Package 'HDLSSkST'

February 2, 2022

Type Package

Title Distribution-Free Exact High Dimensional Low Sample Size k-Sample Tests

Version 2.1.0

Date 2022-02-01

Maintainer Biplab Paul <paul.biplab497@gmail.com>

Description

Testing homogeneity of k multivariate distributions is a classical and challenging problem in statistics, and this becomes even more challenging when the dimension of the data exceeds the sample size.
We construct some tests for this purpose which are exact level (size) alpha tests based on clustering.
These tests are easy to implement and distribution-free in finite sample situations. Under appropriate regularity conditions, these tests have the consistency property in HDLSS asymptotic regime, where the dimension of data grows to infinity while the sample size remains fixed. We also consider a multiscale approach, where the results for different number of partitions are aggregated judiciously. Details are in
Biplab Paul, Shyamal K De and Anil K Ghosh (2021) <doi:10.1016/j.jmva.2021.104897>; Soham Sarkar and Anil K Ghosh (2019) <doi:10.1109/TPAMI.2019.2912599>; William M Rand (1971) <doi:10.1080/01621459.1971.10482356>; Cyrus R Mehta and Nitin R Patel (1983) <doi:10.2307/2288652>; Joseph C Dunn (1973) <doi:10.1080/01969727308546046>; Sture Holm (1979) <doi:10.2307/4615733>; Yoav Benjamini and Yosef Hochberg (1995) <doi:10.2307/2346101>.

License GPL (>= 2)

Imports Rcpp (>= 1.0.3), stats, utils

LinkingTo Rcpp

Author Biplab Paul [aut, cre],
Shyamal K. De [aut],
Anil K. Ghosh [aut]

NeedsCompilation yes

Repository CRAN

Date/Publication 2022-02-02 08:00:08 UTC

R topics documented:

HDLSSkST-package 2
AFStest 3
ARItest 5
BenHoch 8
FStest 9
gMADD 11
gMADD_DI 12
Holm 14
MTFStest 15
MTRItest 18
pmf 20
randfun 21
rctab 22
RItest 23

Index **26**

`HDLSSkST-package` *Distribution-Free Exact High Dimensional Low Sample Size k-Sample Tests*

Description

Testing homogeneity of k ($\geq$ 2) multivariate distributions is a classical and challenging problem in statistics, and this becomes even more challenging when the dimension of the data exceeds the sample size. We construct some tests for this purpose which are exact level (size) α tests based on clustering. These tests are easy to implement and distribution-free in finite sample situations. Under appropriate regularity conditions, these tests have the consistency property in HDLSS asymptotic regime, where the dimension of data d grows to ∞ while the sample size remains fixed. We also consider a multiscale approach, where the results for the different number of partitions are aggregated judiciously. This package includes eight tests, namely (i) RI test, (ii) FS test, (iii) MRI test, (iv) MFS test, (v) MTRI test , (vi) MTFS test, (vii) ARI test and (viii) AFS test. In MRI and MFS test, we modified the RI and FS test, respectively, using an estimated clustering number. In the multiscale approach (MTRI and MTFS), we use Holm's step-down-procedure (1979) and Benjamini-Hochberg FDR controlling procedure (1995).

Author(s)

Biplab Paul, Shyamal K. De and Anil K. Ghosh

Maintainer: Biplab Paul<paul.biplab497@gmail.com>

References

Biplab Paul, Shyamal K De and Anil K Ghosh (2021). Some clustering based exact distribution-free k-sample tests applicable to high dimension, low sample size data, *Journal of Multivariate Analysis*, doi:10.1016/j.jmva.2021.104897.

Soham Sarkar and Anil K Ghosh (2019). On perfect clustering of high dimension, low sample size data, *IEEE transactions on pattern analysis and machine intelligence*, doi:10.1109/TPAMI.2019.2912599.

William M Rand (1971). Objective criteria for the evaluation of clustering methods, *Journal of the American Statistical association*, 66(336):846-850, doi:10.1080/01621459.1971.10482356.

Cyrus R Mehta and Nitin R Patel (1983). A network algorithm for performing Fisher's exact test in rxc contingency tables, *Journal of the American Statistical Association*, 78(382):427-434, doi:10.2307/2288652.

Joseph C Dunn (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters, doi:10.1080/01969727308546046.

Sture Holm (1979). A simple sequentially rejective multiple test procedure, *Scandinavian journal of statistics*, 65-70, doi:10.2307/4615733.

Yoav Benjamini and Yosef Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal statistical society: series B (Methodological)* 57.1: 289-300, doi: 10.2307/2346101.

`AFStest` *k-Sample AFS Test of Equal Distributions*

Description

Performs the distribution free exact k-sample test for equality of multivariate distributions in the HDLSS regime. This an aggregate test of the two sample versions of the FS test over $\frac{k(k-1)}{2}$ numbers of two-sample comparisons, and the test statistic is the minimum of these two sample FS test statistics. Holm's step-down-procedure (1979) and Benjamini-Hochberg procedure (1995) are applied for multiple testing.

