Package 'TwoSampleTest.HD'

February 27, 2023

Title A Two-Sample Test for the Equality of Distributions for High-Dimensional Data

Version 1.2

Maintainer Marta Cousido Rocha <martacousido@uvigo.es>

Description For high-dimensional data whose main feature is a large number, p, of variables but a small sample size, the null hypothesis that the marginal distributions of p variables are the same for two groups is tested. We propose a test statistic motivated by the simple idea of comparing, for each of the p variables, the empirical characteristic functions computed from the two samples. If one rejects this global null hypothesis of no differences in distributions between the two groups, a set of permutation p-values is reported to identify which variables are not equally distributed in both groups.

Depends R (>= 3.5.0)

License GPL-2

Encoding UTF-8

RoxygenNote 7.2.3

NeedsCompilation no

Author Marta Cousido Rocha [aut, cre],
José Carlos Soage González [ctr],
Jacobo de Uña Álvarez [aut, ths],
Jeffrey D. Hart [aut]

Repository CRAN

Date/Publication 2023-02-27 09:12:33 UTC

R topics documented:

`TwoSampleTest.HD-package` *A two-sample test for the equality of distributions for high-dimensional data*

Description

This package implements four different tests proposed in Cousido-Rocha et al. (2018). These methods test the (global) null hypothesis of equality of the univariate marginals of the p-variate distributions in the two populations. In other words, the null hypothesis is an intersection of the p null hypotheses corresponding to p different two-sample problems. These methods are particularly well suited to the low sample size, high dimensional setting (n « p). The sample size can be as small as 2. The test accounts for the possibility that the p variables in each data set can be weakly dependent. Three of the methods arise from different approaches to estimate the variance of the same statistic. This statistic averages p individual statistics based on comparing the empirical characteristic functions computed from the two samples. The last method is an alternative global test whose statistic averages the p-values derived from applying permutation tests to the individual statistics mentioned above. When the global null hypothesis is rejected such permutation p-values can also be used to identify which variables contribute to this significance. The standardized version of each test statistic and its p-value are computed among other things.

Details

Package 'TwoSampleTest.HD'

Documentation for package 'TwoSampleTest.HD' version 1.2

- Package: 'TwoSampleTest.HD'
- Version: 1.2
- Maintainer: Marta Cousido Rocha <martacousido@uvigo.es>
- License: GPL-2

Value

- 'TwoSampleTest.HD'

Acknowledgements

This work has received financial support of the Call 2015 Grants for PhD contracts for training of doctors of the Ministry of Economy and Competitiveness, co-financed by the European Social Fund (Ref. BES-2015-074958). The authors acknowledge support from MTM2014-55966-P project, Ministry of Economy and Competitiveness, and MTM2017-89422-P project, Ministry of Economy, Industry and Competitiveness, State Research Agency, and Regional Development Fund, UE. The authors also acknowledge the financial support provided by the SiDOR research group through the grant Competitive Reference Group, 2016-2019 (ED431C 2016/040), funded by the "Consellería de Cultura, Educación e Ordenación Universitaria. Xunta de Galicia".

Author(s)

- Cousido Rocha, Marta.
- Soage González, José Carlos.
- de Uña-Álvarez, Jacobo.
- D. Hart, Jeffrey.

References

Cousido-Rocha, M., de Uña-Álvarez J., and Hart, J. (2018). A two-sample test for the equality of distributions for high-dimensional data. Preprint.

`TwoSampleTest.HD` *A two-sample test for the equality of distributions for high-dimensional data*

Description

Performs the four tests of equality of the p marginal distributions for two groups proposed in Cousido- Rocha et al.(2018). The methods have been designed for the low sample size and high dimensional setting. Furthermore, the possibility that the p variables in each data set can be weakly dependent is considered. The function also reports a set of p permutation p-values, each of which is derived from testing the equality of distributions in the two groups for each of the variables separately. These p-values are useful when a proposed test rejects the global null hypothesis since it makes it possible to identify which variables have contributed to this significance.

Usage

```
TwoSampleTest.HD(
  X,
  Y,
  method = c("spect", "spect_ind", "boot", "us", "us_ind", "perm"),
  I.permutation.p.values = FALSE,
  b_I.permutation.p.values = c("global", "individual")
)
```

Arguments

`X`	A matrix where each row is one of the p-samples in the first group.
`Y`	A matrix where each row is one of the p-samples in the second group.
`method`	the two-sample test. By default the "us" method is computed. See details.
`I.permutation.p.values`	Logical. Default is FALSE. A variable indicating whether to compute the permutation p-values or not when the selected method is not "perm". See details.
`b_I.permutation.p.values`	The method used to compute the individual statistics on which are based the permutation p-values. Default is "global". See details.

