Package 'kerTests'

August 22, 2023

Type Package

Title Generalized Kernel Two-Sample Tests

Version 0.1.4

Author Hoseung Song [aut, cre],
Hao Chen [aut]

Maintainer Hoseung Song <hosong@ucdavis.edu>

Description New kernel-based test and fast tests for testing whether two samples are from the same distribution. They work well particularly for high-dimensional data. Song, H. and Chen, H. (2023) <arXiv:2011.06127>.

License GPL (>= 2)

Encoding UTF-8

NeedsCompilation no

Repository CRAN

Date/Publication 2023-08-22 20:40:02 UTC

R topics documented:

`kerTests` *Generalized Kernel Two-Sample Tests*

Description

This package can be used to determine whether two samples are from the same distribution. The Gaussian kernel with the median heuristic, which is the median of all pairwise distances among observations, is used. To obtain the median heuristic, the function `med_sigma` should be used. The main function is `kertests`

Author(s)

Hoseung Song and Hao Chen

Maintainer: Hoseung Song (hosong@ucdavis.edu)

References

Song, Hoseung, and Hao Chen (2020). Generalized kernel two-sample tests. arXiv:2011.06127

See Also

kertests, med_sigma

Examples

```
## Mean difference in Gaussian distribution.
d = 100
mu = 0.2
sam = 100
n = 200
set.seed(500)
X = matrix(rnorm(d*sam), sam)
Y = matrix(rnorm(d*sam,mu), sam)
sigma = med_sigma(X, Y) # median heuristic
a = kertests(X, Y, sigma, r1=1.2, r2=0.8, perm=1000)
# output results based on the permutation and the asymptotic results
# the test statistic values can be found in a$teststat
# p-values can be found in a$pval
```

---

kertests *Generalized Kernel Two-Sample Tests*

Description

This function provides generalzied kernel-based two-sample tests.

Usage

```
kertests(X, Y, sigma, r1=1.2, r2=0.8, perm=0)
```

Arguments

X	The first samples.
Y	The second samples.
sigma	The bandwidth of Gaussian kernels. The median heuristic should be used.
r1	The constant in the test statistics $Z_{W,r1}$.

`r2` The constant in the test statistics $Z_{W,r2}$.

`perm` The number of permutations performed to calculate the p-value of the test. The default value is 0, which means the permutation is not performed and only approximated p-value based on the asymptotic theory is provided. Doing permutation could be time consuming, so be cautious if you want to set this value to be larger than 10,000.

Value

Returns a list `teststat` with each test statistic value and a list `pval` with p-values of the tests. See below for more details.

`GPK` The value of the test statistic GPK

`ZW1` The value of the test statistic $Z_{W,r1}$.

`ZW2` The value of the test statistic $Z_{W,r2}$.

`ZD` The value of the test statistic Z_D.

`fGPK_appr` The approximated p-value of fGPK based on asymptotic theory.

`fGPKM_appr` The approximated p-value of $fGPK_M$ based on asymptotic theory.

`fGPK_Simes_appr`
The approximated p-value of fGPK based on asymptotic theory with a Simes procedure.

`fGPKM_Simes_appr`
The approximated p-value of $fGPK_M$ based on asymptotic theory with a Simes procedure.

`GPK_perm` The permutation p-value of GPK when argument 'perm' is positive.

`fGPK_perm` The permutation p-value of fGPK when argument 'perm' is positive.

`fGPKM_perm` The permutation p-value of $fGPK_M$ when argument 'perm' is positive.

`fGPK_Simes_perm`
The permutation p-value of fGPK with a Simes procedure when argument 'perm' is positive.

`fGPKM_Simes_perm`
The permutation p-value of $fGPK_M$ with a Simes procedure when argument 'perm' is positive.

See Also

`kerTests-package`, `med_sigma`

Examples

```
## Mean difference in Gaussian distribution.
d = 100
mu = 0.2
sam = 100
n = 200
```

```
set.seed(500)
X = matrix(rnorm(d*sam), sam)
Y = matrix(rnorm(d*sam,mu), sam)

sigma = med_sigma(X, Y) # median heuristic

a = kertests(X, Y, sigma, r1=1.2, r2=0.8, perm=1000)
# output results based on the permutation and the asymptotic results
# the test statistic values can be found in a$teststat
# p-values can be found in a$pval
```

`med_sigma` *Compute the Median Heuristic*

Description

This function provides the most popular bandwidth of the Gaussian kernel, the median heuristic.

Usage

```
med_sigma(X, Y)
```

Arguments

`X`	The first samples.
`Y`	The second samples.

Value

Returns a numeric value, the median heuristic, which is the median of all pairwise distances among pooled observations, as a bandwidth of the kernel.

See Also

`kerTests-package`, `kertests`

Examples

```
## Mean difference in Gaussian distribution.
d = 100
mu = 0.2
sam = 100
n = 200
set.seed(500)
X = matrix(rnorm(d*sam), sam)
Y = matrix(rnorm(d*sam,mu), sam)

sigma = med_sigma(X, Y) # median heuristic (bandwidth)
```

Index