Package 'IndependenceTests'

February 19, 2015

Version 0.2

Date 2012-12-10

Title Nonparametric tests of independence between random vectors.

Author P Lafaye de Micheaux <lafaye@dms.umontreal.ca>, M Bilodeau <bilodeau@dms.umontreal.ca>

Maintainer P Lafaye de Micheaux <lafaye@dms.umontreal.ca>

Depends R (>= 2.3.0), xtable

Description Functions for testing mutual independence between many numerical random vectors or serial independence of a multivariate stationary sequence. The proposed test works when some or all of the marginal distributions are singular with respect to Lebesgue measure.

License GPL (>= 2)

Suggests snow, rsprng, Rmpi

Repository CRAN

Date/Publication 2012-12-11 16:46:04

NeedsCompilation yes

R topics documented:

A.dep.tests 2
dependogram 4
dna 6
highschool 7

Index **8**

`A.dep.tests` *Tests for mutual and serial independence of categorical variables*

Description

The tests are constructed from the Mobius transformation applied to the probability cells in a multi-way contingency table. The Pearson chi-squared test of mutual independence is partitioned into A-dependence statistics over all subsets A of variables. The goal of the partition is to identify subsets of dependent variables when the mutual independence hypothesis is rejected by the Pearson chi-squared test. The methodology can be directly adapted to test for serial independence of d successive observations of a stationary categorical time series.

For categorical time series, especially those of a nominal (non ordinal) nature, the user should be aware that tests of serial independence obtained by methods suited to quantitative sequences by quantification of the labels are not invariant to permutation of the labels contrary to the test described here.

Usage

```
A.dep.tests(Xmat,choice=1,d=0,m=d,freqname="",type="text")
```

Arguments

`Xmat`	Table or matrix of the contingency table data, if choice=1. Vector of the time series data, if choice=2.
`choice`	Integer. 1 for mutual, 2 for serial.
`d`	Integer. Used only if choice=2 for the number of successive obervations.
`m`	Integer. Maximum cardinality of subsets A for which an A-dependence statistic is required. This option is particularly useful for large values of d.
`freqname`	Character. Used only when Xmat is a matrix to indicate the name of the variable for the counts (frequencies).
`type`	"text" or "html"

Value

Returns an object of class list containing the following components:

`TA`	the A-dependence statistics for each subset A of variables.
`fA`	the degrees of freedom of the A-dependence statistics.
`pvalA`	the p-values of the A-dependence statistics.
`X`	summary of the results.
`X2`	test statistic for mutual independence obtained by the sum of the A-dependence statistics.
`Y2`	test statistic for serial independence obtained by the sum of the A-dependence statistics.
`f`	number of degrees of freedom associated with the test statistic X2 or Y2.
`pval`	the p-value associated with the test statistic X2 or Y2.

Author(s)

M. Bilodeau, P. Lafaye de Micheaux

References

Bilodeau, M., Lafaye de Micheaux, P (2009). A-dependence statistics for mutual and serial independence of categorical variables, Journal of Statistical Planning and Inference, 139, 2407-2419.

Agresti A. (2002). Categorical data analysis, Wiley, p. 322

Whisenant, E.C., Rasheed, B.K.A., Ostrer, H., Bhatnagar, Y.M. (1991). Evolution and sequence analysis of a human Y-chromosomal DNA fragment, J. Mol. Evol., 33, 133-141.

Examples

```
# Test of  mutual independence between 3 independent Bernoulli variables.

n <- 100
data <- data.frame(X1=rbinom(n,1,0.3),X2=rbinom(n,1,0.3),X3=rbinom(n,1,0.3))
X <- table(data)
A.dep.tests(X)

# Test of mutual independence between 4 variables which are
# 2-independent and 3-independent but 4-dependent.

n <- 100
W <- sample(x=1:8,size=n,TRUE)
X1 <- W %in% c(1,2,3,5)
X2 <- W %in% c(1,2,4,6)
X3 <- W %in% c(1,3,4,7)
X4 <- W %in% c(2,3,4,8)
data <- data.frame(X1,X2,X3,X4)
X <- table(data)
A.dep.tests(X)

# Test of serial independence of a nucleotide sequence of length
# 4156 described in Whisenant et al. (1991).

data(dna)
x2 <- dna[1]
for (i in 2:length(dna)) x2 <- paste(x2, dna[i], sep = "")
x <- unlist(strsplit(x2, ""))
x[x=="a"|x=="g"] <- "r"
x[x=="c"|x=="t"] <- "y"

out <- A.dep.tests(x,choice=2,d=1501,m=2)$TA[[1]]
plot(100:1500,out[100:1500],xlab="lag j",ylab="T(1,j+1)",pch=19)
abline(h=qchisq(.995,df=1))


# Analysis of a contingency table in Agresti (2002) p. 322

data(highschool)
```

```
A.dep.tests(highschool,freqname="count")
```

---

dependogram *Nonparametric tests of independence between random vectors*

Description

This function can be used for the following two problems: 1) testing mutual independence between many numerical random vectors, and 2) testing for serial independence of a multivariate stationary numerical time series. The proposed test does not assume continuous marginals. It is valid for any probability distribution. It is also invariant with respect to the affine general linear group of transformations on the vectors. This test is based on a characterization of mutual independence defined from probabilities of half-spaces in a combinatorial formula of Mobius. As such, it is a natural generalization of tests of independence between univariate random variables using the empirical distribution function. Without the assumption that each vector is one-dimensional with a continuous cumulative distribution function, any test of independence can not be distribution free. The critical values of the proposed test are thus computed with the bootstrap which was shown to be consistent in this context.

