Package 'funtimes'

August 14, 2021

Type Package

Title Functions for Time Series Analysis

Version 8.1

Date 2021-08-15

Depends R (>= 3.0.0)

License GPL (>= 2)

Imports Jmisc, Kendall, Rdpack, TDA, FNN, dbscan, igraph

Suggests covid19us, Ecdat, ggplot2, knitr, patchwork, randomcoloR, rmarkdown

Description Includes nonparametric estimators and tests for time series analysis. The functions are to test for presence of possibly non-monotonic trends and for synchronism of trends in multiple time series, using bootstrap techniques and robust nonparametric difference-based estimators.

RdMacros Rdpack

RoxygenNote 7.1.1

Encoding UTF-8

VignetteBuilder knitr, rmarkdown

NeedsCompilation no

Author Vyacheslav Lyubchich [aut, cre] (<https://orcid.org/0000-0001-7936-4285>), Yulia R. Gel [aut], Alexander Brenning [ctb], Calvin Chu [ctb], Xin Huang [ctb], Umar Islambekov [ctb], Palina Niamkova [ctb], Dorcas Ofori-Boateng [ctb], Ethan D. Schaeffer [ctb], Srishti Vishwakarma [ctb], Xingyu Wang [ctb]

Maintainer Vyacheslav Lyubchich <lyubchich@umces.edu>

Repository CRAN

Date/Publication 2021-08-14 14:20:02 UTC

R topics documented:

`funtimes-package` *funtimes: Functions for Time Series Analysis*

Description

Advances in multiple aspects of time-series analysis are documented in this package. See available vignettes using
`browseVignettes(package = "funtimes")`

Tests for trends applicable to autocorrelated data, see
`vignette("trendtests",package = "funtimes")`
include bootstrapped versions of t-test and Mann–Kendall test (Noguchi et al. 2011) and bootstrapped version of WAVK test for possibly non-monotonic trends (Lyubchich et al. 2013). The WAVK test is further applied in testing synchronism of trends (Lyubchich and Gel 2016); see an implementation to climate data in Lyubchich (2016). With iterative testing, the synchronism test is also applied for identifying clusters of multiple time series (Ghahari et al. 2017).

Additional clustering methods are implemented using functions `BICC` (Schaeffer et al. 2016) and `DR` (Huang et al. 2018); function `purity` can be used to assess accuracy of clustering if true classes are known.

Changepoint detection methods include modified CUSUM-based bootstrapped test (Lyubchich et al. 2020).

Additional functions include implementation of the Beale's ratio estimator, see `vignette("beales",package = "funtimes")`
Nonparametric comparison of tails of distributions is implemented using small bins defined based on quantiles (Soliman et al. 2015) or intervals in the units in which the data are recorded (Lyubchich and Gel 2017).

For a list of currently deprecated functions, use `?'funtimes-deprecated'`

For a list of defunct (removed) functions, use `?'funtimes-defunct'`

Author(s)

Maintainer: Vyacheslav Lyubchich <lyubchich@umces.edu> (ORCID)

Authors:

- Yulia R. Gel

Other contributors:

- Alexander Brenning [contributor]
- Calvin Chu [contributor]
- Xin Huang [contributor]
- Umar Islambekov [contributor]
- Palina Niamkova [contributor]
- Dorcas Ofori-Boateng [contributor]
- Ethan D. Schaeffer [contributor]
- Srishti Vishwakarma [contributor]
- Xingyu Wang [contributor]

References

Ghahari A, Gel YR, Lyubchich V, Chun Y, Uribe D (2017). "On employing multi-resolution weather data in crop insurance." In *Proceedings of the SIAM International Conference on Data Mining (SDM17) Workshop on Mining Big Data in Climate and Environment (MBDCE 2017)*.

Huang X, Iliev IR, Lyubchich V, Gel YR (2018). "Riding down the bay: space-time clustering of ecological trends." *Environmetrics*, **29**(5–6), e2455. doi: 10.1002/env.2455.

Lyubchich V (2016). "Detecting time series trends and their synchronization in climate data." *Intelligence. Innovations. Investments*, **12**, 132–137.

Lyubchich V, Gel YR (2016). "A local factor nonparametric test for trend synchronism in multiple time series." *Journal of Multivariate Analysis*, **150**, 91–104. doi: 10.1016/j.jmva.2016.05.004.

Lyubchich V, Gel YR (2017). "Can we weather proof our insurance?" *Environmetrics*, **28**(2), e2433. doi: 10.1002/env.2433.

Lyubchich V, Gel YR, El-Shaarawi A (2013). "On detecting non-monotonic trends in environmental time series: a fusion of local regression and bootstrap." *Environmetrics*, **24**(4), 209–226. doi: 10.1002/env.2212.

Lyubchich V, Lebedeva TV, Testa JM (2020). "A data-driven approach to detecting change points in linear regression models." *Environmetrics*, **31**(1), e2591. doi: 10.1002/env.2591.

Noguchi K, Gel YR, Duguay CR (2011). "Bootstrap-based tests for trends in hydrological time series, with application to ice phenology data." *Journal of Hydrology*, **410**(3), 150–161. doi: 10.1016/j.jhydrol.2011.09.008.

Schaeffer ED, Testa JM, Gel YR, Lyubchich V (2016). "On information criteria for dynamic spatio-temporal clustering." In Banerjee A, Ding W, Dy JG, Lyubchich V, Rhines A (eds.), *The 6th International Workshop on Climate Informatics: CI2016*, 5–8. doi: 10.5065/D6K072N6.

Soliman M, Lyubchich V, Gel YR, Naser D, Esterby S (2015). "Evaluating the impact of climate change on dynamics of house insurance claims." In Lakshmanan V, Gilleland E, McGovern A, Tingley M (eds.), *Machine Learning and Data Mining Approaches to Climate Science*, chapter 16, 175–183. Springer, Switzerland. doi: 10.1007/9783319172200_16.

`ARest` *Estimation of Autoregressive (AR) Parameters*

Description

Estimate parameters ϕ of autoregressive time series model

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + e_t,$$

by default using robust difference-based estimator and Bayesian information criterion (BIC) to select the order p. This function is employed for time series filtering in functions `sync_test` and `wavk_test`.

Usage

```
ARest(x, ar.order = NULL, ar.method = "HVK", BIC = TRUE)
```

Arguments

`x`	a vector containing a univariate time series. Missing values are not allowed.
`ar.order`	order of autoregressive model when `BIC = FALSE`, or the maximal order for BIC-based filtering. Default is `round(10*log10(length(x)))`, where `x` is the time series.

`ar.method` method of estimating autoregression coefficients. Default `"HVK"` delivers robust difference-based estimates by Hall and Van Keilegom (2003). Alternatively, options of `ar` function can be used, such as `"burg"`, `"ols"`, `"mle"`, and `"yw"`.

`BIC` logical value indicates whether the order of autoregressive filter should be selected by Bayesian information criterion (BIC). If `TRUE` (default), models of orders p = 0,1,...,`ar.order` or p = 0,1,...,`round(10*log10(length(x)))` are considered, depending on whether `ar.order` is defined or not (`x` is the time series).

Details

The same formula for BIC is used consistently for all methods:

$$BIC = n \ln(\hat{\sigma}^2) + k \ln(n),$$

where n = `length(x)`, $k = p + 1$.

Value

A vector of estimated AR coefficients. Returns `numeric(0)` if the final $p = 0$.

Author(s)

Vyacheslav Lyubchich

References

Hall P, Van Keilegom I (2003). "Using difference-based methods for inference in nonparametric regression with time series errors." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **65**(2), 443–456. doi: 10.1111/14679868.00395.

See Also

`ar`, `HVK`, `sync_test`, `wavk_test`

Examples

```
# Simulate a time series Y:
Y <- arima.sim(n = 200, list(order = c(2, 0, 0), ar = c(-0.7, -0.1)))
plot.ts(Y)

# Estimate the coefficients:
ARest(Y) # HVK, by default
ARest(Y, ar.method = "yw") # Yule--Walker
ARest(Y, ar.method = "burg") # Burg
```

AuePolyReg_test *Testing for Change Points in Time Series via Polynomial Regression*

Description

The function uses a nonlinear polynomial regression model in which it tests for the null hypothesis of structural stability in the regression parameters against the alternative of a break at an unknown time. The method is based on the extreme value distribution of a maximum-type test statistic which is asymptotically equivalent to the maximally selected likelihood ratio. The resulting testing approach is easily tractable and delivers accurate size and power of the test, even in small samples (Aue et al. 2008).

Usage

```
AuePolyReg_test(
  y,
  a.order,
  alpha = 0.05,
  crit.type = c("asymptotic", "bootstrap"),
  bootstrap.method = c("nonparametric", "parametric"),
  num.bootstrap = 1000
)
```

Arguments

y	a vector that contains univariate time series observations. Missing values are not allowed.
a.order	order of the autoregressive model which must be a nonnegative integer number.
alpha	significance level for testing hypothesis of no change point. Default value is 0.05.
crit.type	method of obtaining critical values: "asymptotic" (default) or "bootstrap".
bootstrap.method	type of bootstrap if crit.type = "bootstrap": "nonparametric" (default) or "parametric".
num.bootstrap	number of bootstrap replications if crit.type = "bootstrap". Default number is 1000.

Value

A list with the following components:

index	time point where the change point has occurred.
stat	test statistic.
crit.val	critical region value (CV(alpha, n)).
p.value	p-value of the change point test.

Author(s)

Palina Niamkova, Dorcas Ofori-Boateng, Yulia R. Gel

References

Aue A, Horvath L, Huskova M, Kokoszka P (2008). "Testing for changes in polynomial regression." *Bernoulli*, **14**(3), 637–660. doi: 10.3150/08BEJ122.

