

Package ‘traj’

December 7, 2023

Title Clustering of Functional Data Based on Measures of Change

Version 2.0.1

Description Implements the three-step procedure proposed by Leffondree et al. (2004) to identify clusters of individual longitudinal trajectories. The procedure involves (1) calculating a number of measures of change capturing various features of the trajectories; (2) using a Principal Component Analysis based dimension reduction algorithm to select a subset of measures and (3) using the K-means clustering algorithm to identify clusters of trajectories.

License MIT + file LICENSE

URL <https://CRAN.R-project.org/package=traj>

Encoding UTF-8

RoxygenNote 7.1.2

Imports stats, cluster, psych

Depends R (>= 2.10)

NeedsCompilation no

Author Marie-Pierre Sylvestre [aut],
Laurence Boulanger [aut, cre],
Gillis Delmas Tchouangue Dinkou [ctb],
Dan Vatnik [ctb]

Maintainer Laurence Boulanger <laurence.boulanger@umontreal.ca>

Repository CRAN

Date/Publication 2023-12-07 17:50:02 UTC

R topics documented:

plot.trajClusters	2
Step1Measures	3
Step2Selection	6
Step3Clusters	7
trajdata	9

Index	10
--------------	-----------

plot.trajClusters *Plots trajClusters objects*

Description

Up to 5 kinds of plots are currently available: a plot of the cluster-specific median and mean trajectories, a random sample of trajectories from each cluster and scatter plots of the measures on which the clustering was based. When the GAP criterion was used in Step3Clusters to determine the optimal number of clusters, a plot of the GAP statistic as a function of the number of clusters is provided.

Usage

```
## S3 method for class 'trajClusters'
plot(x, sample.size = 5, ask = TRUE, which.plots = NULL, spline = FALSE, ...)
```

Arguments

x	object of class trajClusters as returned by Step3Cluster.
sample.size	the number of random trajectories to be randomly sampled from each cluster. Defaults to 5.
ask	logical. If TRUE, the user is asked before each plot. Defaults to TRUE.
which.plots	either NULL or a vector of integers. If NULL, every available plot is displayed. If a vector is supplied, only the corresponding plots will be displayed.
spline	logical. If TRUE, each trajectory will be smoothed using smoothing splines and the median and mean trajectories will be plotted from the smoothed trajectories. Defaults to FALSE.
...	other parameters to be passed through to plotting functions.

See Also

[Step3Clusters](#)

Examples

```
## Not run:

m = Step1Measures(trajdata, ID = TRUE)
s = Step2Selection(m)
c3 = Step3Clusters(s, nclusters = 3)

plot(c3)

#The pointwise mean trajectories correspond to the third and fourth displayed plots.

c4 = Step3Clusters(s, nclusters = 4)
```

```
plot(c4, which.plots = 3:4)
```

```
## End(Not run)
```

Step1Measures	<i>Compute Measures for Identifying Patterns of Change in Longitudinal Data</i>
---------------	---

Description

Step1Measures computes up to 26 measures for each longitudinal trajectory. See Details for the list of measures.

Usage

```
Step1Measures(Data, Time = NULL, ID = FALSE, measures = 1:23, midpoint = NULL)
```

```
## S3 method for class 'trajMeasures'
print(x, ...)
```

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

Arguments

Data	a matrix or data frame in which each row contains the longitudinal data (trajectories).
Time	either NULL, a vector or a matrix/data frame of the same dimension as Data. If a vector, matrix or data frame is supplied, its entries are assumed to be measured at the times of the corresponding cells in Data. When set to NULL (the default), the times are assumed equidistant.
ID	logical. Set to TRUE if the first columns of Data and Time corresponds to an ID variable identifying the trajectories. Defaults to FALSE.
measures	a vector containing the numerical identifiers of the measures to compute (see "Details" section below). The default, 1:23, corresponds to measures 1-23 and thus excludes the measures which require specifying a midpoint.
midpoint	specifies which column of Time to use as the midpoint in measures 24-26. Can be NULL, an integer or a vector of integers of length the number of rows in Time. The default is NULL, in which case the midpoint is the time closest to the median of the Time vector specific to each trajectory.
x	object of class trajMeasures.
...	further arguments passed to or from other methods.
object	object of class trajMeasures.

Details

Each trajectory must have a minimum of 3 observations otherwise it will be omitted from the analysis.

