Package ‘changedetection’

June 17, 2019

Type Package

Title Nonparametric Change Detection in Multivariate Linear Relationships

Version 0.2.0

Date 2019-06-17

Author Olga Gorskikh, Pekka Malo, Pauliina Ilmonen, Lauri Viitasaari, Joni Virta

Maintainer Olga Gorskikh <olga.a.gorskikh@gmail.com>

Description Contains implementation of the Nonparametric Splitting Algorithm (NSA), which estimates a set of structural change points (change dates) within a multivariate time-wise linear regression. Additionally, it contains utility functions to estimate corresponding changing linear model, moving energy distance and a change-detection test. For more information, see Malo et. al (2019) <arXiv:1805.08512v2>.

Encoding UTF-8

License GPL (>= 2)

LazyData true

RoxygenNote 6.1.1

Depends R (>= 2.10)

Imports Rdpack, L1pack, glmnet

RdMacros Rdpack

NeedsCompilation no

Repository CRAN

Date/Publication 2019-06-17 10:00:02 UTC

R topics documented:

`changes` *A list of change locations occured in the structure having given responses ‘y‘ and regressors ‘x‘*

Description

The main function of the package which estimates a set of structural change points for a dataset following multivariate (or univariate) linear model.

We consider the following problem. Assume we have observations $y_t \in R^q$, $x_t \in R^p$ over time $t = 1, ..., N$ which follow linear model

$$y_t = x_t'\beta(t) + \epsilon_t$$

where $\epsilon_t \in R^d$ stands for the model noise and β is a $p \times d$-dimensional piecewise-constant function, i.e. $\beta(t) \in R^{p \times d}$ is a weight matrix for every t. A change point is defined as a time instant $\hat{t}$ where β shifts. More precisely, a set of points separating sequential regimes is a set of change points of the model and the objective of `detectChanges` function is to find this set.

The approach is based on the splitting procedure NSA (Malo et al. 2019) and energy distance analysis (Rizzo and Szekely 2010).

Usage

```
changes(x, y, tau = NULL, l = NULL, R = 1000, pzero = 0.05,
  gamma = 0.5, alpha = 1)
```

Arguments

`x` matrix of regressors with variables in columns and observations in rows

`y` matrix of responses with variables in columns and observations in rows

`tau` length of splitting periods (default: l*10, which is dictated by The general rule of thumb (Harrell 2010))

`l` approximate number of contributing variables (default: overall number of regressors)

`R` number of bootstrap rounds (default: ‘1000‘)

`pzero` trust level for bootstrap (default: ‘0.05‘)

`gamma` tau reduction rate used in NSA (default: ‘0.5‘)

`alpha` parameter for energy distance formula (default: ‘1‘)

Value

A set of change locations indexes `changes`.

References

Harrell F (2010). “Regression Modelling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis.” *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*.

Malo P, Viitasaari L, Gorskikh O, Ilmonen P (2019). “Non-parametric Structural Change Detection in Multivariate Systems.” *arXiv:1805.08512v2*.

Rizzo M, Szekely GJ (2010). “Disco Analysis: A Nonparametric Extension of Analysis of Variance.” *The Annals of Applied Statistics*, **4**(2), 1034-1055.

Examples

```
T<-60
change<-35
x<-rnorm(n=T, m=0, sd=1)
e<-scale(rt(n=T,3), scale=FALSE)
y1<-5*x[1:(change-1)]+e[1:(change-1)]
y2<--2*x[change:T]+e[change:T]
y<-c(y1,y2)

changes(x=as.data.frame(x),
       y=as.data.frame(y),
       tau=20, R=100)
```

`changeTest` *A test showing whether two datasets have similar linear structure*

Description

The function performs a nonparametric test showing whether two datasets have similar linear structure or not. The test is based on applying energy distance (Rizzo and Szekely 2010) to residuals estimated for each dataset separately but only by one model (either first or second). It is implemented as a permutation test with `R` rounds and corresponding `pzero` (Malo et al. 2019).

Usage

```
changeTest(x1, y1, x2, y2, l = NULL, R = 1000, pzero = 0.05,
  alpha = 1)
```

Arguments

`x1` matrix of first period regressors with variables in columns and observations in rows

`y1` matrix of first period responses with variables in columns and observations in rows

`x2`	matrix of second period regressors with variables in columns and observations in rows
`y2`	matrix of second period responses with variables in columns and observations in rows
`l`	approximate number of contributing variables (default: overall number of regressors)
`R`	number of bootstrap rounds (default: ‘1000‘)
`pzero`	trust level for bootstrap (default: ‘0.05‘)
`alpha`	parameter for energy distance formula (default: ‘1‘)

Value

`TRUE` or `FALSE`

References

Malo P, Viitasaari L, Gorskikh O, Ilmonen P (2019). “Non-parametric Structural Change Detection in Multivariate Systems.” *arXiv:1805.08512v2*.

Rizzo M, Szekely GJ (2010). “Disco Analysis: A Nonparametric Extension of Analysis of Variance.” *The Annals of Applied Statistics*, **4**(2), 1034-1055.