See Also

`mcusum.test` change point test for regression

Examples

```
## Not run:
#Example 1:

#Simulate some time series:
set.seed(23450)
series_1 = rnorm(137, 3, 5)
series_2 = rnorm(213, 0, 1)
series_val = c(series_1, series_2)
AuePolyReg_test(series_1, 1) # no change (asymptotic)
AuePolyReg_test(series_val,1) # one change (asymptotic)

#Example 2:

#Consider a time series with annual number of world terrorism incidents from 1970 till 2016:
c.data = Ecdat::terrorism["incidents"]
incidents.ts <- ts(c.data, start = 1970, end = 2016)

#Run a test for change points:
AuePolyReg_test(incidents.ts, 2) # one change (asymptotic)
AuePolyReg_test(incidents.ts, 2, 0.05,"bootstrap", "parametric", 200)
# one change (bootstrap)
incidents.ts[44] #number of victims at the value of change point
year <- 197 + 44 - 1  # year when the change point occurred
plot(incidents.ts) # see the visualized data

#The structural change point occurred at the 44th value which corresponds to 2013,
#with 11,990 identified incidents in that year. These findings can be explained with
#a recent rise of nationalism and  extremism due to appearance of the social media,
#Fisher (2019): White Terrorism Shows 'Stunning' Parallels to Islamic State's Rise.
#The New York Times.

## End(Not run)
```

`beales` *Beale's Estimator and Sample Size*

Description

Beale's ratio estimator (Beale 1962) for estimating population total and confidence intervals, with an option of calculating sample size for a required relative error (`p`) or margin of error (`d`).

Usage

```
beales(x, y, level = 0.95, N = NULL, p = NULL, d = NULL, verbose = TRUE)
```

Arguments

`x`	a numeric vector with quantities of interest, such as river discharge per month. Missing values (`NA`) are allowed.
`y`	a numeric vector with quantities of interest for which the total shall be estimated, such as total nutrient loads per month. Missing values (`NA`) are allowed. Lengths of `x` and `y` mush be the same.
`level`	confidence level, from 0 to 1. Default is `0.95`, that is, 95% confidence.
`N`	population size for which the estimate of the total `y` required. By default, `length(x)` is used.
`p`	optional argument specifying the required relative error, from 0 to 1, for computing corresponding sample size. For example, `p = 0.15` defines a 15% relative error.
`d`	optional argument specifying the required margin of error for computing corresponding sample size. If both `p` and `d` are specified, only `p` is used.
`verbose`	logical value defining whether the output should be printed out in words. Default is set to `TRUE` to give such output.

Value

A list with the following components:

`estimate`	Beale's estimate of the population total for the variable `y`.
`se`	standard error of the estimate.
`CI`	a vector of length 2 with a confidence interval (lower and upper value) for the estimate.
`level`	confidence level for the interval.
`N`	population size.
`n`	the actual sample size.
`p`	the relative error used for sample size calculations. Reported only if `p` was specified in the input.
`d`	the margin of error used for sample size calculations. Reported only if `d` was specified and `p` was not specified in the input.
`nhat`	estimated sample size for the given `level` and error (`p` or `d`).

Author(s)

Vyacheslav Lyubchich, thanks to Dave Lorenz for pointing out an error in version 7 and below of the package

References

Beale EML (1962). "Some uses of computers in operational research." *Industrielle Organisation*, **31**(1), 27–28.

See Also

`vignette("beales",package = "funtimes")`

Examples

```
#Some hypothetical data for monthly river discharge
#and corresponding nutrient loads:
discharge <- c(NA, 50, 90, 100, 80, 90, 100, 90, 80, 70, NA, NA)
loads <- c(33, 22, 44, 48, NA, 44, 49, NA, NA, 36, NA, NA)

#Example 1:
#Estimate total annual load (12 months),
#with 90% confidence intervals
beales(discharge, loads, level = 0.9)

#Example 2:
#Calculate sample size required for 90% confidence intervals
#with a margin of error 30 units
beales(discharge, loads, level = 0.9, d = 30)
```

BICC *BIC-Based Spatio-Temporal Clustering*

Description

Apply the algorithm of unsupervised spatio-temporal clustering, TRUST (Ciampi et al. 2010), with automatic selection of its tuning parameters `Delta` and `Epsilon` based on Bayesian information criterion, BIC (Schaeffer et al. 2016).

Usage

```
BICC(X, Alpha = NULL, Beta = NULL, Theta = 0.8, p, w, s)
```

Arguments

X	a matrix of time series observed within a slide (time series in columns).
Alpha	lower limit of the time series domain, passed to CSlideCluster.
Beta	upper limit of the time series domain, passed to CSlideCluster.
Theta	connectivity parameter, passed to CSlideCluster.
p	number of layers (time series observations) in each slide.
w	number of slides in each window.
s	step to shift a window, calculated in number of slides. The recommended values are 1 (overlapping windows) or equal to w (non-overlapping windows).

Details

This is the upper-level function for time series clustering. It exploits the functions CWindowCluster and CSlideCluster to cluster time series based on closeness and homogeneity measures. Clustering is performed multiple times with a range of equidistant values for the parameters Delta and Epsilon, then optimal parameters Delta and Epsilon along with the corresponding clustering results are shown (see Schaeffer et al. 2016, for more details).

The total length of time series (number of levels, i.e., nrow(X)) should be divisible by p.

Value

A list with the following elements:

delta.opt	optimal value for the clustering parameter Delta.
epsilon.opt	optimal value for the clustering parameter Epsilon.
clusters	vector of length ncol(X) with cluster labels.
IC	values of the information criterion (BIC) for each considered combination of Delta (rows) and Epsilon (columns).
delta.all	vector of considered values for Delta.
epsilon.all	vector of considered values for Epsilon.

Author(s)

Ethan Schaeffer, Vyacheslav Lyubchich

References

Ciampi A, Appice A, Malerba D (2010). "Discovering trend-based clusters in spatially distributed data streams." In *International Workshop of Mining Ubiquitous and Social Environments*, 107–122.

Schaeffer ED, Testa JM, Gel YR, Lyubchich V (2016). "On information criteria for dynamic spatio-temporal clustering." In Banerjee A, Ding W, Dy JG, Lyubchich V, Rhines A (eds.), *The 6th International Workshop on Climate Informatics: CI2016*, 5–8. doi: 10.5065/D6K072N6.

See Also

CSlideCluster, CWindowCluster, purity

Examples

```
# Fix seed for reproducible simulations:
set.seed(1)

##### Example 1
# Similar to Schaeffer et al. (2016), simulate 3 years of monthly data
#for 10 locations and apply clustering:
# 1.1 Simulation
T <- 36 #total months
N <- 10 #locations
phi <- c(0.5) #parameter of autoregression
burn <- 300 #burn-in period for simulations
X <- sapply(1:N, function(x)
    arima.sim(n = T + burn,
              list(order = c(length(phi), 0, 0), ar = phi)))[(burn + 1):(T + burn),]
colnames(X) <- paste("TS", c(1:dim(X)[2]), sep = "")

# 1.2 Clustering
# Assume that information arrives in year-long slides or data chunks
p <- 12 #number of time layers (months) in a slide
# Let the upper level of clustering (window) be the whole period of 3 years, so
w <- 3 #number of slides in a window
s <- w #step to shift a window, but it does not matter much here as we have only one window of data
tmp <- BICC(X, p = p, w = w, s = s)

# 1.3 Evaluate clustering
# In these simulations, it is known that all time series belong to one class,
#since they were all simulated the same way:
classes <- rep(1, 10)
# Use the information on the classes to calculate clustering purity:
purity(classes, tmp$clusters[1,])

##### Example 2
# 2.1 Modify time series and update classes accordingly:
# Add a mean shift to a half of the time series:
X2 <- X
X2[, 1:(N/2)] <- X2[, 1:(N/2)] + 3
classes2 <- rep(c(1, 2), each = N/2)

# 2.2 Re-apply clustering procedure and evaluate clustering purity:
tmp2 <- BICC(X2, p = p, w = w, s = s)
tmp2$clusters
purity(classes2, tmp2$clusters[1,])
```

`ccf_boot` *Cross-Correlation Function of Time Series with Sieve Bootstrap p-values*

Description

Account for possible autocorrelation of time series when assessing statistical significance of their cross-correlation. A sieve bootstrap approach is used to generate multiple copies of the time series with the same autoregressive dependence, under the null hypothesis of the two time series under investigation being uncorrelated. Significance of cross-correlation coefficients is assessed based on the distribution of their bootstrapped counterparts. Both Pearson and Spearman types of coefficients are obtained, but plot is provided for only one type, with significant correlations shown using filled circles.

Usage

```
ccf_boot(
  x,
  y,
  lag.max = NULL,
  plot = c("Pearson", "Spearman", "none"),
  level = 0.95,
  B = 1000,
  ...
)
```

Arguments

`x, y`	univariate numeric time series objects or numeric vectors for which to compute cross-correlation. Different time attributes in `ts` objects are acknowledged, see Example 2 below.
`lag.max`	maximum lag at which to calculate the cross-correlation. Will be automatically limited as in `ccf`.
`plot`	choose whether to plot results for Pearson correlation (default, or use `plot = "Pearson"`), Spearman correlation (use `plot = "Spearman"`), or suppress plotting (use `plot = "none"`). Both Pearson and Spearman results are given in the output, irregardless of the `plot` setting.
`level`	confidence level, from 0 to 1. Default is 0.95, that is, 95% confidence.
`B`	number of bootstrap simulations to obtain empirical critical values. Default is 1000.
`...`	other parameters passed to the function `ARest` to control how autoregressive dependencies are estimated. The same set of parameters is used separately on `x` and `y`.

Value

A data frame with the following columns:

`Lag`	lags for which the following values were obtained.
`rP`	observed Pearson correlations.
`pP`	bootstrap p-value for Pearson correlations.
`lowerP, upperP`	lower and upper quantiles (for the confidence level set by `level`) of the bootstrapped Pearson correlations.
`rS`	observed Spearman correlations.
`pS`	bootstrap p-value for Spearman correlations.
`lowerS, upperS`	lower and upper quantiles (for the confidence level set by `level`) of the bootstrapped Spearman correlations.