The 26 measures and their numerical identifiers are listed below. Please refer to the vignette for the specific formulas used to compute them.

1. Range
2. Mean of the function
3. Functional standard deviation (SD)
4. Coefficient of variation (ratio of measure 3 to measure 2)
5. Overall change (initial value - final value)
6. Mean change per unit time
7. Overall change relative to initial value
8. Overall change relative to functional mean (ratio of measure 5 to measure 2)
9. Slope of the linear model
10. R^2 : Proportion of variance explained by the linear model
11. Maximum value of the speed
12. Functional SD of the speed
13. Mean absolute speed
14. Maximum absolute speed
15. Maximum absolute speed relative to the functional mean (ratio of measure 14 to measure 2)
16. Maximum absolute speed relative to the slope (ratio of measure 14 to measure 9)
17. Functional SD of the speed relative to the slope
(ratio of measure 12 to measure 9)
18. Mean acceleration

19. Mean absolute acceleration
20. Maximum of the absolute acceleration
21. Maximum of the absolute acceleration relative to the functional (ratio of measure 20 to measure 2)
22. Maximum of the absolute acceleration relative to the mean absolute speed (ratio of measure 20 to measure 13)
23. Mean absolute acceleration relative to the mean absolute speed (ratio of measure 19 to measure 13)
24. Early change relative to later change
25. Early change relative to overall change
26. Later change relative to overall change

Value

An object of class `trajMeasures`; a list containing the values of the measures, a table of the outliers which have been capped, as well as a curated form of the function's arguments.

References

Leffondre K, Abrahamowicz M, Regeasse A, Hawker GA, Badley EM, McCusker J, Belzile E. Statistical measures were proposed for identifying longitudinal patterns of change in quantitative health indicators. *J Clin Epidemiol.* 2004 Oct;57(10):1049-62. doi: 10.1016/j.jclinepi.2004.02.012. PMID: 15528056.

Nishiyama T, Improved Chebyshev inequality: new probability bounds with known supremum of PDF, arXiv:1808.10770v2 stat.ME <https://doi.org/10.48550/arXiv.1808.10770>

Examples

```
## Not run:
m1 = Step1Measures(trajdata, ID = TRUE, measures = 24:26, midpoint = NULL)
m2 = Step1Measures(trajdata, ID = TRUE, measures = 24:26, midpoint = 3)

identical(s1$measures, s2$measures)

## End(Not run)
```

Description

This function applies the following dimension reduction algorithm to the measures computed by [Step1Measures](#):

1. Use principal component analysis (PCA) on the measure to form factors summarizing the variability in the measures;
2. Drop the factors whose variance is smaller than any one of the normalized measures;
3. Performs a varimax rotation on the remaining factors;
4. For each rotated factor, select the measure that has the highest correlation (aka factor loading) with it and that hasn't yet been selected;
5. Drop the remaining measures.

Usage

```
Step2Selection(trajMeasures, num.select = NULL, discard = NULL, select = NULL)
```

```
## S3 method for class 'trajSelection'
print(x, ...)
```

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

Arguments

trajMeasures	object of class trajMeasures as returned by Step1Measures .
num.select	an optional positive integer indicating the number of factors to keep in the second stage of the algorithm. Defaults to NULL so that all factors with variance greater than any one of the normalized measures are selected.
discard	an optional vector of positive integers corresponding to the measures to be dropped from the analysis. See Step1Measures for the list of measures. Defaults to NULL.
select	an optional vector of positive integers corresponding to the measures to forcefully select. Defaults to NULL. If a vector is supplied, the five-steps selection algorithm described above is bypassed and the corresponding measures are selected instead. Can be NULL or a numeric vector corresponding to the numerical identifier of measures present in trajMeasures. If a numeric vector is supplied, then four-steps selection algorithm described above is bypassed and the corresponding measures are selected instead.
x	object of class trajSelection.
...	further arguments passed to or from other methods.
object	object of class trajSelection.

Details

In the presence of highly correlated measures (Pearson correlation > 0.98), the function selects the highest-ranking measure on the list (see [Step1Measures](#)) and discards the others. Because the K-means algorithm is sensitive to outliers, measures which are quotients (i.e. 4, 7, 8, 15-17, 21-26) are prevented from taking extremely large or infinite values (caused by division by 0). Nishiyama's improved Chebychev bound is used to determine extreme values for each measure, corresponding to a 0.3% probability threshold. Extreme values beyond the threshold are capped to the 0.3% probability threshold. Measures corresponding to quotients which would be of the form 0/0 are set to 1. PCA is applied on the remaining measures using the [principal](#) function from the psych package.