Usage

```
dependogram(X,vecd.ou.p,N=10,B=2000,alpha=0.05,display=TRUE,graphics=TRUE,nbclus=1)
```

Arguments

X	Data.frame or matrix with observations corresponding to rows and variables to columns.
vecd.ou.p	For the mutual independence problem 1), a vector giving the sizes of each sub-vector. For the serial independence problem 2), an integer indicating the number of consecutive observations.
N	Integer. Number of points of the discretization to obtain directions on the sphere in order to evaluate the value of the test statistic.
B	Integer. Number of bootstrap replications. Note that B can be slightly modified if nbclus>1
alpha	Double. Level of the test.
display	Logical. TRUE to display the values of the A-dependence statistic.
graphics	Logical. TRUE to plot the dependogram.
nbclus	Integer. Number of nodes in the cluster. Used only for parallel computations.

Value

A list with the following components:

In the mutual independence case:

`norm.RnA`	... should be completed ...
`Rn`	... should be completed ...
`rA`	... should be completed ...
`r`	... should be completed ...
`RnAsstar`	... should be completed ...

In the serial case:

`norm.SnA`	... should be completed ...
`Sn`	... should be completed ...
`sA`	... should be completed ...
`s`	... should be completed ...
`RnAsstar`	... should be completed ...

Author(s)

M. Bilodeau, P. Lafaye de Micheaux

Examples

```
n <- 100
W1 <- rpois(n,1)
W3 <- rpois(n,1)
W4 <- rpois(n,1)
W6 <- rpois(n,1)
W2 <- rpois(n,3)
W5 <- rpois(n,3)
X1 <- W1 + W2
X2 <- W2 + W3
X3 <- W4 + W5
X4 <- W5 + W6
X <- cbind(X1,X2,X3,X4)
dependogram(X,vecd.ou.p=c(1,1,1,1),N=10,B=20,alpha=0.05,display=TRUE,graphics=TRUE)
```

```
n <- 50
Sigma <- matrix(c(1, 0, 0, 0, 0, 0,
                  0, 1, 0, 0, 0, 0,
                  0, 0, 1, 0,.4,.5,
                  0, 0, 0, 1,.1,.2,
                  0, 0,.4,.1, 1, 0,
                  0, 0,.5,.2, 0, 1),nrow=6,ncol=6,byrow=TRUE)
W <- chol(Sigma)
X1 <- cbind(W[1,],W[2,])
X2 <- cbind(W[3,],W[4,])
X3 <- cbind(W[5,],W[6,])
X <- cbind(X1,X2,X3)

dependogram(X,vecd.ou.p=c(2,2,2),N=10,B=20,alpha=0.05,display=TRUE,graphics=TRUE)



n <- 100
W <- sample(x=1:8,size=n,TRUE)
X1 <- W
X2 <- W
X3 <- W
X4 <- W
X <- cbind(X1,X2,X3,X4)
dependogram(X,vecd.ou.p=c(1,1,1,1),N=10,B=20,alpha=0.05,display=TRUE,graphics=TRUE)


n <- 100
W <- rbinom(n,1,0.8)
Y <- W[1:(n-3)]*W[4:n]
dependogram(W,vecd.ou.p=4,N=10,B=20,alpha=0.05,display=TRUE,graphics=TRUE)
dependogram(Y,vecd.ou.p=4,N=10,B=20,alpha=0.05,display=TRUE,graphics=TRUE)


n <- 75
U <- matrix(rnorm(2*n),nrow=n,ncol=2)
W <- U[1:(n-1),] + sqrt(2)*U[2:n,]
Y <- W/apply(W,MARGIN=1,FUN=function(x) sqrt(x[1]^2+x[2]^2))

dependogram(Y,vecd.ou.p=3,N=10,B=20,alpha=0.05,display=TRUE,graphics=TRUE)
```

---

dna *dna sequence*

Description

This data from Whisenant et al. (1991) is a nucleotides sequence of 4156 base pairs (bp). The categorical variable represents the nucleotide which is either one of the two purines (r), adenine (a) or guanine (g), or one of the two pyrimidines (y), cytosine (c) or thymine (t).

Usage

```
data(dna)
```

Format

A character vector of length 70 representing 70 consecutive segments of a dna strands.

References

Whisenant, E.C., Rasheed, B.K.A., Ostrer, H., Bhatnagar, Y.M. (1991). Evolution and sequence analysis of a human Y-chromosomal DNA fragment, J. Mol. Evol., 33, 133-141.

Examples

```
data(dna)
```

---

highschool *Highschool data on alcohol, cigarette and marijuana use for high-school seniors*

Description

Data from a 1992 survey by the Wright State University School of Medecine and the United Health Services in Dayton, Ohio. The survey asked 2276 students in their final year of highschool in a nonurban area near Dayton, Ohio, whether they had ever used alcohol, cigarettes, or marijuana.

Usage

```
data(highschool)
```

Format

A data frame of 8 observations on the variables:

alcohol a factor vector with components "yes" or "no".

cigarette a factor vector with components "yes" or "no".

marijuana a factor vector with components "yes" or "no".

count a numeric vector of frequencies.

References

Agresti A. (2002). Categorical data analysis, Wiley, p. 322.

Examples

```
data(highschool)
```

Index

∗Topic **datasets**
dna, 6
highschool, 7

A.dep.tests, 2

dependogram, 4
dna, 6

highschool, 7