Author(s)

Vyacheslav Lyubchich

See Also

`ARest`, `ar`, `ccf`, `HVK`

Examples

```
## Not run:
# Fix seed for reproducible simulations:
set.seed(1)

# Example 1
# Simulate independent normal time series of same lengths
x <- rnorm(100)
y <- rnorm(100)
ccf(x, y) # default CCF with parametric confidence band
tmp <- ccf_boot(x, y) # CCF with bootstrap
tmp$rP; tmp$rS # can always extract results for both Pearson and Spearman correlations

# Example 2
# Simulated ts objects of different lengths and starts (incomplete overlap)
x <- arima.sim(list(order = c(1, 0, 0), ar = 0.5), n = 30)
x <- ts(x, start = 2001)
y <- arima.sim(list(order = c(2, 0, 0), ar = c(0.5, 0.2)), n = 40)
y <- ts(y, start = 2020)
ts.plot(x, y, col = 1:2, lty = 1:2) # show how x and y are aligned
ccf(x, y)
ccf_boot(x, y, plot = "Spearman") # CCF with bootstrap
# Notice that only +-7 lags can be calculated in both cases because of the small
# overlap of the time series. If save these time series as plain vectors, the time
# information would be lost, and time series will be misaligned.
ccf(as.numeric(x), as.numeric(y))
```

```
# Example 3
# Box & Jenkins time series of sales and a leading indicator, see ?BJsales
plot.ts(cbind(BJsales.lead, BJsales))
# Each of the BJ time series looks as having a stochastic linear trend, so apply differences:
plot.ts(cbind(diff(BJsales.lead), diff(BJsales)))
# Get cross-correlation of the differenced series:
ccf_boot(diff(BJsales.lead), diff(BJsales), plot = "Spearman")
# The leading indicator "stands out" with significant correlations at negative lags,
# showing it can be used to predict the sales 2-3 time steps ahead (that is,
# diff(BJsales.lead) at times t-2 and t-3 is strongly correlated with diff(BJsales) at
# current time t).

## End(Not run)
```

`CSlideCluster` *Slide-Level Time Series Clustering*

Description

Cluster time series at a slide level, based on Algorithm 1 of Ciampi et al. (2010).

Usage

```
CSlideCluster(X, Alpha = NULL, Beta = NULL, Delta = NULL, Theta = 0.8)
```

Arguments

`X`	a matrix of time series observed within a slide (time series in columns).
`Alpha`	lower limit of the time series domain. Default is `quantile(X)[2] - 1.5*(quantile(X)[4] -quantile(X)[2])`.
`Beta`	upper limit of the time series domain. Default is `quantile(X)[2] + 1.5*(quantile(X)[4] -quantile(X)[2])`.
`Delta`	closeness parameter, a real value in $[0, 1]$. Default is `0.1*(Beta -Alpha)`.
`Theta`	connectivity parameter, a real value in $[0, 1]$. Default is 0.8.

Value

A vector of length `ncol(X)` with cluster labels.

Author(s)

Vyacheslav Lyubchich

References

Ciampi A, Appice A, Malerba D (2010). "Discovering trend-based clusters in spatially distributed data streams." In *International Workshop of Mining Ubiquitous and Social Environments*, 107–122.

See Also

CSlideCluster, CWindowCluster, and BICC

Examples

```
set.seed(123)
X <- matrix(rnorm(50), 10, 5)
CSlideCluster(X)
```

`cumsumCPA_test` *Change Point Detection in Time Series via a Linear Regression with Temporally Correlated Errors*

Description

The function tests for a change point in parameters of a linear regression model with errors exhibiting a general weakly dependent structure. The approach extends earlier methods based on cumulative sums derived under assumption of independent errors. The approach applies smoothing when the time series is dominated by high frequencies. To detect multiple changes, it is recommended to employ a binary or wild segmentation (Gombay 2010).

Usage

```
cumsumCPA_test(
  y,
  a.order,
  crit.type = c("asymptotic", "bootstrap"),
  bootstrap.method = c("nonparametric", "parametric"),
  num.bootstrap = 1000
)
```

Arguments

`y`	a numeric time series vector. Missing values are not allowed.
`a.order`	order of the autoregressive model which must be a nonnegative integer number.
`crit.type`	a string parameter allowing to choose "asymptotic" or "bootstrap" options.
`bootstrap.method`	a string parameter allowing to choose "nonparametric" or "parametric" method of bootstrapping. "nonparametric" - resampling of the estimated residuals (with replacement); "parametric" - sampling innovations from a normal distribution.
`num.bootstrap`	number of bootstrap replications if `crit.type = "bootstrap"`. Default number is 1000.

Value

A list with the following components:

index	time point where the change has occurred.
stat	test statistic.
p.value	p-value of the change point test.

Author(s)

Palina Niamkova, Dorcas Ofori-Boateng, Yulia R. Gel

References

Gombay E (2010). "Change detection in linear regression with time series errors." *Canadian Journal of Statistics*, **38**(1), 65–79.

See Also

mcusum.test for change point test for regression

Examples

```
## Not run:
#Example 1:

#Simulate some time series:
series_1 = rnorm(157, 2, 1)
series_2 = rnorm(43, 7, 10)
main_val = c(series_1, series_2)

#Now perform a change point detection:
cumsumCPA_test(series_1, 1) # no change
cumsumCPA_test(main_val, 1) # one change, asymptotic critical region
cumsumCPA_test(main_val, 1, "bootstrap", "parametric") # one change, parametric bootstrap
cumsumCPA_test(main_val, 1, "bootstrap", "nonparametric") # one change, nonparametric
#bootstrap

#Example 2:

#Consider time series with ratio of real GDP per family to the median income. This is a
#skewness and income inequality measure for the US families from 1947 till 2012.
e.data = (Ecdat::incomeInequality['mean.median'])
incomeInequality.ts = ts(e.data, start = 1947, end = 2012, frequency = 1)

#Now perform a change point detection:
cumsumCPA_test(incomeInequality.ts, 0)
cumsumCPA_test(incomeInequality.ts, 0, "bootstrap", "parametric")
cumsumCPA_test(incomeInequality.ts, 0, "bootstrap", "nonparametric")
incomeInequality.ts[13] # median income
Ecdat::incomeInequality$Year[13] + 1 # year of change point
```

```
#The first change point occurs at the 13th time point, that is 1960, where the ratio of real
#GDP per family to the median income is 1.940126. This ratio shows that in 1960 the national
#wealth was not distributed equally between all the population and that most people earn
#almost twice less than the equal share of the all produced goods and services by the nation.

#Note: To look for the other possible change points, run the same function for the
#segment of time series after value 13.

## End(Not run)
```

`CWindowCluster` *Window-Level Time Series Clustering*

Description

Cluster time series at a window level, based on Algorithm 2 of Ciampi et al. (2010).

Usage

```
CWindowCluster(
  X,
  Alpha = NULL,
  Beta = NULL,
  Delta = NULL,
  Theta = 0.8,
  p,
  w,
  s,
  Epsilon = 1
)
```

Arguments

`X`	a matrix of time series observed within a slide (time series in columns).
`Alpha`	lower limit of the time series domain, passed to `CSlideCluster`.
`Beta`	upper limit of the time series domain, passed to `CSlideCluster`.
`Delta`	closeness parameter, passed to `CSlideCluster`.
`Theta`	connectivity parameter, passed to `CSlideCluster`.
`p`	number of layers (time series observations) in each slide.
`w`	number of slides in each window.
`s`	step to shift a window, calculated in number of slides. The recommended values are 1 (overlapping windows) or equal to `w` (non-overlapping windows).
`Epsilon`	a real value in $[0, 1]$ used to identify each pair of time series that are clustered together over at least `w*Epsilon` slides within a window; see Definition 7 by Ciampi et al. (2010). Default is 1.

Details

This is the upper-level function for time series clustering. It exploits the function CSlideCluster to cluster time series within each slide based on closeness and homogeneity measures. Then, it uses slide-level cluster assignments to cluster time series within each window.

The total length of time series (number of levels, i.e., nrow(X)) should be divisible by p.

Value

A vector (if X contains only one window) or matrix with cluster labels for each time series (columns) and window (rows).

Author(s)

Vyacheslav Lyubchich

References

Ciampi A, Appice A, Malerba D (2010). "Discovering trend-based clusters in spatially distributed data streams." In *International Workshop of Mining Ubiquitous and Social Environments*, 107–122.

See Also

CSlideCluster, CWindowCluster, and BICC

Examples

```
#For example, weekly data come in slides of 4 weeks
p <- 4 #number of layers in each slide (data come in a slide)

#We want to analyze the trend clusters within a window of 1 year
w <- 13 #number of slides in each window
s <- w  #step to shift a window

#Simulate 26 autoregressive time series with two years of weekly data (52*2 weeks),
#with a 'burn-in' period of 300.
N <- 26
T <- 2*p*w

set.seed(123)
phi <- c(0.5) #parameter of autoregression
X <- sapply(1:N, function(x) arima.sim(n = T + 300,
     list(order = c(length(phi), 0, 0), ar = phi)))[301:(T + 300),]
colnames(X) <- paste("TS", c(1:dim(X)[2]), sep = "")

tmp <- CWindowCluster(X, Delta = NULL, Theta = 0.8, p = p, w = w, s = s, Epsilon = 1)

#Time series were simulated with the same parameters, but based on the clustering parameters,
#not all time series join the same cluster. We can plot the main cluster for each window, and
#time series out of the cluster:
par(mfrow = c(2, 2))
ts.plot(X[c(1:(p*w)), tmp[1,] == 1], ylim = c(-4, 4),
```

```
        main = "Time series cluster 1 in window 1")
ts.plot(X[c(1:(p*w)), tmp[1,] != 1], ylim = c(-4, 4),
        main = "The rest of the time series in window 1")
ts.plot(X[c(1:(p*w)) + s*p, tmp[2,] == 1], ylim = c(-4, 4),
        main = "Time series cluster 1 in window 2")
ts.plot(X[c(1:(p*w)) + s*p, tmp[2,] != 1], ylim = c(-4, 4),
        main = "The rest of the time series in window 2")
```

DR *Downhill Riding (DR) Procedure*

Description

Downhill riding procedure for selecting optimal tuning parameters in clustering algorithms, using an (in)stability probe.

Usage

```
DR(X, method, minPts = 3, theta = 0.9, B = 500, lb = -30, ub = 10)
```

Arguments

X	a $n \times k$ matrix where columns are k objects to be clustered, and each object contains n observations (objects could be a set of time series).
method	the clustering method to be used – currently either "TRUST" (Ciampi et al. 2010) or "DBSCAN" (Ester et al. 1996). If method is DBSCAN, then set MinPts and optimal ϵ is selected using DR. If method is TRUST, then set theta and optimal δ is selected using DR.
minPts	the minimum number of samples in an ϵ-neighborhood of a point to be considered as a core point. The minPts is to be used only with DBSCAN method. Default value is 3.
theta	connectivity parameter $\theta \in (0, 1)$, which is to be used only with TRUST method. Default value is 0.9.
B	number of random splits in calculating the Average Cluster Deviation (ACD). Default value is 500.
lb, ub	end points for a range of search for the optimal parameter.

Details

Parameters lb,ub are end points for a range of search for the optimal parameter. The parameter candidates are calculated in a way such that $P := 1.1^x, x \in lb, lb + 0.5, lb + 1.0, ..., ub$. Although the default range of search is sufficiently wide, in some cases lb,ub can be further extended if a warning message is given.