Value

An object of class trajSelection; a list containing the values of the selected measures, the output of the principal component analysis as well as a curated form of the arguments.

References

Leffondre K, Abrahamowicz M, Regeasse A, Hawker GA, Badley EM, McCusker J, Belzile E. Statistical measures were proposed for identifying longitudinal patterns of change in quantitative health indicators. J Clin Epidemiol. 2004 Oct;57(10):1049-62. doi: 10.1016/j.jclinepi.2004.02.012. PMID: 15528056.

See Also

[principal](#) [Step1Measures](#)

Examples

```
## Not run:
m = Step1Measures(trajdata, ID = TRUE)
s = Step2Selection(m)

s$RC$loadings

s2 = Step2Selection(m, select = c(10, 12, 8, 4))

## End(Not run)
```

Description

Classifies the trajectories by applying the K-mean clustering algorithm to the measures selected by Step2Selection.

Usage

```

Step3Clusters(
  trajSelection,
  nclusters = NULL,
  nstart = 50,
  iter.max = 20,
  K.max = 8,
  B = 500,
  d.power = 2,
  spaceH0 = "scaledPCA",
  method = "Tibs2001SEmax",
  SE.factor = 1
)

## S3 method for class 'trajClusters'
print(x, ...)

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

```

Arguments

<code>trajSelection</code>	object of class <code>trajSelection</code> as returned by <code>Step2Selection</code> .
<code>nclusters</code>	either <code>NULL</code> or the desired number of clusters. If <code>NULL</code> , the number of clusters is determined using the GAP criterion as implemented in the clusGap function.
<code>nstart</code>	to be passed to the <code>nstart</code> argument of kmeans .
<code>iter.max</code>	to be passed to the <code>iter.max</code> argument of kmeans .
<code>K.max</code>	to be passed to the <code>K.max</code> argument of clusGap .
<code>B</code>	to be passed to the <code>SE.factor</code> argument of clusGap .
<code>d.power</code>	to be passed to the <code>B</code> argument of clusGap .
<code>spaceH0</code>	to be passed to the <code>spaceH0</code> argument of clusGap .
<code>method</code>	to be passed to the <code>method</code> argument of clusGap .
<code>SE.factor</code>	to be passed to the <code>SE.factor</code> argument of clusGap .
<code>x</code>	object of class <code>trajClusters</code>
<code>...</code>	further arguments passed to or from other methods.
<code>object</code>	object of class <code>trajClusters</code>

Value

An object of class `trajClusters`; a list containing the result of the clustering, the output of the `clusGap` function, as well as a curated form of the arguments.

See Also

[Step2Selection](#)

Examples

```
## Not run:

m = Step1Measures(trajdata, ID = TRUE)
s = Step2Selection(m)

s$RC$loadings

s2 = Step2Selection(m, select = c(10, 12, 8, 4))

c3.part <- Step3Clusters(s2, nclusters = 3)$partition
c4.part <- Step3Clusters(s2, nclusters = 4)$partition
c5.part <- Step3Clusters(s2, nclusters = 5)$partition

## End(Not run)
```

trajdata

*Fictitious functional data for illustration purposes in traj.***Description**

Fictitious functional data for illustration purposes in traj.

Usage

```
data("trajdata")
```

Format

This data set contains 130 rows and 7 columns, the first of which is an ID variable. The other columns represent six measurements $f(x_1), \dots, f(x_6)$ of some function f made at $x_1 < \dots < x_6$.

Index

* datasets

trajdata, [9](#)

clusGap, [8](#)

kmeans, [8](#)

plot.trajClusters, [2](#)

principal, [7](#)

print.trajClusters (Step3Clusters), [7](#)

print.trajMeasures (Step1Measures), [3](#)

print.trajSelection (Step2Selection), [6](#)

Step1Measures, [3](#), [6](#), [7](#)

Step2Selection, [6](#), [8](#)

Step3Clusters, [2](#), [7](#)

summary.trajClusters (Step3Clusters), [7](#)

summary.trajMeasures (Step1Measures), [3](#)

summary.trajSelection (Step2Selection),
[6](#)

trajdata, [9](#)