

Package ‘clustcurv’

October 25, 2024

Type Package

Title Determining Groups in Multiples Curves

URL <https://github.com/noramvillanueva/clustcurv>

BugReports <https://github.com/noramvillanueva/clustcurv/issues>

Version 2.0.2

Date 2024-10-09

Maintainer Nora M. Villanueva <nmvillanueva@uvigo.es>

Description A method for determining groups in multiple curves with an automatic selection of their number based on k-means or k-medians algorithms. The selection of the optimal number is provided by bootstrap methods. The methodology can be applied both in regression and survival framework. Implemented methods are:
Grouping multiple survival curves described by Villanueva et al. (2018) <[doi:10.1002/sim.8016](https://doi.org/10.1002/sim.8016)>.

Depends R (>= 3.5.0)

License MIT + file LICENSE

Encoding UTF-8

LazyData true

Imports doParallel, doRNG, foreach, ggfortify, ggplot2, Gmedian, grDevices, npregfast, RColorBrewer, survival

Suggests covr, knitr, plotly, rmarkdown, testthat (>= 3.0.0)

Config/testthat/edition 3

RoxygenNote 7.3.2

VignetteBuilder knitr

NeedsCompilation no

Author Nora M. Villanueva [aut, cre] (<<https://orcid.org/0000-0001-8085-2745>>),
Marta Sestelo [aut] (<<https://orcid.org/0000-0003-4284-6509>>)

Repository CRAN

Date/Publication 2024-10-25 08:20:07 UTC

Contents

autoplot.clustcurves	2
barnacle5	4
clustcurv	5
kregcurves	6
ksurvcurves	7
regclustcurves	8
summary.clustcurves	10
summary.kcurves	12
survclustcurves	13

Index	16
--------------	-----------

autoplot.clustcurves	<i>Visualization of clustcurves objects with ggplot2 graphics</i>
----------------------	---

Description

Useful for drawing the estimated functions grouped by color and the centroids (mean curve of the curves pertaining to the same group).

Usage

```
## S3 method for class 'clustcurves'
autoplot(
  object = object,
  groups_by_colour = TRUE,
  centers = FALSE,
  conf.int = FALSE,
  censor = FALSE,
  xlab = "Time",
  ylab = "Survival",
  interactive = FALSE,
  ...
)
```

Arguments

object	Object of clustcurves class.
groups_by_colour	A specification for the plotting groups by color.
centers	Draw the centroids (mean of the curves pertaining to the same group) into the plot. By default it is FALSE.
conf.int	Only for survival curves. Logical flag indicating whether to plot confidence intervals.
censor	Only for survival curves. Logical flag indicating whether to plot censors.

<code>xlab</code>	A title for the x axis.
<code>ylab</code>	A title for the y axis.
<code>interactive</code>	Logical flag indicating if an interactive plot with plotly is produced.
<code>...</code>	Other options.

Details

See help page of the function `ggfortify::autoplot.survfit()`.

Value

A ggplot object, so you can use common features from ggplot2 package to manipulate the plot.

Author(s)

Nora M. Villanueva and Marta Sestelo.

Examples

```
library(survival)
library(clustcurv)
library(ggplot2)
library(ggfortify)

# Survival

c12 <- ksurvcurves(time = veteran$time, status = veteran$status,
x = veteran$celltype, k = 2, algorithm = "kmeans")

autoplot(c12)
autoplot(c12, groups_by_colour = FALSE)
autoplot(c12, centers = TRUE)

# Regression

r2 <- kregcurves(y = barnacle5$DW, x = barnacle5$RC,
z = barnacle5$F, k = 2, algorithm = "kmeans")

autoplot(r2)
autoplot(r2, groups_by_colour = FALSE)
autoplot(r2, groups_by_colour = FALSE, interactive = TRUE)
autoplot(r2, centers = TRUE)
```

`barnacle5`*Barnacle data*

Description

This barnacle data set gives the measurements of the variables dry weight (in g.) and rostro-carinal length (in mm) for 5000 barnacles collected along the intertidal zone from five sites of the Atlantic coast of Galicia (Spain).

Usage`barnacle5`**Format**

`barnacle5` is a data frame with 5000 cases (rows) and 3 variables (columns).

Note that barnacle data set from the `npregfast` package gives the same three variables (columns) but for two sites, thus 2000 cases (rows).

DW Dry weight (in g.)

RC Rostro-carinal length (in mm).

F Factor indicating the sites of harvest: `laxe`, `lens`, `barca`, `laxe`, and `lens`.