Usage

```
AFStest(M, sizes, randomization = TRUE, clust_alg = "knwClustNo", kmax = 4,
multTest = "Holm", s_psi = 1, s_h = 1, lb = 1, n_sts = 1000, alpha = 0.05)
```

Arguments

`M`	$n \times d$ observations matrix of pooled sample, the observations should be grouped by their respective classes
`sizes`	vector of sample sizes
`randomization`	logical; if TRUE (default), randomization test and FALSE, non-randomization test
`clust_alg`	`"knwClustNo"`(default) or `"estclustNo"`; modified K-means algorithm used for clustering

kmax	maximum value of total number of clusters to estimate total number of clusters for two-sample comparition, default: 4
multTest	"HOlm"(default) or "BenHoch"; different multiple tests
s_psi	function required for clustering, 1 for t^2, 2 for $1-\exp(-t)$, 3 for $1-\exp(-t^2)$, 4 for $\log(1+t)$, 5 for t
s_h	function required for clustering, 1 for $\sqrt{t}$, 2 for t
lb	each observation is partitioned into some numbers of smaller vectors of same length lb, default: 1
n_sts	number of simulation of the test statistic, default: 1000
alpha	numeric, confidence level α, default: 0.05

Value

AFStest returns a list containing the following items:

AFSStat	value of the observed test statistic
AFCutoff	cut-off of the test
randomGamma	randomized coefficient of the test
decisionAFS	if returns 1, reject the null hypothesis and if returns 0, fails to reject the null hypothesis
multipleTest	indicates where two populations are different according to multiple tests

Author(s)

Biplab Paul, Shyamal K. De and Anil K. Ghosh

Maintainer: Biplab Paul<paul.biplab497@gmail.com>

References

Biplab Paul, Shyamal K De and Anil K Ghosh (2021). Some clustering based exact distribution-free k-sample tests applicable to high dimension, low sample size data, *Journal of Multivariate Analysis*, doi:10.1016/j.jmva.2021.104897.

Cyrus R Mehta and Nitin R Patel (1983). A network algorithm for performing Fisher's exact test in rxc contingency tables, *Journal of the American Statistical Association*, 78(382):427-434, doi:10.2307/2288652.

Sture Holm (1979). A simple sequentially rejective multiple test procedure, *Scandinavian journal of statistics*, 65-70, doi:10.2307/4615733.

Yoav Benjamini and Yosef Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal statistical society: series B (Methodological)* 57.1: 289-300, doi: 10.2307/2346101.

Examples

```
# muiltivariate normal distribution:
# generate data with dimension d = 500
set.seed(151)
n1=n2=n3=n4=10
d = 500
I1 <- matrix(rnorm(n1*d,mean=0,sd=1),n1,d)
I2 <- matrix(rnorm(n2*d,mean=0.5,sd=1),n2,d)
I3 <- matrix(rnorm(n3*d,mean=1,sd=1),n3,d)
I4 <- matrix(rnorm(n4*d,mean=1.5,sd=1),n4,d)
X <- as.matrix(rbind(I1,I2,I3,I4))
#AFS test:
results <- AFStest(M=X, sizes = c(n1,n2,n3,n4))

 ## outputs:
 results$AFSStat
 #[1] 5.412544e-06

 results$AFCutoff
 #[1] 0.0109604

 results$randomGamma
 #[1] 0

 results$decisionAFS
 #[1] 1

 results$multipleTest
 #  Population.1 Population.2 rejected pvalues
 #1            1            2     TRUE       0
 #2            1            3     TRUE       0
 #3            1            4     TRUE       0
 #4            2            3     TRUE       0
 #5            2            4     TRUE       0
 #6            3            4     TRUE       0
```

ARItest *k-Sample ARI Test of Equal Distributions*

Description

Performs the distribution free exact k-sample test for equality of multivariate distributions in the HDLSS regime. This an aggregate test of the two sample versions of the RI test over $\frac{k(k-1)}{2}$ numbers of two-sample comparisons, and the test statistic is the minimum of these two sample RI test statistics. Holm's step-down-procedure (1979) and Benjamini-Hochberg procedure (1995) are applied for multiple testing.

Usage

```
ARItest(M, sizes, randomization = TRUE, clust_alg = "knwClustNo", kmax = 4,
multTest = "Holm", s_psi = 1, s_h = 1, lb = 1, n_sts = 1000, alpha = 0.05)
```

Arguments

M	$n \times d$ observations matrix of pooled sample, the observations should be grouped by their respective classes
sizes	vector of sample sizes
randomization	logical; if TRUE (default), randomization test and FALSE, non-randomization test
clust_alg	"knwClustNo"(default) or "estclustNo"; modified K-means algorithm used for clustering
kmax	maximum value of total number of clusters to estimate total number of clusters for two-sample comparition, default: 4
multTest	"HOlm"(default) or "BenHoch"; different multiple tests
s_psi	function required for clustering, 1 for t^2, 2 for $1-\exp(-t)$, 3 for $1-\exp(-t^2)$, 4 for $\log(1+t)$, 5 for t
s_h	function required for clustering, 1 for $\sqrt{t}$, 2 for t
lb	each observation is partitioned into some numbers of smaller vectors of same length lb, default: 1
n_sts	number of simulation of the test statistic, default: 1000
alpha	numeric, confidence level α, default: 0.05

Value

ARItest returns a list containing the following items:

ARIStat	value of the observed test statistic
Cutoff	cut-off of the test
randomGamma	randomized coefficient of the test
decisionARI	if returns 1, reject the null hypothesis and if returns 0, fails to reject the null hypothesis
multipleTest	indicates where two populations are different according to multiple tests

Author(s)

Biplab Paul, Shyamal K. De and Anil K. Ghosh

Maintainer: Biplab Paul<paul.biplab497@gmail.com>

References

Biplab Paul, Shyamal K De and Anil K Ghosh (2021). Some clustering based exact distribution-free k-sample tests applicable to high dimension, low sample size data, *Journal of Multivariate Analysis*, doi:10.1016/j.jmva.2021.104897.

