Package 'multicross'

May 25, 2020

Type Package

Title A Graph-Based Test for Comparing Multivariate Distributions in the Multi Sample Framework

Version 2.1.0

Author Somabha Mukherjee <somabha@wharton.upenn.edu>
Divyansh Agarwal <divyansh.agarwal@pennmedicine.upenn.edu>
Bhaswar Bhattacharya <bhaswar@wharton.upenn.edu>
Nancy R. Zhang <nzh@wharton.upenn.edu>

Maintainer Divyansh Agarwal <divyansh.agarwal@pennmedicine.upenn.edu>

Description We introduce a nonparametric, graphical test based on optimal matching for assessing whether multiple unknown multivariate probability distributions are equal. This method is consistent, and does not make any distributional assumptions on the data. Our procedure combines data that belong to different classes or groups to create a graph on the pooled data, and then utilizes the number of edges connecting data points from different classes to examine equality of distributions among the classes. The functions available through this package implement the work described here: <arXiv:1906.04776>.

Depends R (>= 3.5.0),

License GPL (>= 2)

Encoding UTF-8

LazyData true

Imports stats (>= 3.5.0), MASS (>= 7.3-49), Matrix (>= 1.2-17), nbpMatching (>= 1.5.1), crossmatch (>= 1.3-1),

Suggests ape

RoxygenNote 7.1.0

NeedsCompilation no

Repository CRAN

Date/Publication 2020-05-25 20:50:03 UTC

R topics documented:

`mcm` *Multisample generalization of Rosenbaum's crossmatch test*

Description

In this packcage, we present a framework inspired by Rosenbaum's crossmatch idea to tackle the nonparametric, multisample problem wherein one is concerned with testing the equality of K unknown multivariate probability distributions. We implement two tests: the first is a multisample generalization of Rosenbaum's crossmatch (MCM), and the other further introduces a Malahnobis-type modification to the test (MMCM).

Usage

```
mcm(data_list, level)
```

Arguments

`data_list`	is list of multifeature matrices corresponding to the K different classes, so each element of the list is a matrix, for a total of K matrices. Each matrix contains observations as the rows and features as the columns
`level`	is the level alpha for hypothesis testing

Value

The p-value corresponding to rejection of the alternative, along with the decision of the hypothesis testing (Null being accepted versus rejected)

Examples

```
# Simulation Example when the user wants to test whether K=3 multivariate distributions are equal:
X1 = MASS::mvrnorm(10,rep(0,4),diag(2,4),tol=1e-6, empirical=FALSE, EISPACK=FALSE)
X2 = MASS::mvrnorm(10,rep(0,4),diag(1,4),tol=1e-6, empirical=FALSE, EISPACK=FALSE)
X3 = MASS::mvrnorm(10,rep(0,4),diag(3,4),tol=1e-6, empirical=FALSE, EISPACK=FALSE)
mcm(list(X1,X2,X3),0.05)
```

`mhcccreate` *Creates the null covariance matrix for mmcm, corresponding to the scenario when all K distributions are the same*

Description

Creates the null covariance matrix for mmcm, corresponding to the scenario when all K distributions are the same

Usage

```
mhcccreate(nvec)
```

Arguments

`nvec`	is a vector containing the sizes of the K different classes

Value

The inputs for the Multisample Mahalanobis Crossmatch Test

`mhccexecutelong` *Calculates the pairwise crosscounts for the K classes being examined*

Description

Calculates the pairwise crosscounts for the K classes being examined

Usage

```
mhccexecutelong(nvec, apmat)
```

Arguments

`nvec`	is a vector containing the sizes of the K different classes
`apmat`	is the data matrix containing pooled data from each of the K classess, on which optimal non-bipartite matching is performed.

Value

The inputs for the Multisample Mahalanobis Crossmatch Tests

`mmcm` *Use the Mahalnobis-type multisample test based on optimal matching to compare K different multivariate distributions*

Description

Use the Mahalnobis-type multisample test based on optimal matching to compare K different multivariate distributions

Usage

```
mmcm(data_list, level)
```

Arguments

`data_list`	is list of multifeature matrices corresponding to the K different classes, so each element of the list is a matrix, for a total of K matrices.
`level`	is the cutoff value (alpha) for hypothesis testing

Value

The p-value corresponding to rejection of the alternative, along with the decision of the hypothesis testing (Null being accepted versus rejected)

Examples

```
# Simulation Example when the user wants to test whether K=3 multivariate distributions are equal:
X1 = MASS::mvrnorm(10,rep(0,4),diag(2,4),tol=1e-6, empirical=FALSE, EISPACK=FALSE)
X2 = MASS::mvrnorm(10,rep(0,4),diag(1,4),tol=1e-6, empirical=FALSE, EISPACK=FALSE)
X3 = MASS::mvrnorm(10,rep(0,4),diag(3,4),tol=1e-6, empirical=FALSE, EISPACK=FALSE)
mmcm(list(X1,X2,X3), 0.05)
```

`multigene` *Given two input matrices with the same number of observations but differrent number of variables, this function returns the largest canonical correlation between variables of matrix 1 (X) and those of matrix 2 (Y).*

Description

Given two input matrices with the same number of observations but differrent number of variables, this function returns the largest canonical correlation between variables of matrix 1 (X) and those of matrix 2 (Y).

Usage

```
multigene(X, Y)
```

Arguments

X	is a data matrix with observations as rows and features/genes as columns.
Y	is a data matrix with observations as rows (same observations as X) and a different set of features/genes as columns.

Value

The largest canonical correlation between X and Y as described in <https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Canonical_Correlation.pdf>.

---

select_class *When the MCM/MMCM tests reject the null, class selection can help determine which of the K classes are the likely contributors for rejection*

Description

When the MCM/MMCM tests reject the null, class selection can help determine which of the K classes are the likely contributors for rejection

Usage

```
select_class(data_list, level)
```

Arguments

data_list	is list of multifeature matrices corresponding to the K different classes, so each element of the list is a matrix, for a total of K matrices.
level	is the cutoff value (alpha) for hypothesis testing

Value

A table of pairwise comparisons among the K classes, to further probe which class influences the rejection of the null the most. No p-value adjustment is made to these reported p-values

Examples

```
# Simulation Example when the user wants to test whether K=3 multivariate distributions are equal:
X1 = MASS::mvrnorm(10,rep(0,4),diag(2,4),tol=1e-6, empirical=FALSE, EISPACK=FALSE)
X2 = MASS::mvrnorm(10,rep(0,4),diag(1,4),tol=1e-6, empirical=FALSE, EISPACK=FALSE)
X3 = MASS::mvrnorm(10,rep(0,4),diag(3,4),tol=1e-6, empirical=FALSE, EISPACK=FALSE)
select_class(list(X1,X2,X3), 0.05)
```

`split_mat` *Split a data frame or matrix into subsets based on a particular categorical variable*

Description

Split a data frame or matrix into subsets based on a particular categorical variable

Usage

```
split_mat(obj, by)
```

Arguments

`obj`	is a data frame or matrix to be split into subsets, divided by the categorical variable
`by`	is a character-string that specifies the columns that need to be subsetted

Value

A list containing the subsetted data sets. The names of the list corresponds to the value of the subsetted list

Index