For more discussion on properties of the considered clustering algorithms and the DR procedure see Huang et al. (2016) and Huang et al. (2018).

Value

A list containing the following components:

P_opt	the value of optimal parameter. If method is DBSCAN, then P_opt is optimal ϵ. If method is TRUST, then P_opt is optimal δ.
ACD_matrix	a matrix that returns ACD for different values of a tuning parameter. If method is DBSCAN, then the tuning parameter is ϵ. If method is TRUST, then the tuning parameter is δ.

Author(s)

Xin Huang, Yulia R. Gel

References

Ciampi A, Appice A, Malerba D (2010). "Discovering trend-based clusters in spatially distributed data streams." In *International Workshop of Mining Ubiquitous and Social Environments*, 107–122.

Ester M, Kriegel H, Sander J, Xu X (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise." In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD)*, volume 96(34), 226–231.

Huang X, Iliev IR, Brenning A, Gel YR (2016). "Space-time clustering with stability probe while riding downhill." In *Proceedings of the 2nd SIGKDD Workshop on Mining and Learning from Time Series (MiLeTS)*.

Huang X, Iliev IR, Lyubchich V, Gel YR (2018). "Riding down the bay: space-time clustering of ecological trends." *Environmetrics*, **29**(5–6), e2455. doi: 10.1002/env.2455.

See Also

BICC, dbscan

Examples

```
## Not run:
## example 1
## use iris data to test DR procedure

data(iris)
require(clue)  # calculate NMI to compare the clustering result with the ground truth
require(scatterplot3d)

Data <- scale(iris[,-5])
ground_truth_label <- iris[,5]

# perform DR procedure to select optimal eps for DBSCAN
# and save it in variable eps_opt
eps_opt <- DR(t(Data), method="DBSCAN", minPts = 5)$P_opt
```

```
# apply DBSCAN with the optimal eps on iris data
# and save the clustering result in variable res
res <- dbscan(Data, eps = eps_opt, minPts =5)$cluster

# calculate NMI to compare the clustering result with the ground truth label
clue::cl_agreement(as.cl_partition(ground_truth_label),
                   as.cl_partition(as.numeric(res)), method = "NMI")
# visualize the clustering result and compare it with the ground truth result
# 3D visualization of clustering result using variables Sepal.Width, Sepal.Length,
# and Petal.Length
scatterplot3d(Data[,-4],color = res)
# 3D visualization of ground truth result using variables Sepal.Width, Sepal.Length,
# and Petal.Length
scatterplot3d(Data[,-4],color = as.numeric(ground_truth_label))


## example 2
## use synthetic time series data to test DR procedure

require(funtimes)
require(clue)
require(zoo)

# simulate 16 time series for 4 clusters, each cluster contains 4 time series
set.seed(114)
samp_Ind <- sample(12,replace=F)
time_points <- 30
X <- matrix(0,nrow=time_points,ncol = 12)
cluster1 <- sapply(1:4,function(x) arima.sim(list(order = c(1, 0, 0), ar = c(0.2)),
                                             n = time_points, mean = 0, sd = 1))
cluster2 <- sapply(1:4,function(x) arima.sim(list(order = c(2 ,0, 0), ar = c(0.1, -0.2)),
                                             n = time_points, mean = 2, sd = 1))
cluster3 <- sapply(1:4,function(x) arima.sim(list(order = c(1, 0, 1), ar = c(0.3), ma = c(0.1)),
                                             n = time_points, mean = 6, sd = 1))

X[,samp_Ind[1:4]] <- t(round(cluster1, 4))
X[,samp_Ind[5:8]] <- t(round(cluster2, 4))
X[,samp_Ind[9:12]] <- t(round(cluster3, 4))


# create ground truth label of the synthetic data
ground_truth_label = matrix(1, nrow = 12, ncol = 1)
for(k in 1:3){
    ground_truth_label[samp_Ind[(4*k - 4 + 1):(4*k)]] = k
}

# perform DR procedure to select optimal delta for TRUST
# and save it in variable delta_opt
delta_opt <- DR(X, method = "TRUST")$P_opt

# apply TRUST with the optimal delta on the synthetic data
# and save the clustering result in variable res
res <- CSlideCluster(X, Delta = delta_opt, Theta = 0.9)
```

```
# calculate NMI to compare the clustering result with the ground truth label
clue::cl_agreement(as.cl_partition(as.numeric(ground_truth_label)),
                   as.cl_partition(as.numeric(res)), method = "NMI")

# visualize the clustering result and compare it with the ground truth result
# visualization of the clustering result obtained by TRUST
plot.zoo(X, type = "l", plot.type = "single", col = res, xlab = "Time index", ylab = "")
# visualization of the ground truth result
plot.zoo(X, type = "l", plot.type = "single", col = ground_truth_label,
         xlab = "Time index", ylab = "")

## End(Not run)
```

`GombayCPA_test` *Change Point Detection in Autoregressive Time Series*

Description

The function detects change points in autoregressive (AR) models for time series. Changes can be detected in any of p + 2 (mean, var, phi) autoregressive parameters where p is the order of the AR model. The test statistic is based on the efficient score vector (Gombay 2008).

Usage

```
GombayCPA_test(
  y,
  a.order,
  alternatives = c("two-sided", "greater", "lesser", "temporary"),
  crit.type = c("asymptotic", "bootstrap"),
  num.bootstrap = 1000
)
```

Arguments

`y`	a vector that contains univariate time series observations. Missing values are not allowed.
`a.order`	order of the autoregressive model which must be a nonnegative integer number.
`alternatives`	a string parameter that specifies a type of the test (i.e., "two-sided", "greater", "lesser", and "temporary"). The option "temporary" examines the temporary change in one of the parameters (Gombay 2008).
`crit.type`	method of obtaining critical values: "asymptotic" (default) or "bootstrap".
`num.bootstrap`	number of bootstrap replications if `crit.type = "bootstrap"`. Default number is 1000.

Details

The function allows for testing for a temporary change and for a change in a specific model parameters. Critical values can be estimated via asymptotic distribution `"asymptotic"` (i.e., the default option) or via sieve bootstrap `"bootstrap"`. The function employs internal function `change.point` and sieve bootstrap `change.point.sieve` function.

Value

A list with the following components:

`index`	points of change for each parameter. The value of the `"alternatives"` determines the return: "temporary" - returns max, min and abs.max points; "greater" - returns max points; "lesser" - returns min points; "two-sided" - returns abs.max.
`stats`	test statistic values for change points in: mean, var, phi.
`p.values`	p-value of the change point test.

Author(s)

Palina Niamkova, Dorcas Ofori-Boateng, Yulia R. Gel

References

Gombay E (2008). "Change detection in autoregressive time series." *Journal of Multivariate Analysis*, **99**(3), 451–464. doi: 10.1016/j.jmva.2007.01.003.

See Also

`mcusum.test` change point test for regression and `terrorism` dataset used in the Example 2

Examples

```
## Not run:
#Example 1:

#Simulate some time series:
series_1 = arima.sim(n = 100, list(order = c(2,0,0), ar = c(-0.7, -0.1)))
series_2 = arima.sim(n = 200, list(order = c(2,0,0), ar = c(0.1, -0.6)))
main_series = c(series_1, series_2)

result11 = GombayCPA_test(series_1, 2, "two-sided")
result11 #== No change point ===#

result12 = GombayCPA_test(main_series, 2, "two-sided")
result12  #=== One change at phi values ===#

result13 = GombayCPA_test(main_series, 2, "two-sided", "bootstrap")
result13  #=== One change at phi values ===#
```

```
#Example 2:

#From the package 'Ecdat' consider a time series with annual world number of victims of
#terrorism in the US from 1970 till 2016:
c.data = Ecdat::terrorism['nkill.us']
nkill.us.ts <- ts(c.data, start = 1970, end = 2016)

#Now perform a change point detection with one sided tests:
GombayCPA_test(nkill.us.ts, 0, "lesser")
GombayCPA_test(nkill.us.ts, 0, "greater")
nkill.us.ts[32]
year = 1970 + 31
print(year)
plot(nkill.us.ts)

#In both cases we find that the change point is located at the position 31 or 32. We can
# examine it further by checking the value of this position (using: nkill.us.ts[32]) as well as
# by plotting the graph (using: plot(nkill.us.ts)). The detected change point corresponds to
#the year of 2001, when the 9/11 attack happened.

## End(Not run)
```

HVK *HVK Estimator*

Description

Estimate coefficients in non-parametric autoregression using the difference-based approach by Hall and Van Keilegom (2003).

Usage

```
HVK(X, m1 = NULL, m2 = NULL, ar.order = 1)
```

Arguments

`X`	univariate time series. Missing values are not allowed.
`m1, m2`	subsidiary smoothing parameters. Default `m1 = round(length(X)^(0.1))`, `m2 = round(length(X)^(0.5))`.
`ar.order`	order of the non-parametric autoregression (specified by user).

Details

First, autocovariances are estimated using formula (2.6) by Hall and Van Keilegom (2003):

$$\hat{\gamma}(0) = \frac{1}{m_2 - m_1 + 1} \sum_{m=m_1}^{m_2} \frac{1}{2(n-m)} \sum_{i=m+1}^{n} \{(D_m X)_i\}^2,$$

$$\hat{\gamma}(j) = \hat{\gamma}(0) - \frac{1}{2(n-j)} \sum_{i=j+1}^{n} \{(D_j X)_i\}^2,$$

where n = `length(X)` is sample size, D_j is a difference operator such that $(D_j X)_i = X_i - X_{i-j}$. Then, Yule–Walker method is used to derive autoregression coefficients.

Value

Vector of length `ar.order` with estimated autoregression coefficients.

Author(s)

Yulia R. Gel, Vyacheslav Lyubchich, Xingyu Wang

References

Hall P, Van Keilegom I (2003). "Using difference-based methods for inference in nonparametric regression with time series errors." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **65**(2), 443–456. doi: 10.1111/14679868.00395.

See Also

`ar`, `ARest`

Examples

```
X <- arima.sim(n = 300, list(order = c(1, 0, 0), ar = c(0.6)))
HVK(as.vector(X), ar.order = 1)
```

---

i.tails *Interval-Based Tails Comparison*

Description

Compare right tails of two sample distributions using an interval-based approach (IBA); see Chu et al. (2015) and Lyubchich and Gel (2017).

Usage

```
i.tails(x0, x1, d = NULL)
```

Arguments

`x0, x1`	vectors of the same length (preferably). Tail in `x1` is compared against the tail in `x0`.
`d`	a threshold defining the tail. The threshold is the same for both `x0` and `x1`. Default is `quantile(x0,probs = 0.99)`.