William M Rand (1971). Objective criteria for the evaluation of clustering methods, *Journal of the American Statistical association*, 66(336):846-850, doi:10.1080/01621459.1971.10482356.

Sture Holm (1979). A simple sequentially rejective multiple test procedure, *Scandinavian journal of statistics*, 65-70, doi:10.2307/4615733.

Yoav Benjamini and Yosef Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal statistical society: series B (Methodological)* 57.1: 289-300, doi: 10.2307/2346101.

Examples

```
# muiltivariate normal distribution:
# generate data with dimension d = 500
set.seed(151)
n1=n2=n3=n4=10
d = 500
I1 <- matrix(rnorm(n1*d,mean=0,sd=1),n1,d)
I2 <- matrix(rnorm(n2*d,mean=0.5,sd=1),n2,d)
I3 <- matrix(rnorm(n3*d,mean=1,sd=1),n3,d)
I4 <- matrix(rnorm(n4*d,mean=1.5,sd=1),n4,d)
X <- as.matrix(rbind(I1,I2,I3,I4))
#ARI test:
results <- ARItest(M=X, sizes = c(n1,n2,n3,n4))

 ## outputs:
 results$ARIStat
 #[1] 0

 results$ARICutoff
 #[1] 0.3368421

 results$randomGamma
 #[1] 0

 results$decisionARI
 #[1] 1

 results$multipleTest
 #  Population.1 Population.2 rejected pvalues
 #1            1            2     TRUE       0
 #2            1            3     TRUE       0
 #3            1            4     TRUE       0
 #4            2            3     TRUE       0
 #5            2            4     TRUE       0
 #6            3            4     TRUE       0
```

BenHoch *Benjamini-Hochbergs step-up-procedure (1995)*

Description

Benjamini-Hochbergs step-up-procedure (1995) for multiple tests.

Usage

```
BenHoch(pvalues, alpha)
```

Arguments

pvalues	vector of p-values
alpha	numeric, false discovery rate controling level α, default: 0.05

Value

a vector of 0s and 1s. 0: fails to reject the corresponding hypothesis and 1: reject the corresponding hypothesis

Author(s)

Biplab Paul, Shyamal K. De and Anil K. Ghosh

Maintainer: Biplab Paul<paul.biplab497@gmail.com>

References

Yoav Benjamini and Yosef Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal statistical society: series B (Methodological)* 57.1: 289-300, doi: 10.2307/2346101.

Examples

```
# Benjamini-Hochbergs step-up-procedure:
pvalues <- c(0.50,0.01,0.001,0.69,0.02,0.05,0.0025)
alpha <- 0.05
BenHoch(pvalues, alpha)

## outputs:
#[1] 0 1 1 0 1 0 1
```

FStest *k-Sample FS Test of Equal Distributions*

Description

Performs the distribution free exact k-sample test for equality of multivariate distributions in the HDLSS regime.

Usage

```
FStest(M, labels, sizes, n_clust, randomization = TRUE, clust_alg = "knwClustNo",
kmax = 2 * n_clust, s_psi = 1, s_h = 1, lb = 1, n_sts = 1000, alpha = 0.05)
```

Arguments

M	$n \times d$ observations matrix of pooled sample, the observations should be grouped by their respective classes
labels	length n vector of membership index of observations
sizes	vector of sample sizes
n_clust	number of the Populations
randomization	logical; if TRUE (default), randomization test and FALSE, non-randomization test
clust_alg	"knwClustNo"(default) or "estclustNo"(for MFS test); modified K-means algorithm used for clustering
kmax	maximum value of total number of clusters to estimate total number of clusters in the whole observations, default: 2*n_clust
s_psi	function required for clustering, 1 for t^2, 2 for $1-\exp(-t)$, 3 for $1-\exp(-t^2)$, 4 for $\log(1+t)$, 5 for t
s_h	function required for clustering, 1 for $\sqrt{t}$, 2 for t
lb	each observation is partitioned into some numbers of smaller vectors of same length lb, default: 1
n_sts	number of simulation of the test statistic, default: 1000
alpha	numeric, confidence level α, default: 0.05

Value

FStest returns a list containing the following items:

estClustLabel	a vector of length n of estimated class membership index of all observations
obsCtyTab	observed contingency table
ObservedProb	value of the observed test statistic
FCutoff	cut-off of the test
randomGamma	randomized coefficient of the test

estPvalue	estimated p-value of the test
decisionF	if returns 1, reject the null hypothesis and if returns 0, fails to reject the null hypothesis
estClustNo	total number of the estimated classes

Author(s)

Biplab Paul, Shyamal K. De and Anil K. Ghosh

Maintainer: Biplab Paul<paul.biplab497@gmail.com>

References

Biplab Paul, Shyamal K De and Anil K Ghosh (2021). Some clustering based exact distribution-free k-sample tests applicable to high dimension, low sample size data, *Journal of Multivariate Analysis*, doi:10.1016/j.jmva.2021.104897.