Author(s)

Marta Sestelo

References

Sestelo, M. and Roca-Pardinas, J. (2011). A new approach to estimation of length-weight relationship of *Pollicipes pollicipes* (Gmelin, 1789) on the Atlantic coast of Galicia (Northwest Spain): some aspects of its biology and management. *Journal of Shellfish Research*, 30(3), 939–948.

Sestelo, M., Villanueva, N.M., Meira-Machado, L., Roca-Pardinas, J. (2017). `npregfast`: An R Package for Nonparametric Estimation and Inference in Life Sciences. *Journal of Statistical Software*, 82(12), 1-27.

Examples

```
data(barnacle5)
head(barnacle5)
```

`clustcurv`*clustcurv: Determining Groups in Multiple Curves.*

Description

This package provides a method for determining groups in multiple curves with an automatic selection of their number based on k-means or k-medians algorithms. The selection of the optimal number is provided by bootstrap methods. The methodology can be applied both in regression and survival framework.

Details

Package: `clustcurv`
Type: `Package`
License: `MIT + file LICENSE`

`clustcurv` is designed along lines similar to those of other R packages. This software helps the user determine groups in multiple curves (survival and regression curves). In addition, it enables both numerical and graphical outputs to be displayed (by means of `ggplot2`). The package provides the `kclustcurv()` function that groups the curves given a number `k` and the `autoclustcurv()` function that selects the optimal number of groups automatically through a bootstrap-based test. The `autoplot()` function let the user draw the resulted estimated curves coloured by groups.

For a listing of all routines in the `clustcurv` package type: `library(help="clustcurv")`.

Author(s)

Nora M. Villanueva and Marta Sestelo

References

Villanueva, N. M., Sestelo, M., and Meira-Machado, J. (2019). A method for determining groups in multiple survival curves. *Statistics in Medicine*, 8(5):866-877

See Also

Useful links:

- <https://github.com/noramvillanueva/clustcurv>
- Report bugs at <https://github.com/noramvillanueva/clustcurv/issues>

kregcurves

k-groups of multiple regression curves

Description

Function for grouping regression curves, given a number k , based on the k -means or k -medians algorithm.

Usage

```
kregcurves(y, x, z, k, kbin = 50, h = -1, algorithm = "kmeans", seed = NULL)
```

Arguments

<code>y</code>	Response variable.
<code>x</code>	Dependent variable.
<code>z</code>	Categorical variable indicating the population to which the observations belong.
<code>k</code>	An integer specifying the number of groups of curves to be performed.
<code>kbin</code>	Size of the grid over which the survival functions are to be estimated.
<code>h</code>	The kernel bandwidth smoothing parameter.
<code>algorithm</code>	A character string specifying which clustering algorithm is used, i.e., k -means ("kmeans") or k -medians ("kmedians").
<code>seed</code>	Seed to be used in the procedure.

Value

A list containing the following items:

<code>measure</code>	Value of the test statistic.
<code>levels</code>	Original levels of the variable <code>fac</code> .
<code>cluster</code>	A vector of integers (from 1:k) indicating the cluster to which each curve is allocated.
<code>centers</code>	An object containing the fitted centroids (mean of the curves pertaining to the same group).
<code>curves</code>	An object containing the fitted regression curves for each population.

Author(s)

Nora M. Villanueva and Marta Sestelo.

Examples

```
library(clustcurv)

# Regression: 2 groups k-means
r2 <- kregcurves(y = barnacle5$DW, x = barnacle5$RC,
z = barnacle5$F, k = 2, algorithm = "kmeans")

data.frame(level = r2$level, cluster = r2$cluster)
```

ksurvcurves

*k-groups of multiple survival curves***Description**

Function for grouping survival curves, given a number k , based on the k -means or k -medians algorithm.

Usage

```
ksurvcurves(
  time,
  status = NULL,
  x,
  k,
  kbin = 50,
  algorithm = "kmeans",
  seed = NULL
)
```

Arguments

time	Survival time.
status	Censoring indicator of the survival time of the process; 0 if the total time is censored and 1 otherwise.
x	Categorical variable indicating the population to which the observations belong.
k	An integer specifying the number of groups of curves to be performed.
kbin	Size of the grid over which the survival functions are to be estimated.
algorithm	A character string specifying which clustering algorithm is used, i.e., k -means("kmeans") or k -medians("kmedians").
seed	Seed to be used in the procedure.

Value

A list containing the following items:

measure	Value of the test statistics.
levels	Original levels of the variable x.
cluster	A vector of integers (from 1:k) indicating the cluster to which each curve is allocated.
centers	An object of class <code>survfit</code> containing the centroids (mean of the curves pertaining to the same group).
curves	An object of class <code>survfit</code> containing the survival curves for each population.