Details

Sturges' formula is used to calculate number of intervals (k) for `x0` $\geq$ `d`, then interval width is derived. The tails, `x0` $\geq$ `d` and `x1` $\geq$ `d`, are divided into the intervals. Number of `x1`-values within each interval is compared with the number of `x0`-values within the same interval (this difference is reported as `Nk`).

Value

A list with two elements:

`Nk`	vector that tells how many more `x1`-values compared with `x0`-values there are within each interval.
`Ck`	vector of the intervals' centers.

Author(s)

Calvin Chu, Yulia R. Gel, Vyacheslav Lyubchich

References

Chu C, Gel YR, Lyubchich V (2015). "Climate change from an insurance perspective: a case study of Norway." In Dy JG, Emile-Geay J, Lakshmanan V, Liu Y (eds.), *The 5th International Workshop on Climate Informatics: CI2015*.

Lyubchich V, Gel YR (2017). "Can we weather proof our insurance?" *Environmetrics*, **28**(2), e2433. doi: 10.1002/env.2433.

See Also

`q.tails`

Examples

```
x0 <- rnorm(1000)
x1 <- rt(1000, 5)
i.tails(x0, x1)
```

`mcusum_test` *Change Point Test for Regression*

Description

Apply change point test by Horvath et al. (2017) for detecting at-most-m changes in regression coefficients, where test statistic is a modified cumulative sum (CUSUM), and critical values are obtained with sieve bootstrap (Lyubchich et al. 2020).

Usage

```
mcusum_test(
  e,
  k,
  m = length(k),
  B = 1000,
  shortboot = FALSE,
  ksm = FALSE,
  ksm.arg = list(kernel = "gaussian", bw = "sj"),
  ...
)
```

Arguments

e	vector of regression residuals (a stationary time series).
k	an integer vector or scalar with hypothesized change point location(s) to test.
m	an integer specifying the maximum number of change points being confirmed as statistically significant (from those specified in k) would be $\leq m$. Thus, m must be in 1,...,k.
B	number of bootstrap simulations to obtain empirical critical values. Default is 1000.
shortboot	if TRUE, then a heuristic is used to perform the test with a reduced number of bootstrap replicates. Specifically, B/4 replicates are used, which may reduce computing time by up to 75% when the number of retained null hypotheses is large. A p-value of 999 is reported whenever a null hypothesis is retained as a result of this mechanism.
ksm	logical value indicating whether a kernel smoothing to innovations in sieve bootstrap shall be applied (default is FALSE, that is, the original estimated innovations are bootstrapped, without the smoothing).
ksm.arg	used only if ksm = TRUE. A list of arguments for kernel smoothing to be passed to density function. Default settings specify the use of Gaussian kernel and the "sj" rule to choose the bandwidth.
...	additional arguments passed to ARest (for example, ar.method).

Details

The sieve bootstrap is applied by approximating regression residuals e with an AR(p) model using function ARest, where the autoregressive coefficients are estimated with ar.method, and order p is selected based on ar.order and BIC settings (see ARest). At the next step, B autoregressive processes are simulated under the null hypothesis of no change points. The distribution of test statistics M_T computed on each of those bootstrapped series is used to obtain bootstrap-based p-values for the test (Lyubchich et al. 2020).

In the current implementation, p-value is calculated using equation 4.10 of Davison and Hinkley (1997): p.value = (1 + n) / (B + 1), where n is number of bootstrapped statistics greater or equal to the observed statistic.

The test statistic corresponds to the maximal value of the modified CUSUM over all up to `m` combinations of hypothesized change points specified in `k`. The change points that correspond to that maximum are reported in `estimate$khat`, and their number is reported as `parameter`.

Value

A list of class `"htest"` containing the following components:

`method`	name of the method.
`data.name`	name of the data.
`statistic`	obseved value of the test statistic.
`parameter`	`mhat` is the final number of change points, from those specified in the input `k`, for which the test statistic is reported. See the corresponding locations, `khat`, in the `estimate`.
`p.value`	bootstrapped p-value of the test.
`alternative`	alternative hypothesis.
`estimate`	list with elements: `AR_order` and `AR_coefficients` (the autoregressive order and estimated autoregressive coefficients used in sieve bootstrap procedure), `khat` (final change points, from those specified in the input `k` for which the test statistic is reported), and `B` (the number of bootstrap replications).

Author(s)

Vyacheslav Lyubchich

References

Davison AC, Hinkley DV (1997). *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge.

Horvath L, Pouliot W, Wang S (2017). "Detecting at-most-m changes in linear regression models." *Journal of Time Series Analysis*, **38**, 552–590. doi: 10.1111/jtsa.12228.

Lyubchich V, Lebedeva TV, Testa JM (2020). "A data-driven approach to detecting change points in linear regression models." *Environmetrics*, **31**(1), e2591. doi: 10.1002/env.2591.

Examples

```
##### Model 1 with normal errors, by Horvath et al. (2017)
T <- 100 #length of time series
X <- rnorm(T, mean = 1, sd = 1)
E <- rnorm(T, mean = 0, sd = 1)
SizeOfChange <- 1
TimeOfChange <- 50
Y <- c(1 * X[1:TimeOfChange] + E[1:TimeOfChange],
      (1 + SizeOfChange)*X[(TimeOfChange + 1):T] + E[(TimeOfChange + 1):T])
ehat <- lm(Y ~ X)$resid
mcusum_test(ehat, k = c(30, 50, 70))
```

```
#Same, but with bootstrapped innovations obtained from a kernel smoothed distribution:
mcusum_test(ehat, k = c(30, 50, 70), ksm = TRUE)
```

`notrend_test` *Sieve Bootstrap Based Test for the Null Hypothesis of no Trend*

Description

A combination of time series trend tests for testing the null hypothesis of no trend, versus the alternative hypothesis of a linear trend (Student's t-test), or monotonic trend (Mann–Kendall test), or more general, possibly non-monotonic trend (WAVK test).

Usage

```
notrend_test(
  x,
  B = 1000,
  test = c("t", "MK", "WAVK"),
  ar.method = "HVK",
  ar.order = NULL,
  BIC = TRUE,
  factor.length = c("user.defined", "adaptive.selection"),
  Window = NULL,
  q = 3/4,
  j = c(8:11)
)
```

Arguments

`x`	a vector containing a univariate time series. Missing values are not allowed.
`B`	number of bootstrap simulations to obtain empirical critical values. Default is 1000.
`test`	trend test to implement: Student's t-test (`"t"`, default), Mann–Kendall test (`"MK"`), or WAVK test (`"WAVK"`, see `WAVK`).
`ar.method`	method of estimating autoregression coefficients. Default `"HVK"` delivers robust difference-based estimates by Hall and Van Keilegom (2003). Alternatively, options of `ar` function can be used, such as `"burg"`, `"ols"`, `"mle"`, and `"yw"`.
`ar.order`	order of autoregressive model when `BIC = FALSE`, or the maximal order for BIC-based filtering. Default is `round(10*log10(length(x)))`, where `x` is the time series.
`BIC`	logical value indicates whether the order of autoregressive filter should be selected by Bayesian information criterion (BIC). If `TRUE` (default), models of orders $p = 0,1,...,$`ar.order` or $p = 0,1,...,$`round(10*log10(length(x)))` are considered, depending on whether `ar.order` is defined or not (`x` is the time series).

`factor.length`	method to define the length of local windows (factors). Used only if `test = "WAVK"`. Default option `"user.defined"` allows to set only one value of the argument `Window`. The option `"adaptive.selection"` sets `method = "boot"` and employs heuristic m-out-of-n subsampling algorithm (Bickel and Sakov 2008) to select an optimal window from the set of possible windows `length(x)*q^j` whose values are mapped to the largest previous integer and greater than 2. Vector `x` is the time series tested.
`Window`	length of the local window (factor), default is `round(0.1*length(x))`. Used only if `test = "WAVK"`. This argument is ignored if `factor.length = "adaptive.selection"`.
`q`	scalar from 0 to 1 to define the set of possible windows when `factor.length = "adaptive.selection"`. Used only if `test = "WAVK"`. Default is $3/4$. This argument is ignored if `factor.length = "user.defined"`.
`j`	numeric vector to define the set of possible windows when `factor.length = "adaptive.selection"`. Used only if `test = "WAVK"`. Default is `c(8:11)`. This argument is ignored if `factor.length = "user.defined"`.

Details

This function tests the null hypothesis of no trend versus different alternatives. To set some other shape of trend as the null hypothesis, use `wavk_test`. Note that `wavk_test` employs hybrid bootstrap, which is alternative to the sieve bootstrap employed by the current function.

Value

A list with class `"htest"` containing the following components:

`method`	name of the method.
`data.name`	name of the data.
`statistic`	value of the test statistic.
`p.value`	p-value of the test.
`alternative`	alternative hypothesis.
`estimate`	list with the following elements: employed AR order and estimated AR coefficients.
`parameter`	window that was used in WAVK test, included in the output only if `test = "WAVK"`.

Author(s)

Vyacheslav Lyubchich

References

Bickel PJ, Sakov A (2008). "On the choice of m in the m out of n bootstrap and confidence bounds for extrema." *Statistica Sinica*, **18**(3), 967–985.

Hall P, Van Keilegom I (2003). "Using difference-based methods for inference in nonparametric regression with time series errors." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **65**(2), 443–456. doi: 10.1111/14679868.00395.

See Also

ar, HVK, WAVK, wavk_test, vignette("trendtests",package = "funtimes")

Examples

```
## Not run:
# Fix seed for reproducible simulations:
set.seed(1)

#Simulate autoregressive time series of length n with smooth linear trend:
n <- 200
tsTrend <- 1 + 2*(1:n/n)
tsNoise <- arima.sim(n = n, list(order = c(2, 0, 0), ar = c(0.5, -0.1)))
U <- tsTrend + tsNoise
plot.ts(U)

#Use t-test
notrend_test(U)

#Use Mann--Kendall test and Yule-Walker estimates of the AR parameters
notrend_test(U, test = "MK", ar.method = "yw")

#Use WAVK test for the H0 of no trend, with m-out-of-n selection of the local window:
notrend_test(U, test = "WAVK", factor.length = "adaptive.selection")
# Sample output:
## Sieve-bootstrap WAVK trend test
##
##data:  U
##WAVK test statistic = 21.654, moving window = 15, p-value < 2.2e-16
##alternative hypothesis: (non-)monotonic trend.
##sample estimates:
##$AR_order
##[1] 1
##
##$AR_coefficients
##    phi_1
##0.4041848

## End(Not run)
```

---

purity *Clustering Purity*

Description

Calculate purity of the clustering results. For example, see Schaeffer et al. (2016).

