Package ‘gR2’

November 25, 2018

Type Package

Title Generalized R Square Measures for a Mixture of Bivariate Linear Dependences

Version 0.1.0

Maintainer Jingyi Jessica Li <jli@stat.ucla.edu>

Description This R package contains a function to compute the supervised and unsupervised sample generalized R square measures. The function also implements the K-lines clustering algorithm and allows an automatic choice of K for the unsupervised case. Statistical inference of the supervised and unsupervised population generalied R square measures is also included.

License GPL-2

Encoding UTF-8

LazyData true

Imports lmodel2,
mvtnorm,
parallel

RoxygenNote 6.1.1

R topics documented:

---

gR2 *gR2.*

Description

gR2 is the main function that calculates the sample generalized R square measures, i.e., point estimates, for the supervised scenario and the unsupervised scenario with or without K specified. It also performs statistical inference for the population measures, i.e., parameters of interest.

Usage

```
gR2(x, y, z = NULL, K = NULL, cand.Ks = 1:4, inference = FALSE,
  conf.level = 0.95, method = c("general", "binorm"), nstart = 30,
  mc.cores = detectCores() - 1)
```

Arguments

`x` a numeric vector.

`y` a numeric vector of the same length as `x`.

`z` an optional numeric vector containing integer values that indicate line memberships under the supervised scenario. The length of `z` must be the same as that of `x` and `y`.

`K` an optional number of lines under the unsupervised scenario, when `z=NULL`.

`cand.Ks` `1:4` (default) or a numeric vector containing the candidate values of `K`, when `K=NULL` under the unsupervised scenario (`z=NULL`).

`inference` logical. Should statistical inference be performed? Default is `FALSE`. If `TRUE`, a confidence interval of level `conf.level` will be computed for the population generalized R square measure, and a p-value will be computed for a one-sided test (greater than) against the null hypothesis that the population generalized R square measure is equal to zero.

`conf.level` the confidence level for the returned confidence interval. Default is `0.95`.

`method` a character string indicating which asymptotic distribution of the sample generalized measure is to be used for the inference. It must be one of `"general"` (the general asymptotic distribution) or `"binorm"` (the asymptotic distribution under the binormal distribution). The default is `"general"`.

`nstart` the number of initial starts for the K-lines algorithm under the unsupervised scenario (`z=NULL`).

`mc.cores` the number of cores to use, i.e. at most how many child processes will be run simultaneously. Must be at least one, and parallelization requires at least two cores. The default is the number of CPU cores minus one.

Details

`x`, `y` and `z` (if not `NULL`) must be numeric vectors of the same length. If `z` is not `NULL`, the supervised scenario is considered; otherwise, the unsupervised scenario is considered.

Under the supervised scenario, `K` will not be used.

Under the unsupervised scenario, if `is.null(K)` is `FALSE`, `K` must be a positive integer. If users are interested in the existence of a mixture of two linear relationships, set `K=2`. We do not recommend specifying `K` larger than `4` for interpretability consideration. If `is.null(K)` is `TRUE`, the function will automatically choose a `K` value from candidate values (default set to `{1,2,3,4}`) using the Akaike information criterion (AIC). The function will output two plots: (1) a scree plot showing how the within-cluster average squared perpendicular distances vary with `K` values, and (2) a plot showing how AIC changes with `K`. Users can decide whether the chosen `K` value is reasonable by checking these two plots.

If `inference` is `FALSE` (default), the function only outputs a point estimate, i.e., a sample generalized R square measure. Under the unsupervised scenario, the inferred line memberships by the K-lines clustering will also be output. If `inference` is `TRUE`, the function will additionally output a `conf.level`-level confidence interval of the population generalized R square measure, as well as a p-value for a one-sided test against the null hypothesis that the population generalized R square is equal to zero.

Value

`gR2` returns a list with the following components:

estimate — The sample generalized R square measure.

conf.level — The confidence level for the returned confidence interval (if inference is TRUE).

conf.int — A numeric vector with two elements indicating the lower and upper bounds of the confidence interval (if inference is TRUE).

p.val — The p-value for testing against the null hypothesis that the population generalized R square measure is zero (if inference is TRUE).

K — The number of line relationships (if z is NULL).

membership — The inferred line memberships of the data points in x and y (if z is NULL).

Author(s)

Jingyi Jessica Li, <jli@stat.ucla.edu>

References

Li, J.J., Tong, X., and Bickel, P.J. (2018). Generalized R2 Measures for a Mixture of Bivariate Linear Dependences. arXiv.

Examples

```
# generate data from a bivariate normal mixture model
library(mvtnorm)
library(parallel)
n = 200 # sample size
K = 2 # number of components (lines)
p_s = c(0.5, 0.5) # proportions of components
mu_s = list(c(0,-2), c(0,2)) # mean vectors
Sigma_s = list(rbind(c(1,0.8),c(0.8,1)), rbind(c(1,0.8),c(0.8,1))) # covariance matrices
z = sample(1:K, size=n, prob=p_s, replace=TRUE) # line memberships
data = matrix(0, nrow=n, ncol=2)
for (i in 1:K) {
  idx = which(z==i)
  data[idx,] = rmvnorm(n=length(idx), mean=mu_s[[i]], sigma=Sigma_s[[i]])
}
x = data[,1]
y = data[,2]

# supervised sample generalized R square
gR2(x, y, z) # without inference
gR2(x, y, z, inference=TRUE) # with inference

# unsupervised sample generalized R square
#gR2(x, y, K=2, mc.cores=2) # with K specified, without inference
#gR2(x, y, inference=TRUE, mc.cores=2) # without K specified, with inference
```

Index