Author(s)

Nora M. Villanueva and Marta Sestelo.

Examples

```
library(clustcurv)
library(survival)
data(veteran)

# Survival: 2 groups k-means
s2 <- ksurvcurves(time = veteran$time, status = veteran$status,
x = veteran$celltype, k = 2, algorithm = "kmeans")

data.frame(level = s2$level, cluster = s2$cluster)

# Survival: 2 groups k-medians
s22 <- ksurvcurves(time = veteran$time, status = veteran$status,
x = veteran$celltype, k = 2, algorithm = "kmedians")

data.frame(level = s22$level, cluster = s22$cluster)
```

regclustcurves

Clustering multiple regression curves

Description

Function for grouping regression curves based on the k-means or k-medians algorithm. It returns the number of groups and the assignment.

Usage

```
regclustcurves(
  y,
  x,
  z,
  kvector = NULL,
  kbin = 50,
  h = -1,
  nboot = 100,
  algorithm = "kmeans",
  alpha = 0.05,
  cluster = FALSE,
  ncores = NULL,
  seed = NULL,
  multiple = FALSE,
  multiple.method = "holm"
)
```

Arguments

y	Response variable.
x	Dependent variable.
z	Categorical variable indicating the population to which the observations belong.
kvector	A vector specifying the number of groups of curves to be checking.
kbin	Size of the grid over which the survival functions are to be estimated.
h	The kernel bandwidth smoothing parameter.
nboot	Number of bootstrap repeats.
algorithm	A character string specifying which clustering algorithm is used, i.e., k-means("kmeans") or k-medians ("kmedians").
alpha	Significance level of the testing procedure. Defaults to 0.05.
cluster	A logical value. If TRUE (default), the testing procedure is parallelized. Note that there are cases (e.g., a low number of bootstrap repetitions) that R will gain in performance through serial computation. R takes time to distribute tasks across the processors also it will need time for binding them all together later on. Therefore, if the time for distributing and gathering pieces together is greater than the time need for single-thread computing, it does not worth parallelize.
ncores	An integer value specifying the number of cores to be used in the parallelized procedure. If NULL (default), the number of cores to be used is equal to the number of cores of the machine - 1.
seed	Seed to be used in the procedure.
multiple	A logical value. If TRUE (not default), the resulted pvalues are adjusted by using one of several methods for multiple comparisons.
multiple.method	Correction method. See Details.

Details

The adjustment methods include the Bonferroni correction ("bonferroni") in which the p-values are multiplied by the number of comparisons. Less conservative corrections are also included by Holm (1979) ('holm'), Hochberg (1988) ('hochberg'), Hommel (1988) ('hommel'), Benjamini & Hochberg (1995) ('BH' or its alias 'fdr'), and Benjamini & Yekutieli (2001) ('BY'), respectively. A pass-through option ('none') is also included.

Value

A list containing the following items:

table	A data frame containing the null hypothesis tested, the values of the test statistic and the obtained pvalues.
levels	Original levels of the variable z.
cluster	A vector of integers (from 1:k) indicating the cluster to which each curve is allocated.
centers	An object containing the centroids (mean of the curves pertaining to the same group).
curves	An object containing the fitted curves for each population.

Author(s)

Nora M. Villanueva and Marta Sestelo.

Examples

```
library(clustcurv)

# Regression framework
res <- regclustcurves(y = barnacle5$DW, x = barnacle5$RC, z = barnacle5$F,
algorithm = 'kmeans', nboot = 2, cluster = TRUE, ncores = 2)
```

summary.clustcurves *Summarizing fits of kclustcurves class produced by
survclustcurves and regclustcurves*

Description

Takes a clustcurves object and produces various useful summaries from it.

Usage

```
## S3 method for class 'clustcurves'
summary(object, ...)
```

Arguments

object a clustcurves object as produced by survclustcurves and regclustcurves
... additional arguments.

Details

print.clustcurves tries to be smart about summary.clustcurves.

Value

summary.clustcurves computes and returns a list of summary information for a clustcurves object.

levels Levels of the factor.
cluster A vector containing the assignment of each factor's level to its group.
table A data.frame containing the results from the hypothesis test.

Author(s)

Nora M. Villanueva and Marta Sestelo.

Examples

```
library(clustcurv)
library(survival)
data(veteran)

# Survival framework
ressurv <- survclustcurves(time = veteran$time, status = veteran$status,
x = veteran$celltype, algorithm = 'kmeans', nboot = 2)

summary(ressurv)

# Regression framework
resreg <- regclustcurves(y = barnacle5$DW, x = barnacle5$RC, z = barnacle5$F,
algorithm = 'kmeans', nboot = 2)

summary(resreg)
```

summary.kcurves	<i>Summarizing fits of kcurves class produced by ksurvcurves and kregcurves</i>
-----------------	---

Description

Takes a kcurves object and produces various useful summaries from it.