Usage

```
purity(classes, clusters)
```

Arguments

`classes`	a vector with labels of true classes.
`clusters`	a vector with labels of assigned clusters for which purity is to be tested. Should be of the same length as `classes`.

Details

Following Manning et al. (2008), each cluster is assigned to the class which is most frequent in the cluster, then

$$Purity(\Omega, C) = \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j|,$$

where $\Omega = \{\omega_1, \ldots, \omega_K\}$ is the set of identified clusters and $C = \{c_1, \ldots, c_J\}$ is the set of classes. That is, within each class $j = 1, \ldots, J$ find the size of the most populous cluster from the $K - j$ unassigned clusters. Then, sum together the $\min(K, J)$ sizes found and divide by N, where N = `length(classes)` = `length(clusters)`.

If $\max_j |\omega_k \cap c_j|$ is not unique for some j, it is assigned to the class which second maximum is the smallest, to maximize the $Purity$ (see 'Examples').

Number of unique elements in `classes` and `clusters` may differ.

Value

A list with two elements:

`pur`	purity value.
`out`	table with $\min(K, J)$ = `min(length(unique(classes)),length(unique(clusters)))` rows and the following columns: `ClassLabels`, `ClusterLabels`, and `ClusterSize`.

Author(s)

Vyacheslav Lyubchich

References

Manning CD, Raghavan P, Schutze H (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York.

Schaeffer ED, Testa JM, Gel YR, Lyubchich V (2016). "On information criteria for dynamic spatio-temporal clustering." In Banerjee A, Ding W, Dy JG, Lyubchich V, Rhines A (eds.), *The 6th International Workshop on Climate Informatics: CI2016*, 5–8. doi: 10.5065/D6K072N6.

Examples

```
# Fix seed for reproducible simulations:
# RNGkind(sample.kind = "Rounding") #run this line to have same seed across R versions > R 3.6.0
set.seed(1)

##### Example 1
#Create some classes and cluster labels:
classes <- rep(LETTERS[1:3], each = 5)
clusters <- sample(letters[1:5], length(classes), replace = TRUE)

#From the table below:
# - cluster 'b' corresponds to class A;
# - either of the clusters 'd' and 'e' can correspond to class B,
#   however, 'e' should be chosen, because cluster 'd' also highly
#   intersects with Class C. Thus,
# - cluster 'd' corresponds to class C.
table(classes, clusters)
##       clusters
##classes a b c d e
##      A 0 3 1 0 1
##      B 1 0 0 2 2
##      C 1 2 0 2 0

#The function does this choice automatically:
purity(classes, clusters)

#Sample output:
##$pur
##[1] 0.4666667
##
##$out
##  ClassLabels ClusterLabels ClusterSize
##1           A             b           3
##2           B             e           2
##3           C             d           2


##### Example 2
#The labels can be also numeric:
classes <- rep(1:5, each = 3)
clusters <- sample(1:3, length(classes), replace = TRUE)
purity(classes, clusters)
```

`q.tails` *Quantile-Based Tails Comparison*

Description

Compare right tails of two sample distributions using a quantile-based approach (QBA); see Soliman et al. (2014), Soliman et al. (2015), and Lyubchich and Gel (2017).

Usage

```
q.tails(x0, x1, q = 0.99)
```

Arguments

x0, x1	vectors of the same length (preferably). Tail in x1 is compared against the tail in x0.
q	a quantile defining the right tail for both x0 and x1. Values above the thresholds quantile(x0,probs = q) and quantile(x1,probs = q) are considered as the respective right tails.

Details

Sturges' formula is used to calculate number of intervals (k) to split the upper $100(1-q)\backslash$ (the right tails). Then, each tail is divided into equally-filled intervals with a quantile step $d = (1-q)/k$. Pk reports the difference between corresponding intervals' centers obtained from x0 and x1.

Value

A list with two elements:

d	the step in probabilities for defining the quantiles.
Pk	vector of differences of the intervals' centers.

Author(s)

Vyacheslav Lyubchich, Yulia R. Gel

References

Lyubchich V, Gel YR (2017). "Can we weather proof our insurance?" *Environmetrics*, **28**(2), e2433. doi: 10.1002/env.2433.

Soliman M, Lyubchich V, Gel YR, Naser D, Esterby S (2015). "Evaluating the impact of climate change on dynamics of house insurance claims." In Lakshmanan V, Gilleland E, McGovern A, Tingley M (eds.), *Machine Learning and Data Mining Approaches to Climate Science*, chapter 16, 175–183. Springer, Switzerland. doi: 10.1007/9783319172200_16.

Soliman M, Naser D, Lyubchich V, Gel YR, Esterby S (2014). "Evaluating the impact of climate change on dynamics of house insurance claims." In Ebert-Uphoff I (ed.), *The 4th International Workshop on Climate Informatics: CI2014*.

See Also

i.tails

Examples

```
x0 <- rnorm(1000)
x1 <- rt(1000, 5)
q.tails(x0, x1)
```

---

sync_cluster *Time Series Clustering based on Trend Synchronism*

Description

Cluster time series with a common parametric trend using the sync_test function (Lyubchich and Gel 2016; Ghahari et al. 2017).

Usage

```
sync_cluster(formula, rate = 1, alpha = 0.05, ...)
```

Arguments

formula	an object of class "formula", specifying the type of common trend for clustering the time series in a T by N matrix of time series (time series in columns) which is passed to sync_test. Variable t should be used to specify the form of the trend, where t is specified within the function automatically as a regular sequence of length T on the interval (0,1]. See Examples.
rate	rate of removal of time series. Default is 1 (i.e., if hypothesis of synchronism is rejected one time series is removed at a time to re-test the remaining time series). Integer values above 1 are treated as number of time series to be removed. Values from 0 to 1 are treated as a fraction of time series to be removed.
alpha	significance level for testing hypothesis of a common trend (using sync_test) of the parametric form specified in formula.
...	arguments to be passed to sync_test, for example, number of bootstrap replications (B).

Details

The sync_cluster function recursively clusters time series having a pre-specified common parametric trend until there are no time series left. Starting with the given N time series, the sync_test function is used to test for a common trend. If null hypothesis of common trend is not rejected by sync_test, the time series are grouped together (i.e., assigned to a cluster). Otherwise, the time series with the largest contribution to the test statistics are temporarily removed (the number of time series to remove depends on the rate of removal) and sync_test is applied again. The contribution to the test statistic is assessed by the WAVK test statistic calculated for each time series.

Value

A list with the elements:

`cluster` an integer vector indicating the cluster to which each time series is allocated. A label `'0'` is assigned to time series which do not have a common trend with other time series (that is, all time series labeled with `'0'` are separate one-element clusters).

`elements` a list with names of the time series in each cluster.

The further elements combine results of `sync_test` for each cluster with at least two elements (that is, single-element clusters labeled with `'0'` are excluded):

`estimate` a list with common parametric trend estimates obtained by `sync_test` for each cluster. The length of this list is `max(cluster)`.

`pval` a list of p-values of `sync_test` for each cluster. The length of this list is `max(cluster)`.

`statistic` a list with values of `sync_test` test statistic for each cluster. The length of this list is `max(cluster)`.

`ar_order` a list of AR filter orders used in `sync_test` for each time series. The results are grouped by cluster in the list of length `max(cluster)`.

`window_used` a list of local windows used in `sync_test` for each time series. The results are grouped by cluster in the list of length `max(cluster)`.

`all_considered_windows` a list of all windows considered in `sync_test` and corresponding test results, for each cluster. The length of this list is `max(cluster)`.

`WAVK_obs` a list of WAVK test statistics obtained in `sync_test` for each time series. The results are grouped by cluster in the list of length `max(cluster)`.

Author(s)

Srishti Vishwakarma, Vyacheslav Lyubchich

References

Ghahari A, Gel YR, Lyubchich V, Chun Y, Uribe D (2017). "On employing multi-resolution weather data in crop insurance." In *Proceedings of the SIAM International Conference on Data Mining (SDM17) Workshop on Mining Big Data in Climate and Environment (MBDCE 2017)*.

Lyubchich V, Gel YR (2016). "A local factor nonparametric test for trend synchronism in multiple time series." *Journal of Multivariate Analysis*, **150**, 91–104. doi: 10.1016/j.jmva.2016.05.004.

See Also

`BICC`, `DR`, `sync_test`

Examples

```
## Not run:
## Simulate 4 autoregressive time series,
## 3 having a linear trend and 1 without a trend:
set.seed(123)
T = 100 #length of time series
N = 4 #number of time series
X = sapply(1:N, function(x) arima.sim(n = T,
           list(order = c(1, 0, 0), ar = c(0.6))))
X[,1] <- 5 * (1:T)/T + X[,1]
plot.ts(X)

# Finding clusters with common linear trends:
LinTrend <- sync_cluster(X ~ t)

## Sample Output:
##[1] "Cluster labels:"
##[1] 0 1 1 1
##[1] "Number of single-element clusters (labeled with '0'): 1"

## plotting the time series of the cluster obtained
for(i in 1:max(LinTrend$cluster)) {
    plot.ts(X[, LinTrend$cluster == i],
            main = paste("Cluster", i))
}


## Simulating 7 autoregressive time series,
## where first 4 time series have a linear trend added
set.seed(234)
T = 100 #length of time series
a <- sapply(1:4, function(x) -10 + 0.1 * (1:T) +
            arima.sim(n = T, list(order = c(1, 0, 0), ar = c(0.6))))
b <- sapply(1:3, function(x) arima.sim(n = T,
            list(order = c(1, 0, 0), ar = c(0.6))))
Y <- cbind(a, b)
plot.ts(Y)

## Clustering based on linear trend with rate of removal = 2
# and confidence level for the synchronism test 90%
LinTrend7 <- sync_cluster(Y ~ t, rate = 2, alpha = 0.1, B = 99)

## Sample output:
##[1] "Cluster labels:"
##[1] 1 1 1 0 2 0 2
##[1] "Number of single-element clusters (labeled with '0'): 2"

## End(Not run)
```

sync_test *Time Series Trend Synchronism Test*

Description

Non-parametric test for synchronism of parametric trends in multiple time series (Lyubchich and Gel 2016). The method tests whether N observed time series exhibit the same trend of some pre-specified smooth parametric form.