Examples

```
T<-60
change<-35
x<-rnorm(n=T, m=0, sd=1)
e<-scale(rt(n=T,3), scale=FALSE)
y1<-5*x[1:(change-1)]+e[1:(change-1)]
y2<--2*x[change:T]+e[change:T]
y<-c(y1,y2)

testResult <- changeTest(x1=as.data.frame(x[1:T]),
                         y1=as.data.frame(y[1:T]),
                         x2=as.data.frame(x[31:T]),
                         y2=as.data.frame(y[31:T]),
                         R=200)
```

changingModel *Changeable linear structure of the data with given change points*

Description

The function estimates a set of linear models within the given dataset splitted by the given change points. The models are calculated as L1 regression based on a set of valuable predictors selected by lasso estimator.

Usage

```
changingModel(x, y, changes, l = NULL)
```

Arguments

`x`	matrix of regressors with variables in columns and observations in rows
`y`	matrix of responses with variables in columns and observations in rows
`changes`	a set of structural change points
`l`	approximate number of contributing variables (default : overall number of regressors)

Value

a set of linear models along with corresponding contributing variables indexes

Examples

```
T<-60
change<-35
x<-rnorm(n=T, m=0, sd=1)
e<-scale(rt(n=T,3), scale=FALSE)
y1<-5*x[1:(change-1)]+e[1:(change-1)]
y2<--2*x[change:T]+e[change:T]
y<-c(y1,y2)

model <- changingModel(x=as.data.frame(x),
                       y=as.data.frame(y),
                       c(change))
```

`energyDistance` *Energy distance between two datasets*

Description

Estimate energy distance which is a metric that measures the distance between the distributions of random vectors (Rizzo and Szekely 2010). The energy distance is zero if an only if the distributions are identical, otherwise it will diverge.

Usage

```
energyDistance(x1, y1, x2, y2, l = NULL, alpha = 1)
```

Arguments

`x1` matrix of first period regressors with variables in columns and observations in rows

`y1` matrix of first period responses with variables in columns and observations in rows

`x2` matrix of second period regressors with variables in columns and observations in rows

`y2` matrix of second period responses with variables in columns and observations in rows

`l` approximate number of contributing variables (default: overall number of regressors)

`alpha` parameter for energy distance formula (default: ‘1‘)

Value

energy distance value

References

Rizzo M, Szekely GJ (2010). “Disco Analysis: A Nonparametric Extension of Analysis of Variance.” *The Annals of Applied Statistics*, **4**(2), 1034-1055.

Examples

```
T<-60
change<-35
x<-rnorm(n=T, m=0, sd=1)
e<-scale(rt(n=T,3), scale=FALSE)
y1<-5*x[1:(change-1)]+e[1:(change-1)]
y2<--2*x[change:T]+e[change:T]
y<-c(y1,y2)

ed <- energyDistance(x1=as.data.frame(x[1:30]),
                     y1=as.data.frame(y[1:30]),
                     x2=as.data.frame(x[31:T]),
                     y2=as.data.frame(y[31:T]))
```

`movingEnergyDistance` *Moving energy distance*

Description

Estimates energy distance (Rizzo and Szekely 2010) for each point starting from `tau+1` to `T-tau`, where `T` is a data length. In these terms, energy distance for a point means energy distance between the dataset containing `tau` observations to the left and the dataset containing `tau` observations to the right of the original point. Hence, we are considering a so-called ’moving frame’ of length `2tau`. The resulting array shows how the energy distance behaves along the period to analyze.

Usage

```
movingEnergyDistance(x, y, l = NULL, tau = NULL, alpha = 1)
```

Arguments

`x` matrix of regressors with variables in columns and observations in rows

`y` matrix of responses with variables in columns and observations in rows

`l` approximate number of contributing variables (Default : overall number of regressors)

`tau` length of a splitting period (Default: l*10, which is dictated by The general rule of thumb (Harrell 2010))

`alpha` parameter for energy distance formula (default: ‘1‘)

Value

a list of energy distnce values for pairs of adjacent data segments of length tau (moving-frame construction)

References

Harrell F (2010). “Regression Modelling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis.” *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*.

Rizzo M, Szekely GJ (2010). “Disco Analysis: A Nonparametric Extension of Analysis of Variance.” *The Annals of Applied Statistics*, **4**(2), 1034-1055.

Examples

```
T<-60
change<-35
x<-rnorm(n=T, m=0, sd=1)
e<-scale(rt(n=T,3), scale=FALSE)
y1<-5*x[1:(change-1)]+e[1:(change-1)]
y2<--2*x[change:T]+e[change:T]
y<-c(y1,y2)

med <- movingEnergyDistance(x=as.data.frame(x),
                            y=as.data.frame(y))
```

`sampledataset` *Sample dataset*

Description

This is artificial data following changing multivariate liner structure.

Usage

```
sampledataset
```

Format

A data frame with 600 rows and 13 columns:

- y_1, y_2, y_3 response variables
- $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_10$ predictors

Index