Usage

```
## S3 method for class 'kcurves'
summary(object, ...)
```

Arguments

object	a kcurves object as produced by ksurvcurves and kregcurves
...	additional arguments.

Details

print.kcurves tries to be smart about summary.kcurves.

Value

summary.kcurves computes and returns a list of summary information for a kcurves object.

levels	Levels of the factor.
cluster	A vector containing the assignment of each factor's level to its group.

Author(s)

Nora M. Villanueva and Marta Sestelo.

Examples

```
library(clustcurv)
library(survival)
data(veteran)

# Survival: 2 groups k-means
s2 <- ksurvcurves(time = veteran$time, status = veteran$status,
x = veteran$celltype, k = 2, algorithm = "kmeans")

summary(s2)

# Regression: 2 groups k-means
r2 <- kregcurves(y = barnacle5$DW, x = barnacle5$RC,
z = barnacle5$F, k = 2, algorithm = "kmeans")
```

```
summary(r2)
```

survclustcurves *Clustering multiple survival curves*

Description

Function for grouping survival curves based on the k-means or k-medians algorithm. It returns the number of groups and the assignment.

Usage

```
survclustcurves(
  time,
  status = NULL,
  x,
  kvector = NULL,
  kbin = 50,
  nboot = 100,
  algorithm = "kmeans",
  alpha = 0.05,
  cluster = FALSE,
  ncores = NULL,
  seed = NULL,
  multiple = FALSE,
  multiple.method = "holm"
)
```

Arguments

time	Survival time.
status	Censoring indicator of the survival time of the process; 0 if the total time is censored and 1 otherwise.
x	Categorical variable indicating the population to which the observations belong.
kvector	A vector specifying the number of groups of curves to be checking.
kbin	Size of the grid over which the survival functions are to be estimated.
nboot	Number of bootstrap repeats.
algorithm	A character string specifying which clustering algorithm is used, i.e., k-means("kmeans") or k-medians ("kmedians").
alpha	Significance level of the testing procedure. Defaults to 0.05.

cluster	A logical value. If TRUE (default), the testing procedure is parallelized. Note that there are cases (e.g., a low number of bootstrap repetitions) that R will gain in performance through serial computation. R takes time to distribute tasks across the processors also it will need time for binding them all together later on. Therefore, if the time for distributing and gathering pieces together is greater than the time need for single-thread computing, it does not worth parallelize.
ncores	An integer value specifying the number of cores to be used in the parallelized procedure. If NULL (default), the number of cores to be used is equal to the number of cores of the machine - 1.
seed	Seed to be used in the procedure.
multiple	A logical value. If TRUE (not default), the resulted pvalues are adjusted by using one of several methods for multiple comparisons.
multiple.method	Correction method. See Details.

Details

The adjustment methods include the Bonferroni correction ("bonferroni") in which the p-values are multiplied by the number of comparisons. Less conservative corrections are also included by Holm (1979) ('holm'), Hochberg (1988) ('hochberg'), Hommel (1988) ('hommel'), Benjamini & Hochberg (1995) ('BH' or its alias 'fdr'), and Benjamini & Yekutieli (2001) ('BY'), respectively. A pass-through option ('none') is also included.

Value

A list containing the following items:

table	A data frame containing the null hypothesis tested, the values of the test statistic and the obtained pvalues.
levels	Original levels of the variable x.
cluster	A vector of integers (from 1:k) indicating the cluster to which each curve is allocated.
centers	An object containing the centroids (mean of the curves pertaining to the same group).
curves	An object containing the fitted curves for each population.

Author(s)

Nora M. Villanueva and Marta Sestelo.

Examples

```
library(clustcurv)
library(survival)
data(veteran)

# Survival framework
res <- survclustcurves(time = veteran$time, status = veteran$status,
```

```
x = veteran$celltype, algorithm = 'kmeans', nboot = 2)
```

Index

`autoplot.clustcurves`, [2](#)

`barnacle5`, [4](#)

`clustcurv`, [5](#)

`clustcurv-package (clustcurv)`, [5](#)

`ggfortify::autoplot.survfit()`, [3](#)

`kregcurves`, [6](#)

`ksurvcurves`, [7](#)

`regclustcurves`, [8](#)

`summary.clustcurves`, [10](#)

`summary.kcurves`, [12](#)

`survclustcurves`, [13](#)