Cyrus R Mehta and Nitin R Patel (1983). A network algorithm for performing Fisher's exact test in rxc contingency tables, *Journal of the American Statistical Association*, 78(382):427-434, doi:10.2307/2288652.

Examples

```
# muiltivariate normal distribution:
# generate data with dimension d = 500
set.seed(151)
n1=n2=n3=n4=10
k = 4
d = 500
I1 <- matrix(rnorm(n1*d,mean=0,sd=1),n1,d)
I2 <- matrix(rnorm(n2*d,mean=0.5,sd=1),n2,d)
I3 <- matrix(rnorm(n3*d,mean=1,sd=1),n3,d)
I4 <- matrix(rnorm(n4*d,mean=1.5,sd=1),n4,d)
levels <- c(rep(0,n1), rep(1,n2), rep(2,n3), rep(3,n4))
X <- as.matrix(rbind(I1,I2,I3,I4))
#FS test:
results <- FStest(M=X, labels=levels, sizes = c(n1,n2,n3,n4), n_clust = k)

## outputs:
results$estClustLabel
#[1] 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3

results$obsCtyTab
#      [,1] [,2] [,3] [,4]
#[1,]   10    0    0    0
#[2,]    0   10    0    0
#[3,]    0    0   10    0
#[4,]    0    0    0   10

results$ObservedProb
#[1] 2.125236e-22
```

```
results$FCutoff
#[1] 1.115958e-07

results$randomGamma
#[1] 0

results$estPvalue
#[1] 0

results$decisionF
#[1] 1
```

---

gMADD *Modified K-Means Algorithm by Using a New Dissimilarity Measure, MADD*

Description

Performs modified K-means algorithm by using a new dissimilarity measure, called MADD, and provides estimated cluster (class) labels or memberships of observations.

Usage

```
gMADD(s_psi, s_h, n_clust, lb, M)
```

Arguments

s_psi	function required for clustering, 1 for t^2, 2 for $1-\exp(-t)$, 3 for $1-\exp(-t^2)$, 4 for $\log(1+t)$, 5 for t
s_h	function required for clustering, 1 for $\sqrt{t}$, 2 for t
n_clust	total number of the classes in the whole observations
lb	each observation is partitioned into some numbers of smaller vectors of same length lb
M	$n \times d$ observations matrix of pooled sample, the observations should be grouped by their respective classes

Value

a vector of length n of estimated cluster (class) labels of observations

Author(s)

Biplab Paul, Shyamal K. De and Anil K. Ghosh

Maintainer: Biplab Paul<paul.biplab497@gmail.com>

References

Biplab Paul, Shyamal K De and Anil K Ghosh (2021). Some clustering based exact distribution-free k-sample tests applicable to high dimension, low sample size data, *Journal of Multivariate Analysis*, doi:10.1016/j.jmva.2021.104897.

Soham Sarkar and Anil K Ghosh (2019). On perfect clustering of high dimension, low sample size data, *IEEE transactions on pattern analysis and machine intelligence*, doi:10.1109/TPAMI.2019.2912599.

Examples

```
# Modified K-means algorithm:
# muiltivariate normal distribution
# generate data with dimension d = 500
set.seed(151)
n1=n2=n3=n4=10
d = 500
I1 <- matrix(rnorm(n1*d,mean=0,sd=1),n1,d)
I2 <- matrix(rnorm(n2*d,mean=0.5,sd=1),n2,d)
I3 <- matrix(rnorm(n3*d,mean=1,sd=1),n3,d)
I4 <- matrix(rnorm(n4*d,mean=1.5,sd=1),n4,d)
n_cl <- 4
X <- as.matrix(rbind(I1,I2,I3,I4))
gMADD(1,1,n_cl,1,X)

 ## outputs:
 #[1] 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3
```

`gMADD_DI` *Modified K-Means Algorithm by Using a New Dissimilarity Measure, MADD and DUNN Index*

Description

Performs modified K-means algorithm by using a new dissimilarity measure, called MADD and DUNN index, and provides estimated cluster (class) labels or memberships and corresponding DUNN index of the observations.

Usage

```
gMADD_DI(s_psi, s_h, kmax, lb, M)
```

Arguments

`s_psi`	function required for clustering, 1 for t^2, 2 for $1-\exp(-t)$, 3 for $1-\exp(-t^2)$, 4 for $\log(1+t)$, 5 for t
`s_h`	function required for clustering, 1 for $\sqrt{t}$, 2 for t
`kmax`	maximum value of total number of clusters to estimate total number of clusters in the whole observations

lb	each observation is partitioned into some numbers of smaller vectors of same length lb
M	$n \times d$ observations matrix of pooled sample, the observations should be grouped by their respective classes

Details

DUNN index is used for cluster validation, but here we use it to estimate total number of cluster k by $\hat{k} = argmax_{2 \leq k' \leq k^*} DI(k')$. Here $DI(k')$ represents the DUNN index and we use $k^* = 2 * k$.

Value

a $kmax \times (n + 1)$ matrix of the estimated cluster (class) labels and corresponding DUNN indexes of observations

Note

The result of this gMADD_DI function is a matrix. The 1st row of this matrix doesn't provide anything about estimated class labels or DUNN index of observations since the DUNN index is only defined for $k \geq 2$. The last column of this matrix represents the DUNN indexes. The estimated cluster labels of observations are calculated by finding out the corresponding row of maximum DUNN index.

