Package 'Equalden.HD'

January 20, 2025

Title Testing the Equality of a High Dimensional Set of Densities

Version 1.2.1

Maintainer Marta Cousido Rocha <martacousido@uvigo.es>

Description The equality of a large number k of densities is tested by measuring the L2 distance between the corresponding kernel density estimators and the one based on the pooled sample. The test even works for sample sizes as small as 2.

Depends R (>= 3.4.0)

ByteCompile true

License GPL-2

Encoding UTF-8

LazyData true

RoxygenNote 7.2.3

NeedsCompilation yes

Author Marta Cousido Rocha [aut, cre],
José Carlos Soage González [ctr],
Jacobo de Uña Álvarez [aut, ths],
Jeffrey D. Hart [aut],
Ivan Kojadinovic [cph],
A. Patton [cph],
C. Parmeter [cph],
J. Racine [cph]

Repository CRAN

Date/Publication 2024-10-19 18:50:02 UTC

Contents

`Equalden.HD-package` *Package Equalden.HD*

Description

This package implements three different methods to test the null hypothesis that a large number k of samples have a common density. The sample size can be as small as 2. These methods are particularly well suited to the low sample size, high dimensional setting (n « k). The first method, proposed by Zhan and Hart (2012), was developed to test the null hypothesis when the samples are independent of each other. The other tests, proposed by Cousido-Rocha et al. (2018), are adaptations of the test in Zhan and Hart (2012) for the setting in which the samples are weakly dependent. The standarized version of each test statistic and its p-value are computed among other things.

Details

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

Value

- Equalden.test.HD: Performs the k-sample test proposed in Zhan and Hart (2012) for the setting of low sample size, large dimension and independent samples, and its adaptions to dependent samples proposed in Cousido-Rocha et. al (2018).

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". José Carlos Soage was supported by Red Tecnológica de Matemática Industrial (Red TMATI), Cons. de Cultura, Educación e OU, Xunta de Galicia (ED341D R2016/051) and by Grupos de Referencia Competitiva, Consolidación y Estructuración de Unidades de Investigación Competitivas del SUG, Cons. de Cultura, Educación e OU, Xunta de Galicia (GRC ED431C 2016/040).

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). Testing equality of a large number of densities under mixing conditions. Preprint.
- Zhan, D., Hart, J. (2012). Testing equality of a large number of densities. Biometrika, 99, 1-17.

`Equalden.test.HD` *A test for the equality of a high dimensional set of densities*

Description

Performs the k-sample test proposed by Zhan and Hart (2012) for the low sample size, high dimensional setting with independent samples, and its extensions for dependent samples proposed by Cousido-Rocha et al. (2018).

Usage

```
Equalden.test.HD(X, method = c("indep", "dep.boot", "dep.spect"))
```

Arguments

`X`	A matrix where each row is one of the k-samples.
`method`	the k-sample test. By default the "dep.spect" method is computed. See details.

Details

The function implements the k-sample test proposed by Zhan and Hart (2012), method="indep", and its extensions for dependent data proposed by Cousido-Rocha et al. (2018), method="dep.boot" and "dep.spect". The method proposed by Zhan and Hart (2012) serves to test the null hypothesis that the k-samples have a common distribution. It is suitable when the k samples are independent and the number of samples k is large, and it works for sample sizes as small as 2. The statistic in Zhan and Hart (2012) is based on a comparison between the k sample-specific kernel density estimates and the kernel density estimate computed from the pooled sample. An alternative expression of this statistic shows that it can be interpreted as a difference between the intra-samples variability and the inter-samples variability. This statistic is standardized using a variance estimator which is valid for independent samples. The asymptotic normality (when k tends to infinity) of the standardized version of the statistic is used to compute the corresponding p-value. Cousido-Rocha et al. (2018) proposed two corrections of the test of Zhan and Hart (2012) for dependent samples. These tests standardize the statistic proposed in Zhan and Hart (2012) by using variance estimators which are suitable when the samples are weakly dependent. The method "dep.boot" implements the dependent multiplier bootstrap to estimate the variance, whereas the method "dep.spect" uses a variance estimator based on the spectral analysis theory. Both tests perform similarly, but the "dep.spect" test tends to be computationally more efficient than the "dep.boot" test. Cousido- Rocha et al. (2018)

showed through simulations that, for independent samples, the tests "dep.boot" and "dep.spect" may be more powerful than the test in Zhan and Hart (2012) despite of being protected against possible dependences. On the other hand, the statistic can be written as a sum of k individual statistics, each of them measures the difference between the intra-sample variability of the corresponding variable and the inter- samples variability. Whether the null hypothesis is rejected, an exploratory analysis of such individual statistics can help to guess which genes are not equally distributed.

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.

`m:`
: number of significant lags for the variance estimator if the method is "dep.spect" or "dep.boot". Null if the method is "indep" since no correction for dependence is required in this case.

`k:`
: number of samples or populations.

`n:`
: sample size.

`method:`
: a character string indicating what k-test was performed.

`I.statistics:`
: the k individual statistics.

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

Author(s)

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

References

- Cousido-Rocha, M., de Uña-Álvarez, J., and Hart, J.(2018). Testing equality of a large number of densities under mixing conditions. Preprint.
- Zhan, D., Hart, J. (2012). Testing equality of a large number of densities. Biometrika, 99, 1-17.