Details

The function implements the two-sample tests proposed by Cousido-Rocha, et al. (2018). The methods "spect", "boot" and "us" are based on a global statistic which is the average of p individual statistics corresponding to each of the p variables. Each of these individual statistics measures the difference between the empirical characteristic functions computed from the two samples. An alternative expression shows that each statistic is essentially the intergrated squared difference between kernel density estimates. The global statistic (average) is standardized using one of three different variance estimators. The method "spect" uses a variance estimator based on spectral analysis, the method "boot" implements the block bootstrap to estimate the variance and the method "us" employs a variance estimator derived from U-statistic theory (more details in Cousido-Rocha et al., 2018). The methods "spect" and "boot" are suitable under some assumptions including that the sequence of individual statistics that define the global statistic is strictly stationary, whereas the method "us" avoids this assumption. However the methods "spect" and "boot" have been checked in simulations and they perform well even when the stationarity assumption is violated. The methods "spect" and "us" have their corresponding versions for independent data ("spect ind" and "us ind"), for which the variance estimators are simplified taking into acount the independence of the variables. The asymptotic normality (when p tends to infinity) of the standardized version of the statistic is used to compute the corresponding p-value. On the other hand, Cousido-Rocha et al. (2018) also proposed the method "perm" whose global statistic is the average of the permutation p-values corresponding to the individual statistics mentioned above. This method assumes that the sequence of p-values is strictly stationary, however in simulations it seems that it performs well when this assumption does not hold. In addition to providing an alternative global test, these p-values can be used when the global null hypothesis is rejected and one wishes to identify which of the p variables have contributed to that rejection. The global statistic depends on a parameter which plays a role similar to that of a smoothing parameter or bandwidth in kernel density estimation. For the four global tests this parameter is estimated using the information from all the variables or features. For the individual statistics from which the permutation p-values are computed, there are two possibilities: (i) use the value employed in the global test (b_I.permutation.p.values="global"), (ii) estimate this parameter for each variable separately using only its sample information (b_I.permutation.p.values="individual")).

Value

A list containing the following components:

`standardized statistic:`
: the value of the standardized statistic.

`p.value:` the p-value for the test.

`statistic:` the value of the statistic.

`variance:` the value of the variance estimator.

`p:` number of samples or populations.

`n:` sample size in the first group.

`m:` sample size in the second group.

`method:` a character string indicating which two sample test is performed.

`I.statistics:` the p individual statistics.

`I.permutation.p.values:` the p individual permutation p-values.

`data.name:` a character string giving the name of the data.

Author(s)

- Marta Cousido-Rocha
- José Carlos Soage González
- Jacobo de Uña-Álvarez
- Jeffrey D. Hart

References

Cousido-Rocha, M., de Uña-Álvarez J., and Hart, J. (2018). A two-sample test for the equality of marginal distributions for high-dimensional data. Preprint.

Examples

```
# We consider a simulated data example. We have simulated the following situation.
# We have two groups, for example, 7 patients with tumor 1 and 7 patients with tumor 2.
# For each patient 1000 variables are measured, for example, gene expression levels.
# Besides, the distributions of 100 of the variables are different in the two groups,
# and the differences are in terms of location. The variables are independent to
# simplify the generation of the data sets.
p <- 1000
n = m = 7
inds <- sample(1:4, p, replace = TRUE)
X <- matrix(rep(0, n * p), ncol = n)
for (j in 1:p){
  if (inds[j] == 1){
    X[j, ] <- rnorm(n)
    }
  if (inds[j] == 2){
    X[j, ] <- rnorm(n, sd = 2)
  }
  if (inds[j] == 3){
    X[j, ] <- rnorm(n, mean = 1)
  }
  if (inds[j] == 4){
    X[j, ] <- rnorm(n, mean = 1, sd = 2)
  }
 }
rho <-  0.1
ind <- sample(1:p, rho * p)
li <- length(ind)
indsy <- inds
for (l in 1:li){
```

```
  if (indsy[ind[l]]==1){
     indsy[ind[l]]=3
     } else {
       if (indsy[ind[l]]==2){
          indsy[ind[l]]=4
           } else {
            if (indsy[ind[l]]==3){
               indsy[ind[l]]=1
               } else {
                indsy[ind[l]] = 2
                }
               }
             }
            }
Y <- matrix(rep(0, m * p), ncol = m)
for (j in 1:p){
  if (indsy[j] == 1){
    Y[j,] <- rnorm(m)}
  if (indsy[j] == 2){
    Y[j, ] <- rnorm(m, sd = 2)
  }
  if (indsy[j]==3){
    Y[j, ] <- rnorm(m, mean = 1)
  }
  if (indsy[j] == 4){
    Y[j,] <- rnorm(m, mean = 1, sd = 2)
  }
 }

# Our interest is to test the null hypothesis that the distribution of each of the 1000 variables
# is the same in the two groups.

# We use for this purpose the four methods proposed in Cousido-Rocha et al. (2018).

res1 <- TwoSampleTest.HD(X, Y, method = "spect")
res1
res2 <- TwoSampleTest.HD(X, Y, method = "boot")
res2
res3 <- TwoSampleTest.HD(X, Y, method = "us")
res3
res4 <- TwoSampleTest.HD(X, Y, method = "perm")
res4
# The four methods reject the global null hypothesis.
# Hence, we use the individual permutation p-values
# to identify which variables are not equally distributed in the two groups.
pv<-res4$I.permutation.p.values

# Applying a multiple testing procedure to these p-values
# we can detect the variables with different distributions for the two groups.
# The following plot of the individual permutation p-values is also informative.
# We remark in red the 100 smallest p-values.

pv_sort <- sort(pv)
```

```
cri <- pv_sort[100]
ind <- which(pv <= cri)
plot(1:p, pv, main = "Individual permutation p-values",
     xlab = "Variables", ylab = "p-values")
points(ind, pv[ind], col = "red")
```

Index