Author(s)

Biplab Paul, Shyamal K. De and Anil K. Ghosh

Maintainer: Biplab Paul<paul.biplab497@gmail.com>

References

Biplab Paul, Shyamal K De and Anil K Ghosh (2021). Some clustering based exact distribution-free k-sample tests applicable to high dimension, low sample size data, *Journal of Multivariate Analysis*, doi:10.1016/j.jmva.2021.104897.

Soham Sarkar and Anil K Ghosh (2019). On perfect clustering of high dimension, low sample size data, *IEEE transactions on pattern analysis and machine intelligence*, doi:10.1109/TPAMI.2019.2912599.

Joseph C Dunn (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters, doi:10.1080/01969727308546046.

Examples

```
# Modified K-means algorithm:
# muiltivariate normal distribution
# generate data with dimension d = 500
set.seed(151)
n1=n2=n3=n4=10
d = 500
I1 <- matrix(rnorm(n1*d,mean=0,sd=1),n1,d)
I2 <- matrix(rnorm(n2*d,mean=0.5,sd=1),n2,d)
I3 <- matrix(rnorm(n3*d,mean=1,sd=1),n3,d)
```

```
I4 <- matrix(rnorm(n4*d,mean=1.5,sd=1),n4,d)
n_cl <- 4
N <- n1+n2+n3+n4
X <- as.matrix(rbind(I1,I2,I3,I4))
dvec_di_mat <-  gMADD_DI(1,1,2*n_cl,1,X)
est_no_cl <- which.max(dvec_di_mat[ ,(N+1)])
dvec_di_mat[est_no_cl,1:N]

 ## outputs:
 #[1] 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3
```

`Holm` *Holm's step-down-procedure (1979)*

Description

Holm's step-down-procedure (1979) for mutiple tests.

Usage

```
Holm(pvalues, alpha)
```

Arguments

`pvalues`	vector of p-values
`alpha`	numeric, family wise error rate controling level α, default: 0.05

Value

a vector of 0s and 1s. 0: fails to reject the corresponding hypothesis and 1: reject the corresponding hypothesis

Author(s)

Biplab Paul, Shyamal K. De and Anil K. Ghosh

Maintainer: Biplab Paul<paul.biplab497@gmail.com>

References

Sture Holm (1979). A simple sequentially rejective multiple test procedure, *Scandinavian journal of statistics*, 65-70, doi:10.2307/4615733.

Examples

```
# Holm's step down procedure:
pvalues <- c(0.50,0.01,0.001,0.69,0.02,0.05,0.0025)
alpha <- 0.05
Holm(pvalues, alpha)

## outputs:
#[1] 0 0 1 0 0 0 1
```

`MTFStest` *k-Sample MTFS Test of Equal Distributions*

Description

Performs the distribution free exact k-sample test for equality of multivariate distributions in the HDLSS regime. This test is a multiscale approach based on FS test, where the results for different number of partitions are aggregated judiciously.

Usage

```
MTFStest(M, labels, sizes, k_max, multTest = "Holm", s_psi = 1, s_h = 1,
lb = 1, n_sts = 1000, alpha = 0.05)
```

Arguments

`M`	$n \times d$ observations matrix of pooled sample, the observations should be grouped by their respective classes
`labels`	length n vector of membership index of observations
`sizes`	vector of sample sizes
`k_max`	maximum value of total number of clusters which is required for the test
`multTest`	`"HOlm"`(default) or `"BenHoch"`; different multiple tests
`s_psi`	function required for clustering, 1 for t^2, 2 for $1-\exp(-t)$, 3 for $1-\exp(-t^2)$, 4 for $\log(1+t)$, 5 for t
`s_h`	function required for clustering, 1 for $\sqrt{t}$, 2 for t
`lb`	each observation is partitioned into some numbers of smaller vectors of same length lb, default: 1
`n_sts`	number of simulation of the test statistic, default: 1000
`alpha`	numeric, confidence level α, default: 0.05

Value

MTFStest returns a list containing the following items:

RIvec	a vector of the Rand indices based on different number of clusters
Pvalues	a vector of FS test p-values based on different number of clusters
decisionMTRI	if returns 1, reject the null hypothesis and if returns 0, fails to reject the null hypothesis
contTabs	a list of the observed contingency table based on different number of clusters
mulTestdec	a vector of 0s and 1s. 0: fails to reject the corresponding hypothesis and 1: reject the corresponding hypothesis

Author(s)

Biplab Paul, Shyamal K. De and Anil K. Ghosh

Maintainer: Biplab Paul<paul.biplab497@gmail.com>

References

Biplab Paul, Shyamal K De and Anil K Ghosh (2021). Some clustering based exact distribution-free k-sample tests applicable to high dimension, low sample size data, *Journal of Multivariate Analysis*, doi:10.1016/j.jmva.2021.104897.

Sture Holm (1979). A simple sequentially rejective multiple test procedure, *Scandinavian journal of statistics*, 65-70, doi:10.2307/4615733.

Yoav Benjamini and Yosef Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal statistical society: series B (Methodological)* 57.1: 289-300, doi: 10.2307/2346101.