Examples

```
n <- 2
k <- 100
set.seed(1234)
X <- matrix(rnorm(n * k), ncol = 2)
```

```
res <- Equalden.test.HD(X,  method = "indep")

res
### The statistic and the variance estimator
res$statistic
res$variance
### The number of samples and sample size
res$k
res$n

### Real data analysis. We test the null hypothesis that 1000 randomly selected genes
### measured in patients with BRCA2 mutations have a common distribution. We use the test
### proposed in Cousido-Rocha et al. (2018) since correlation among expression levels of
### different genes on the same individual is expected.
data(Hedenfalk)
X <- Hedenfalk
k <- dim(X)[1]
### We eliminate the additive patients effects by substracting to each column its sample mean.
BRCA2 <- sweep(X[, 8:15], 2, apply(X[, 8:15], 2, mean))
set.seed (1234)
se<-1000
ind <- sample(1:k, se)
res1 <- Equalden.test.HD(BRCA2[ind, ], method = "dep.boot")
res1
res2 <- Equalden.test.HD(BRCA2[ind, ], method = "dep.spect")
res2
### The null hypothesis is rejected using both methods. Then we plot the individual statistics
### and highlight the 100 most extreme values.
cu <- 100
I.statistics.sorted <- sort(res1$I.statistics)
cv <- I.statistics.sorted[se-cu+1]
ind2 <- which(res1$I.statistics >= cv)
plot(1:se, res1$I.statistics, xlim = c(0, se), ylim = c(min(res1$I.statistics),
                                                        max(res1$I.statistics)),
     xlab = "Genes", ylab = "statistic", main = "Individual statistics")
points(ind2, res1$I.statistics[ind2], col = "red")
### We zoom the plot in the following way since some individual statistics report extreme
### negative values in this data.
plot(1:se, res1$I.statistics, xlim = c(0, se), ylim = c(0, max(res1$I.statistics)),
     xlab = "Genes", ylab = "statistic", main = "Individual statistics")
points(ind2, res1$I.statistics[ind2], col = "red")
```

Hedenfalk *Hedenfalk data*

Description

These data come from the breast cancer gene expression study of Hedenfalk et al. (2001). The data set consists on 3226 gene expression levels measured on 7 patients with breast tumors having

BRCA1 mutations (columns 1-7) and on 8 patients with breast tumors having BRCA2 mutations (columns 8-15). The rows correspond to the genes and the columns refer to the patients.

Usage

```
data(Hedenfalk)
```

Format

A matrix with 3226 rows corresponding to the measured genes and 15 columns corresponding to the patients. The first 7 columns contain the measures for the patients with BRCA1 mutations and the patients with BRCA2 mutations are located in the last 8 columns.

References

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M, Bitter, M., Simon, R., Meltzer, P., Guterson, B., Esteller, M., Kallioniemi, O., Wilfond, B., Borg, A., Trent, J., Raffeld, M., Yakhini, Z., BenDor, A., Dougherty, E., Kononen, J., Buberdorf, L., Fehrle, W., Pittaluga, S., Gruvberger, G., Loman, N., Johannsson, O., Olsson, H., and Sauter, G.(2001), Gene-Expression Profiles in Hereditary Breast Cancer. New England Journal of Medicine 344 (8), 539-548.

Examples

```
data(Hedenfalk)
X <- Hedenfalk
k <- dim(X)[1]
### Estimated densities of logged gene expression levels for patients with BRCA1 mutations.
s <- apply(X[, 1:7], 1, density)
### Plot of estimated densities of 6 randomly selected genes.
set.seed(375)
rs <- sample(1:k, 6)
plot(s[[rs[1]]], main = "Kernel estimates for 6 randomly selected genes",
 xlab = "x", ylab = "density", xlim = c(-1, 4.3), ylim = c(0, 2))
for (i in 2:6){
  lines(s[[rs[i]]], col = i)
}
```

Rat *Rat data*

Description

A microarray data set with 8038 logged gene expression levels measured on 5 rats. All rats were subjected to the same treatment. The rows correspond to the genes and the columns refer to the rats.

Usage

```
data(Rat)
```

Format

A matrix with 8038 rows corresponding to the measured genes and 5 columns corresponding to the rats.

References

Davidson, L.A., Nguyen, D.V., Hokanson, R.M., Callaway, E.S., Isett, R.B., Turner, N.D., Dougherty, E.R., Wang, N., Lupton, J.R., Carroll, R.J., and Chapkin, R.S. (2004). Chemopreventive n-3 polyunsaturated fatty acids reprogram genetic signatures during colon cancer initiation and progression in the rat. Cancer Research, 64, 6797-6804.

Examples

```
data(Rat)
X <- Rat
k <- dim(X)[1]
### Estimated densities of logged gene expression levels.
s <- apply(X, 1, density)
### Plot of estimated densities of 6 randomly selected genes.
set.seed(375)
rs <- sample(1:k, 6)
plot(s[[rs[1]]], main = "Kernel estimates for 6 randomly selected genes",
 xlab = "x", ylab = "density", xlim = c(-4, 2), ylim = c(0, 6))
for (i in 2:6){
  lines(s[[rs[i]]], col = i)
}
```

Index