Usage

```
sync_test(
  formula,
  B = 1000,
  Window = NULL,
  q = NULL,
  j = NULL,
  ar.order = NULL,
  ar.method = "HVK",
  BIC = TRUE
)
```

Arguments

formula	an object of class "formula", specifying the form of the common parametric time trend to be tested in a T by N matrix of time series (time series in columns). Variable t should be used to specify the form of the trend, where t is specified within the function as a regular sequence on the interval (0,1]. See 'Examples'.
B	number of bootstrap simulations to obtain empirical critical values. Default is 1000.
Window	scalar or N-vector with lengths of the local windows (factors). If only one value is set, the same Window is applied to each time series. An N-vector gives a specific window for each time series. If Window is not specified, an automatic algorithm for optimal window selection is applied as a default option (see 'Details').
q	scalar from 0 to 1 to define the set of possible windows T*q^j and to automatically select an optimal window for each time series. Default is $3/4$. This argument is ignored if Window is set by user.
j	numeric vector to define the set of possible windows T*q^j and to automatically select an optimal window for each time series. Default is c(8:11). This argument is ignored if Window is set by user.
ar.order	order of autoregressive filter when BIC = FALSE, or the maximal order for BIC-based filtering. Default is round(10*log10(T)). The ar.order can be a scalar or N-vector. If scalar, the same ar.order is applied to each time series. An N-vector specifies a separate ar.order for each time series.

`ar.method` method of estimating autoregression coefficients. Default `"HVK"` delivers robust difference-based estimates by Hall and Van Keilegom (2003). Alternatively, options of `ar` function can be used, such as `"burg"`, `"ols"`, `"mle"`, and `"yw"`.

`BIC` logical value indicates whether the order of autoregressive filter should be selected by Bayesian information criterion (BIC). If `TRUE` (default), models of orders p = 0,1,...,`ar.order` or p = 0,1,...,`round(10*log10(length(x)))` are considered, depending on whether `ar.order` is defined or not (`x` is the time series).

Details

Arguments `Window`, `j`, and `q` are used to set windows for the local regression. Current version of the function assumes two options: (1) user specifies one fixed window for each time series using the argument `Window` (if `Window` is set, `j` and `q` are ignored), and (2) user specifies a set of windows by `j` and `q` to apply this set to each time series and to select an optimal window using a heuristic m-out-of-n subsampling algorithm (Bickel and Sakov 2008). The option of selecting windows automatically for some of the time series, while for other time series the window is fixed, is not available yet. If none of these three arguments is set, default `j` and `q` are used. Values `T*q^j` are mapped to the largest previous integer, then only those greater than 2 are used.

See more details in Lyubchich and Gel (2016) and Lyubchich (2016).

Value

A list of class `"htest"` containing the following components:

`method` name of the method.

`data.name` name of the data.

`statistic` value of the test statistic.

`p.value` p-value of the test.

`alternative` alternative hypothesis.

`estimate` list with elements `common_trend_estimates`, `ar_order_used`, `Window_used`, `wavk_obs`, and `all_considered_windows`. The latter is a table with bootstrap and asymptotic test results for all considered windows, that is, without adaptive selection of the local window.

Author(s)

Yulia R. Gel, Vyacheslav Lyubchich, Ethan Schaeffer, Xingyu Wang

References

Bickel PJ, Sakov A (2008). "On the choice of m in the m out of n bootstrap and confidence bounds for extrema." *Statistica Sinica*, **18**(3), 967–985.

Hall P, Van Keilegom I (2003). "Using difference-based methods for inference in nonparametric regression with time series errors." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **65**(2), 443–456. doi: 10.1111/14679868.00395.

Lyubchich V (2016). "Detecting time series trends and their synchronization in climate data." *Intelligence. Innovations. Investments*, **12**, 132–137.

Lyubchich V, Gel YR (2016). "A local factor nonparametric test for trend synchronism in multiple time series." *Journal of Multivariate Analysis*, **150**, 91–104. doi: 10.1016/j.jmva.2016.05.004.

See Also

ar, HVK, WAVK, wavk_test

Examples

```
#Fix seed for reproducible simulations:
set.seed(1)

# Simulate two autoregressive time series of length n without trend
#(i.e., with zero or constant trend)
# and arrange the series into a matrix:
n <- 200
y1 <- arima.sim(n = n, list(order = c(1, 0, 0), ar = c(0.6)))
y2 <- arima.sim(n = n, list(order = c(1, 0, 0), ar = c(-0.2)))
Y <- cbind(y1, y2)
plot.ts(Y)


#Test H0 of a common linear trend:
## Not run:
    sync_test(Y ~ t, B = 500)

## End(Not run)
# Sample output:
## Non-parametric test for synchronism of parametric trends
##
##data:  Y
##Test statistic = -0.0028999, p-value = 0.7
##alternative hypothesis: common trend is not of the form Y ~ t.
##sample estimates:
##$common_trend_estimates
##               Estimate Std. Error    t value  Pr(>|t|)
##(Intercept) -0.02472566  0.1014069 -0.2438261 0.8076179
##t            0.04920529  0.1749859  0.2811958 0.7788539
##
##$ar.order_used
##         y1 y2
##ar.order  1  1
##
##$Window_used
##       y1 y2
##Window 15  8
##
##$all_considered_windows
## Window    Statistic p-value Asympt. p-value
```

```
##      8 -0.000384583   0.728       0.9967082
##     11 -0.024994408   0.860       0.7886005
##     15 -0.047030164   0.976       0.6138976
##     20 -0.015078579   0.668       0.8714980
##
##$wavk_obs
##[1]  0.05827148 -0.06117136

# Add a time series y3 with a different linear trend and re-apply the test:
y3 <- 1 + 3*((1:n)/n) + arima.sim(n = n, list(order = c(1, 0, 0), ar = c(-0.2)))
Y2 <- cbind(Y, y3)
plot.ts(Y2)
## Not run:
    sync_test(Y2 ~ t, B = 500)
## End(Not run)
# Sample output:
## Non-parametric test for synchronism of parametric trends
##
##data:  Y2
##Test statistic = 0.48579, p-value < 2.2e-16
##alternative hypothesis: common trend is not of the form Y2 ~ t.
##sample estimates:
##$common_trend_estimates
##              Estimate Std. Error  t value     Pr(>|t|)
##(Intercept) -0.3632963 0.07932649 -4.57976 8.219360e-06
##t            0.7229777 0.13688429  5.28167 3.356552e-07
##
##$ar.order_used
##         Y.y1 Y.y2 y3
##ar.order    1    1  0
##
##$Window_used
##       Y.y1 Y.y2 y3
##Window    8   11  8
##
##$all_considered_windows
## Window Statistic p-value Asympt. p-value
##      8 0.4930069       0    1.207378e-05
##     11 0.5637067       0    5.620248e-07
##     15 0.6369703       0    1.566057e-08
##     20 0.7431621       0    4.201484e-11
##
##$wavk_obs
##[1]  0.08941797 -0.07985614  0.34672734

#Other hypothesized trend forms can be specified, for example:
## Not run:
    sync_test(Y ~ 1) #constant trend
    sync_test(Y ~ poly(t, 2)) #quadratic trend
    sync_test(Y ~ poly(t, 3)) #cubic trend

## End(Not run)
```

TopoCBN *Topological Clustering using Betti Numbers*

Description

The function performs unsupervised clustering of multivariate data based on topological data analysis (TDA). The objective is to partition data into non-overlapping clusters, where the definition of a cluster falls under a general framework of density based clustering, e.g., DBSCAN and OPTICS. That is, intuitively the cluster is a subset of points which is path-connected, i.e., any point in the subset can be reached from any other one through a path consisting of points (also belonging to the subset); furthermore, the consecutive points on the path are close enough and their local neighborhoods are similar in shape (Islambekov and Gel 2019). To compare shapes, TopoCBN builds a Vietoris–Rips (VR) filtration upon such neighborhoods around each point and computes topological summaries in the form of the Betti sequences using persistent homology. The closer the Betti sequences to one another for a pair of close-by points, the more similar the shapes of their neighborhoods. Thus, when identifying clusters, TopoCBN utilizes both the distance function and local geometric information around the points. Note that accounting for shape similarity can be viewed as an extension of conventional clustering properties in the density-based clustering framework.

Usage

```
TopoCBN(data, nKNN, filt_len = 25, dist_matrix = FALSE)
```

Arguments

`data`	a point cloud given as an N by d matrix, where N = number of points, d = dimension of Euclidean space or an N by N matrix of pairwise distances.
`nKNN`	number of k nearest neighbors to take around each point.
`filt_len`	filtration length (also length of Betti sequences). Default is 25.
`dist_matrix`	is set to `FALSE` by default, assuming `data` is a point cloud. Set `dist_matrix = TRUE` if data is a matrix of pairwise distances.

Value

A list with the following components:

`assignments`	cluster labels (vector of length N).
`nClust`	number of clusters.
`cSize`	cluster sizes (vector of length `nClust`).

Author(s)

Palina Niamkova, Umar Islambekov, Yulia R. Gel

References

Islambekov U, Gel YR (2019). "Unsupervised space–time clustering using persistent homology." *Environmetrics*, **30**(4), e2539. doi: 10.1002/env.2539.