Examples

```
# muiltivariate normal distribution:
# generate data with dimension d = 500
set.seed(151)
n1=n2=n3=n4=10
d = 500
I1 <- matrix(rnorm(n1*d,mean=0,sd=1),n1,d)
I2 <- matrix(rnorm(n2*d,mean=0.5,sd=1),n2,d)
I3 <- matrix(rnorm(n3*d,mean=1,sd=1),n3,d)
I4 <- matrix(rnorm(n4*d,mean=1.5,sd=1),n4,d)
levels <- c(rep(0,n1), rep(1,n2), rep(2,n3), rep(3,n4))
X <- as.matrix(rbind(I1,I2,I3,I4))
#MTFS test:
results <- MTFStest(X, levels, c(n1,n2,n3,n4), 8)

 ## outputs:
 results$fpmfvec
#[1] 7.254445e-12 6.137740e-16 2.125236e-22 2.125236e-22 2.125236e-22 2.125236e-22 2.125236e-22

 results$Pvalues
```

```
#[1] 0 0 0 0 0 0 0

results$decisionMTFS
#[1] 1

results$contTabs
#$contTabs[[1]]
#     [,1] [,2]
#[1,]   10    0
#[2,]   10    0
#[3,]    0   10
#[4,]    0   10

#$contTabs[[2]]
#    [,1] [,2] [,3]
#[1,]   10    0    0
#[2,]    0   10    0
#[3,]    0    8    2
#[4,]    0    0   10

#$contTabs[[3]]
#     [,1] [,2] [,3] [,4]
#[1,]   10    0    0    0
#[2,]    0   10    0    0
#[3,]    0    0   10    0
#[4,]    0    0    0   10

#$contTabs[[4]]
#     [,1] [,2] [,3] [,4] [,5]
#[1,]   10    0    0    0    0
#[2,]    0   10    0    0    0
#[3,]    0    0    4    6    0
#[4,]    0    0    0    0   10

#$contTabs[[5]]
#    [,1] [,2] [,3] [,4] [,5] [,6]
#[1,]   10    0    0    0    0    0
#[2,]    0   10    0    0    0    0
#[3,]    0    0    4    6    0    0
#[4,]    0    0    0    0    8    2

#$contTabs[[6]]
#     [,1] [,2] [,3] [,4] [,5] [,6] [,7]
#[1,]   10    0    0    0    0    0    0
#[2,]    0    5    5    0    0    0    0
#[3,]    0    0    0    4    6    0    0
#[4,]    0    0    0    0    0    8    2

#$contTabs[[7]]
#     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
#[1,]    8    2    0    0    0    0    0    0
#[2,]    0    0    5    5    0    0    0    0
#[3,]    0    0    0    0    4    6    0    0
```

```
#[4,]    0    0    0    0    0    0    8    2


results$mulTestdec
#[1] 1 1 1 1 1 1 1
```

`MTRItest` *k-Sample MTRI Test of Equal Distributions*

Description

Performs the distribution free exact k-sample test for equality of multivariate distributions in the HDLSS regime. This test is a multiscale approach based on RI test, where the results for different number of partitions are aggregated judiciously.

Usage

```
MTRItest(M, labels, sizes, k_max, multTest = "Holm", s_psi = 1, s_h = 1,
lb = 1, n_sts = 1000, alpha = 0.05)
```

Arguments

`M`	$n \times d$ observations matrix of pooled sample, the observations should be grouped by their respective classes
`labels`	length n vector of membership index of observations
`sizes`	vector of sample sizes
`k_max`	maximum value of total number of clusters which is required for the test
`multTest`	`"HOlm"`(default) or `"BenHoch"`; different multiple tests
`s_psi`	function required for clustering, 1 for t^2, 2 for $1-\exp(-t)$, 3 for $1-\exp(-t^2)$, 4 for $\log(1+t)$, 5 for t
`s_h`	function required for clustering, 1 for $\sqrt{t}$, 2 for t
`lb`	each observation is partitioned into some numbers of smaller vectors of same length lb, default: 1
`n_sts`	number of simulation of the test statistic, default: 1000
`alpha`	numeric, confidence level α, default: 0.05

Value

MTRItest returns a list containing the following items:

`RIvec`	a vector of the Rand indices based on different number of clusters
`Pvalues`	a vector of RI test p-values based on different number of clusters
`decisionMTRI`	if returns 1, reject the null hypothesis and if returns 0, fails to reject the null hypothesis
`contTabs`	a list of the observed contingency table based on different number of clusters
`mulTestdec`	a vector of 0s and 1s. 0: fails to reject the corresponding hypothesis and 1: reject the corresponding hypothesis

Author(s)

Biplab Paul, Shyamal K. De and Anil K. Ghosh

Maintainer: Biplab Paul<paul.biplab497@gmail.com>

References

Biplab Paul, Shyamal K De and Anil K Ghosh (2021). Some clustering based exact distribution-free k-sample tests applicable to high dimension, low sample size data, *Journal of Multivariate Analysis*, doi:10.1016/j.jmva.2021.104897.

Sture Holm (1979). A simple sequentially rejective multiple test procedure, *Scandinavian journal of statistics*, 65-70, doi:10.2307/4615733.

Yoav Benjamini and Yosef Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal statistical society: series B (Methodological)* 57.1: 289-300, doi: 10.2307/2346101.

Examples

```
# muiltivariate normal distribution:
# generate data with dimension d = 500
set.seed(151)
n1=n2=n3=n4=10
d = 500
I1 <- matrix(rnorm(n1*d,mean=0,sd=1),n1,d)
I2 <- matrix(rnorm(n2*d,mean=0.5,sd=1),n2,d)
I3 <- matrix(rnorm(n3*d,mean=1,sd=1),n3,d)
I4 <- matrix(rnorm(n4*d,mean=1.5,sd=1),n4,d)
levels <- c(rep(0,n1), rep(1,n2), rep(2,n3), rep(3,n4))
X <- as.matrix(rbind(I1,I2,I3,I4))
#MTRI test:
results <- MTRItest(X, levels, c(n1,n2,n3,n4), 8)

## outputs:
results$RIvec
#[1] 0.25641026 0.14871795 0.00000000 0.03076923 0.05128205 0.08333333 0.10384615

results$Pvalues
#[1] 0 0 0 0 0 0 0

results$decisionMTRI
#[1] 1

results$contTabs
#$contTabs[[1]]
#     [,1] [,2]
#[1,]   10    0
#[2,]   10    0
#[3,]    0   10
#[4,]    0   10

#$contTabs[[2]]
```

```
#     [,1] [,2] [,3]
#[1,]   10    0    0
#[2,]    0   10    0
#[3,]    0   10    0
#[4,]    0    0   10

#$contTabs[[3]]
#     [,1] [,2] [,3] [,4]
#[1,]   10    0    0    0
#[2,]    0   10    0    0
#[3,]    0    0   10    0
#[4,]    0    0    0   10

#$contTabs[[4]]
#     [,1] [,2] [,3] [,4] [,5]
#[1,]   10    0    0    0    0
#[2,]    0   10    0    0    0
#[3,]    0    0    4    6    0
#[4,]    0    0    0    0   10

#$contTabs[[5]]
#     [,1] [,2] [,3] [,4] [,5] [,6]
#[1,]   10    0    0    0    0    0
#[2,]    0   10    0    0    0    0
#[3,]    0    0    4    6    0    0
#[4,]    0    0    0    0    8    2

#$contTabs[[6]]
#     [,1] [,2] [,3] [,4] [,5] [,6] [,7]
#[1,]   10    0    0    0    0    0    0
#[2,]    0    5    5    0    0    0    0
#[3,]    0    0    0    4    6    0    0
#[4,]    0    0    0    0    0    8    2

#$contTabs[[7]]
#     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
#[1,]    8    2    0    0    0    0    0    0
#[2,]    0    0    5    5    0    0    0    0
#[3,]    0    0    0    0    4    6    0    0
#[4,]    0    0    0    0    0    0    8    2


results$mulTestdec
#[1] 1 1 1 1 1 1 1
```

---

pmf *Generalized Hypergeometric Probability*

Description

A function that provides the probability of observing an $r \times c$ contingency table using generalized hypergeometric probability.

Usage

```
pmf(M)
```

Arguments

M	$r \times c$ contingency table

Value

a single value between 0 and 1

Author(s)

Biplab Paul, Shyamal K. De and Anil K. Ghosh

Maintainer: Biplab Paul<paul.biplab497@gmail.com>

References

Biplab Paul, Shyamal K De and Anil K Ghosh (2021). Some clustering based exact distribution-free k-sample tests applicable to high dimension, low sample size data, *Journal of Multivariate Analysis*, doi:10.1016/j.jmva.2021.104897.

Cyrus R Mehta and Nitin R Patel (1983). A network algorithm for performing Fisher's exact test in rxc contingency tables, *Journal of the American Statistical Association*, 78(382):427-434, doi:10.2307/2288652.

Examples

```
# Generalized hypergeometric probability of rxc Contingency Table:
mat <- matrix(1:20,5,4, byrow = TRUE)
pmf(mat)

## outputs:
#[1] 4.556478e-09
```

`randfun` *Rand Index*

Description

Measures to compare the dissimilarity of exact cluster labels (memberships) and estimated cluster labels (memberships) of the observations.

Usage

```
randfun(lvel, dv)
```

Arguments

lvel	exact cluster labels of the observations
dv	estimated cluster labels of the observations

Value

a single value between 0 and 1

Author(s)

Biplab Paul, Shyamal K. De and Anil K. Ghosh

Maintainer: Biplab Paul<paul.biplab497@gmail.com>

References

William M Rand (1971). Objective criteria for the evaluation of clustering methods, *Journal of the American Statistical association*, 66(336):846-850, doi:10.1080/01621459.1971.10482356.

Examples

```
# Measures of dissimilarity:
exl <- c(rep(0,5), rep(1,5), rep(2,5), rep(3,5))
el <- c(0,0,1,0,0,1,2,1,0,1,2,2,3,2,2,3,2,3,1,3)
randfun(exl,el)

## outputs:
#[1] 0.2368421
```

---

rctab *Generates an $r \times c$ Contingency Table*

Description

A function that generates an $r \times c$ contingency table with the same marginal totals as given $r \times c$ contingency table.

Usage

```
rctab(M)
```

Arguments

M	$r \times c$ contingency table

Value

generated $r \times c$ contingency table

Author(s)

Biplab Paul, Shyamal K. De and Anil K. Ghosh

Maintainer: Biplab Paul<paul.biplab497@gmail.com>

References

Cyrus R Mehta and Nitin R Patel (1983). A network algorithm for performing Fisher's exact test in rxc contingency tables, *Journal of the American Statistical Association*, 78(382):427-434, doi:10.2307/2288652.