See Also

cumsumCPA_test for change point detection in time series via a linear regression with temporally correlated errors

Examples

```
## Not run:
#Example 1:
#Let's import dataset with today's Covid-19 parameters per each state:
data<-covid19us::get_states_current()
#For this example we will keep data for positive cases and deaths today:
data<-data[c(3,9)]
#We also need to replace NA values to integer 0:
data[is.na(data)] = 0

#Now run CBN:
result <- TopoCBN(data,nKNN=12) # can also try with filt_len=50,75,100

#We can obtain the same results using matrix of pairwise distances:
dMatrix <- as.matrix(dist(data))
result <- TopoCBN(dMatrix,nKNN=12,dist_matrix = TRUE)

#Let's plot the results:
set.seed(365)
distinct_clrs=randomcoloR::distinctColorPalette(result$nClust)
clrs<-distinct_clrs[result$assignments] # distinct colors for clusters
plot(data,col=clrs,pch=20,xlab='x',ylab='y',main = 'TopoCBN')
print(result)

#We can see that CBN function identified 6 clusters within our dataset.

#Example 2:
#Let's import dataset with air quality level in  Californian metropolitan areas. The three
#columns of the dataset contains indicator of air quality (the lower the better), value
#added of companies (in thousands of dollars).

data<-as.matrix(Ecdat::Airq[1:3])

#Now apply TopoCBN function to the air quality data:
result <- TopoCBN(data,nKNN=12) # can also try with filt_len=50,75,100

#The same results can be obtained using matrix of pairwise distances:
dMatrix <- as.matrix(dist(data))
result <- TopoCBN(dMatrix,nKNN=12,dist_matrix = TRUE)

#Plot the results:
set.seed(365)
```

```
distinct_clrs=randomcoloR::distinctColorPalette(result$nClust)
clrs<-distinct_clrs[result$assignments] # distinct colors for clusters
plot(data,col=clrs,pch=20,xlab='x',ylab='y',main = 'TopoCBN')
print(result)

#We see that TopoCBN identified 4 clusters within our dataset of the sizes
#1,3,5, and 21. These results suggest that companies with added values under $5,000 may
#have any value of air pollution. However, companies with higher added values (>$5,000)
#correspond to the dramatically increased (deteriorated) levels of air pollution.

## End(Not run)
```

WAVK *WAVK Statistic*

Description

Statistic for testing the parametric form of a regression function, suggested by Wang et al. (2008).

Usage

```
WAVK(z, kn = NULL)
```

Arguments

z	pre-filtered univariate time series (see formula (2.1) by Wang and Van Keilegom 2007): $$Z_i = \left(Y_{i+p} - \sum_{j=1}^{p} \hat{\phi}_{j,n} Y_{i+p-j} \right) - \left(f(\hat{\theta}, t_{i+p}) - \sum_{j=1}^{p} \hat{\phi}_{j,n} f(\hat{\theta}, t_{i+p-j}) \right),$$ where Y_i is observed time series of length n, $\hat{\theta}$ is an estimator of hypothesized parametric trend $f(\theta, t)$, and $\hat{\phi}_p = (\hat{\phi}_{1,n}, \ldots, \hat{\phi}_{p,n})'$ are estimated coefficients of an autoregressive filter of order p. Missing values are not allowed.
kn	length of the local window.

Value

A list with following components:

Tn	test statistic based on artificial ANOVA and defined by Wang and Van Keilegom (2007) as a difference of mean square for treatments (MST) and mean square for errors (MSE): $$T_n = MST - MSE = \frac{k_n}{n-1} \sum_{t=1}^{T} \left(\overline{V}_{t.} - \overline{V}_{..} \right)^2 - \frac{1}{n(k_n - 1)} \sum_{t=1}^{n} \sum_{j=1}^{k_n} \left(V_{tj} - \overline{V}_{t.} \right)^2,$$ where $\{V_{t1}, \ldots, V_{tk_n}\} = \{Z_j : j \in W_t\}$, W_t is a local window, $\overline{V}_{t.}$ and $\overline{V}_{..}$ are the mean of the tth group and the grand mean, respectively.

`Tns` standardized version of `Tn` according to Theorem 3.1 by Wang and Van Keilegom (2007):

$$T_{ns} = \left(\frac{n}{k_n}\right)^{\frac{1}{2}} T_n \Big/ \left(\frac{4}{3}\right)^{\frac{1}{2}} \sigma^2,$$

where n is length and σ^2 is variance of the time series. Robust difference-based Rice's estimator (Rice 1984) is used to estimate σ^2.

`p.value` p-value for `Tns` based on its asymptotic $N(0, 1)$ distribution.

Author(s)

Yulia R. Gel, Vyacheslav Lyubchich

References

Rice J (1984). "Bandwidth choice for nonparametric regression." *The Annals of Statistics*, **12**(4), 1215–1230. doi: 10.1214/aos/1176346788.

Wang L, Akritas MG, Van Keilegom I (2008). "An ANOVA-type nonparametric diagnostic test for heteroscedastic regression models." *Journal of Nonparametric Statistics*, **20**(5), 365–382.

Wang L, Van Keilegom I (2007). "Nonparametric test for the form of parametric regression with time series errors." *Statistica Sinica*, **17**, 369–386.

See Also

`wavk_test`

Examples

```
z <- rnorm(300)
WAVK(z, kn = 7)
```

`wavk_test` *WAVK Trend Test*

Description

Non-parametric test to detect (non-)monotonic parametric trends in time series (based on Lyubchich et al. 2013).

Usage

```
wavk_test(
  formula,
  factor.length = c("user.defined", "adaptive.selection"),
  Window = NULL,
  q = 3/4,
  j = c(8:11),
  B = 1000,
  method = c("boot", "asympt"),
  ar.order = NULL,
  ar.method = "HVK",
  BIC = TRUE,
  out = FALSE
)
```

Arguments

formula	an object of class "formula", specifying the form of the parametric time trend to be tested. Variable t should be used to specify the form, where t is specified within the function as a regular sequence on the interval (0,1]. See Examples.
factor.length	method to define the length of local windows (factors). Default option "user.defined" allows to set only one value of the argument Window. The option "adaptive.selection" sets method = "boot" and employs heuristic m-out-of-n subsampling algorithm (Bickel and Sakov 2008) to select an optimal window from the set of possible windows length(x)*q^j whose values are mapped to the largest previous integer and greater than 2. Vector x is the time series tested.
Window	length of the local window (factor), default is round(0.1*length(x)), where x is the time series tested. This argument is ignored if factor.length = "adaptive.selection".
q	scalar from 0 to 1 to define the set of possible windows when factor.length = "adaptive.selection". Default is $3/4$. This argument is ignored if factor.length = "user.defined".
j	numeric vector to define the set of possible windows when factor.length = "adaptive.selection". Default is c(8:11). This argument is ignored if factor.length = "user.defined".
B	number of bootstrap simulations to obtain empirical critical values. Default is 1000.
method	method of obtaining critical values: from asymptotical ("asympt") or bootstrap ("boot") distribution. If factor.length = "adaptive.selection" the option "boot" is used.
ar.order	order of autoregressive model when BIC = FALSE, or the maximal order for BIC-based filtering. Default is round(10*log10(length(x))), where x is the time series.
ar.method	method of estimating autoregression coefficients. Default "HVK" delivers robust difference-based estimates by Hall and Van Keilegom (2003). Alternatively, options of ar function can be used, such as "burg", "ols", "mle", and "yw".

BIC	logical value indicates whether the order of autoregressive filter should be selected by Bayesian information criterion (BIC). If TRUE (default), models of orders p = 0,1,...,ar.order or p = 0,1,...,round(10*log10(length(x))) are considered, depending on whether ar.order is defined or not (x is the time series).
out	logical value indicates whether full output should be shown. Default is FALSE.

Details

See more details in Lyubchich and Gel (2016) and Lyubchich (2016).

Value

A list with class "htest" containing the following components:

method	name of the method.
data.name	name of the data.
statistic	value of the test statistic.
p.value	p-value of the test.
alternative	alternative hypothesis.
parameter	window that was used.
estimate	list with the following elements: estimated trend coefficients; user-defined or BIC-selected AR order; estimated AR coefficients; and, if factor.length = "adaptive.selection", test results for all considered windows.

Author(s)

Yulia R. Gel, Vyacheslav Lyubchich, Ethan Schaeffer

References

Bickel PJ, Sakov A (2008). "On the choice of m in the m out of n bootstrap and confidence bounds for extrema." *Statistica Sinica*, **18**(3), 967–985.

Hall P, Van Keilegom I (2003). "Using difference-based methods for inference in nonparametric regression with time series errors." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **65**(2), 443–456. doi: 10.1111/14679868.00395.

Lyubchich V (2016). "Detecting time series trends and their synchronization in climate data." *Intelligence. Innovations. Investments*, **12**, 132–137.

Lyubchich V, Gel YR (2016). "A local factor nonparametric test for trend synchronism in multiple time series." *Journal of Multivariate Analysis*, **150**, 91–104. doi: 10.1016/j.jmva.2016.05.004.

Lyubchich V, Gel YR, El-Shaarawi A (2013). "On detecting non-monotonic trends in environmental time series: a fusion of local regression and bootstrap." *Environmetrics*, **24**(4), 209–226. doi: 10.1002/env.2212.

See Also

ar, HVK, WAVK, sync_test, vignette("trendtests",package = "funtimes")

Examples

```
# Fix seed for reproducible simulations:
set.seed(1)

#Simulate autoregressive time series of length n with smooth quadratic trend:
n <- 100
tsTrend <- 1 + 2*(1:n/n) + 4*(1:n/n)^2
tsNoise <- arima.sim(n = n, list(order = c(2, 0, 0), ar = c(-0.7, -0.1)))
U <- tsTrend + tsNoise
plot.ts(U)

#Test H0 of a linear trend, with m-out-of-n selection of the local window:
## Not run:
    wavk_test(U ~ t, factor.length = "adaptive.selection")
## End(Not run)
# Sample output:
## Trend test by Wang, Akritas, and Van Keilegom (bootstrap p-values)
##
##data:  U
##WAVK test statistic = 5.3964, adaptively selected window = 4, p-value < 2.2e-16
##alternative hypothesis: trend is not of the form U ~ t.

#Test H0 of a quadratic trend, with m-out-of-n selection of the local window
#and output of all results:
## Not run:
    wavk_test(U ~ poly(t, 2), factor.length = "adaptive.selection", out = TRUE)
## End(Not run)
# Sample output:
## Trend test by Wang, Akritas, and Van Keilegom (bootstrap p-values)
##
##data:  U
##WAVK test statistic = 0.40083, adaptively selected window = 4, p-value = 0.576
##alternative hypothesis: trend is not of the form U ~ poly(t, 2).
##sample estimates:
##$trend_coefficients
##(Intercept) poly(t, 2)1 poly(t, 2)2
##   3.408530   17.681422    2.597213
##
##$AR_order
##[1] 1
##
##$AR_coefficients
##         phi_1
##[1] -0.7406163
##
##$all_considered_windows
## Window WAVK-statistic p-value
##      4     0.40083181   0.576
```

```
##      5     0.06098625   0.760
##      7    -0.57115451   0.738
##     10    -1.02982929   0.360

# Test H0 of no trend (constant trend) using asymptotic distribution of statistic.
wavk_test(U ~ 1, method = "asympt")
# Sample output:
## Trend test by Wang, Akritas, and Van Keilegom (asymptotic p-values)
##
##data:  U
##WAVK test statistic = 25.999, user-defined window = 10, p-value < 2.2e-16
##alternative hypothesis: trend is not of the form U ~ 1.
```

Index