Examples

```
# Generation of rxc Contingency Table:
   set.seed(151)
   mat <- matrix(1:20,5,4, byrow = TRUE)
   rctab(mat)

   ## outputs:
   #      [,1] [,2] [,3] [,4]
   #[1,]    3    4    0    3
   #[2,]    4    5   10    7
   #[3,]    8    7   12   15
   #[4,]   18   16   13   11
   #[5,]   12   18   20   24
```

RItest *k-Sample RI Test of Equal Distributions*

Description

Performs the distribution free exact k-sample test for equality of multivariate distributions in the HDLSS regime.

Usage

```
RItest(M, labels, sizes, n_clust, randomization = TRUE, clust_alg = "knwClustNo",
kmax = 2 * n_clust, s_psi = 1, s_h = 1, lb = 1, n_sts = 1000, alpha = 0.05)
```

Arguments

M	$n \times d$ observations matrix of pooled sample, the observations should be grouped by their respective classes
labels	length n vector of membership index of observations
sizes	vector of sample sizes
n_clust	number of the Populations
randomization	logical; if TRUE (default), randomization test and FALSE, non-randomization test
clust_alg	"knwClustNo"(default) or "estclustNo"(for MRI test); modified K-means algorithm used for clustering
kmax	maximum value of total number of clusters to estimate total number of clusters in the whole observations, default: 2*n_clust
s_psi	function required for clustering, 1 for t^2, 2 for $1-\exp(-t)$, 3 for $1-\exp(-t^2)$, 4 for $\log(1+t)$, 5 for t
s_h	function required for clustering, 1 for $\sqrt{t}$, 2 for t
lb	each observation is partitioned into some numbers of smaller vectors of same length lb, default: 1
n_sts	number of simulation of the test statistic, default: 1000
alpha	numeric, confidence level α, default: 0.05

Value

RItest returns a list containing the following items:

estClustLabel	a vector of length n of estimated class membership index of all observations
obsCtyTab	observed contingency table
ObservedRI	value of the observed test statistic
RICutoff	cut-off of the test
randomGamma	randomized coefficient of the test
estPvalue	estimated p-value of the test
decisionRI	if returns 1, reject the null hypothesis and if returns 0, fails to reject the null hypothesis
estClustNo	total number of the estimated classes

Author(s)

Biplab Paul, Shyamal K. De and Anil K. Ghosh

Maintainer: Biplab Paul<paul.biplab497@gmail.com>

References

Biplab Paul, Shyamal K De and Anil K Ghosh (2021). Some clustering based exact distribution-free k-sample tests applicable to high dimension, low sample size data, *Journal of Multivariate Analysis*, doi:10.1016/j.jmva.2021.104897.

William M Rand (1971). Objective criteria for the evaluation of clustering methods, *Journal of the American Statistical association*, 66(336):846-850, doi:10.1080/01621459.1971.10482356.

Examples

```
# muiltivariate normal distribution:
# generate data with dimension d = 500
set.seed(151)
n1=n2=n3=n4=10
k = 4
d = 500
I1 <- matrix(rnorm(n1*d,mean=0,sd=1),n1,d)
I2 <- matrix(rnorm(n2*d,mean=0.5,sd=1),n2,d)
I3 <- matrix(rnorm(n3*d,mean=1,sd=1),n3,d)
I4 <- matrix(rnorm(n4*d,mean=1.5,sd=1),n4,d)
levels <- c(rep(0,n1), rep(1,n2), rep(2,n3), rep(3,n4))
X <- as.matrix(rbind(I1,I2,I3,I4))
# RI test:
results <- RItest(M=X, labels=levels, sizes = c(n1,n2,n3,n4), n_clust = k)

 ## outputs:
 results$estClustLabel
 #[1] 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3

 results$obsCtyTab
 #      [,1] [,2] [,3] [,4]
 #[1,]   10    0    0    0
 #[2,]    0   10    0    0
 #[3,]    0    0   10    0
 #[4,]    0    0    0   10

 results$ObservedRI
 #[1] 0

 results$RICutoff
 #[1] 0.3307692

 results$randomGamma
 #[1] 0

 results$estPvalue
 #[1] 0

 results$decisionRI
 #[1] 1
```

Index

∗ **AFStest function**
AFStest, 3
∗ **ARItest function**
ARItest, 5
∗ **BenHoch function**
BenHoch, 8
∗ **FStest function**
FStest, 9
∗ **HDLSS**
HDLSSkST-package, 2
∗ **Holm function**
Holm, 14
∗ **MTFStest function**
MTFStest, 15
∗ **MTRItest function**
MTRItest, 18
∗ **RItest function**
RItest, 23
∗ **gMADD function**
gMADD, 11
∗ **gMADD_DI function**
gMADD_DI, 12
∗ **package**
HDLSSkST-package, 2
∗ **pmf function**
pmf, 20
∗ **randfun function**
randfun, 21
∗ **rctab function**
rctab, 22

AFStest, 3
ARItest, 5

BenHoch, 8

FStest, 9

gMADD, 11
gMADD_DI, 12

HDLSSkST-package, 2
Holm, 14

MTFStest, 15
MTRItest, 18

pmf, 20

randfun, 21
rctab, 22
RItest, 23