

Package ‘traj’

March 16, 2026

Title Feature-Based Clustering of Longitudinal Trajectories

Version 3.0.1

Description Identifies clusters of individual longitudinal trajectories. In the spirit of Leffondre et al. (2004), the procedure involves identifying each trajectory to a point in the space of measures. In this context, a measure is a quantity meant to capture a certain characteristic feature of the trajectory. The points in the space of measures are then clustered using a version of the Spectral Clustering algorithm.

License MIT + file LICENSE

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

Encoding UTF-8

RoxygenNote 7.3.2

Suggests knitr, rmarkdown, testthat (>= 3.0.0)

Config/testthat/edition 3

Imports stats, cluster, clusterCrit, fclust, igraph, e1071

Depends R (>= 2.10)

LazyData true

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 2026-03-16 01:00:02 UTC

Contents

plot.trajClusters	2
trajClusters	3
trajdata	5
trajMeasures	6
trajReduce	8

plot.trajClusters *Plot trajClusters object*

Description

Plots the curves corresponding to (or closest to) the centroids of the clusters and plots a random sample from each groups.

Usage

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

scatterplots(x, ask = TRUE, which.scatter = NULL, N = NULL, ...)

CVIplot(x, ...)
```

Arguments

x	object of class trajClusters as returned by the function trajClusters().
sample.size	the number of trajectories to be randomly sampled from each clusters. If NULL, all the trajectories are included in a random order. 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.
...	other parameters to be passed through to plotting functions.
which.scatter	either NULL or a vector of integers that is a subset of the measure argument used in function trajClusters to produce object x. If NULL, every available scatter plots are displayed. If a vector is supplied, only the corresponding plots will be displayed.
N	the maximum number of points present in each scatter plots. If a non NULL value is specified, N points are sampled randomly in a way that preserves the relative groups sizes. If NULL (the default), all the points are plotted.

Examples

```
## Not run:
data("trajdata")
trajdata.noGrp <- trajdata[, -which(colnames(trajdata) == "Group")] #remove the Group column

m = trajMeasures(trajdata.noGrp, ID = TRUE)
c3 = trajClusters(m, nclusters = 3)

plot(c3)
```

```
#The pointwise mean trajectories correspond to the third and fourth displayed plots.  
c4 = trajClusters(m, nclusters = 4)  
  
plot(c4, which.plots = 3:4)  
  
## End(Not run)
```

trajClusters

Classify the Longitudinal Data Based on the Measures.

Description

Classifies the trajectories by applying a nonparametric clustering algorithm to the measures computed by `trajMeasures()`.

Usage

```
trajClusters(  
  Measures,  
  select = NULL,  
  fuzzy = FALSE,  
  nclusters = NULL,  
  subset.n = NULL,  
  nstart = 50  
)  
  
## S3 method for class 'trajClusters'  
print(x, ...)  
  
## S3 method for class 'trajClusters'  
summary(object, ...)
```

Arguments

Measures	object of class <code>trajMeasures</code> as returned by the function <code>trajMeasures()</code> .
select	an optional vector of positive integers corresponding to the measures to use in the clustering. Defaults to <code>NULL</code> , which uses all the measures contained in <code>Measures</code> .
fuzzy	logical. If <code>FALSE</code> , each trajectory is assigned to a unique group. If <code>TRUE</code> , each trajectory is assigned a "degree of membership" to each group. Defaults to <code>FALSE</code> .

nclusters	The desired number of clusters. If NULL, clustering is carried out for every number of clusters between 2 and (up to) 8 and the "best" number of clusters is used, as judged by the combination of three internal cluster validity indices. See section 'Value' for more details. Defaults to NULL.
subset.n	A positive numerical integer smaller than the number of trajectories. If nclusters is NULL, subset is the number of trajectories, randomly sampled from the complete data set, that will be used to determine the optimal number of clusters in the interest of speeding up the process.
nstart	The number of random starts. Defaults to 50.
x	object of class trajClusters.
...	further arguments passed to or from other methods.
object	object of class trajClusters.

Details

The spectral clustering algorithm presented in Meila (2005) is implemented in which the similarity matrix S is built from a binary K nearest neighbors similarity function ($S = (W + W^T)/2$, where $W_{ij} = 1$ if data point j is among the nearest points to data point i and $W_{ij} = 0$ otherwise).

Value

An object of class trajClusters; a list containing the result of the clustering, as well as a curated form of the arguments. If nclusters is set to NULL, clustering is carried out for each number k of clusters between 2 and (up to) 8 and a plot is produced representing the value of three internal cluster validity indices (C-index, Calinski-Harabasz, Wemmert-Gancarski) as a function of k . As in the 'KmL' package of Genolini et al., these validity indices are presented on a scale from 0 to 1, with 1 corresponding to the highest validity score and 0 corresponding to the lowest. From this, a "best" value of k is determined using a ranked voting system.

References

Genolini, C. et al., kml: K-Means for Longitudinal Data, <https://CRAN.R-project.org/package=kml>
 Meila, M., Spectral Clustering. Handbook of Cluster Analysis, Chapter 7, Chapman and Hall/CRC, 2005.

Examples

```
## Not run:
data("trajdata")
trajdata.noGrp <- trajdata[, -which(colnames(trajdata) == "Group")] # remove the Group column

m = trajMeasures(trajdata.noGrp, ID = TRUE, measures = 1:19)

s2.3 <- trajClusters(m, nclusters = 3)
plot(s2.3)

#'s2.4 <- trajClusters(m, nclusters = 4)
plot(s2.4)
```

```
#'s2.5 <- trajClusters(m, nclusters = 5)
plot(s2.5)

groups <- s2.4 <- trajClusters(m, nclusters = 4)$partition

## End(Not run)
```

trajdata

trajdata

Description

An artificially created data set with 130 trajectories split into four groups, labelled A, B, C, D according to the data generating process.

Usage

```
trajdata
```

Format

This data frame has 130 rows and the following 7 columns:

ID An identification variable that runs from 1 to 130.

Group A character variable that's either "A", "B", "C" or "D" depending on which of the four data generating process the trajectory is coming from.

X1 The observation of the trajectory at time $t = 1$.

X2 The observation of the trajectory at time $t = 2$.

X3 The observation of the trajectory at time $t = 3$.

X4 The observation of the trajectory at time $t = 4$.

X5 The observation of the trajectory at time $t = 5$.

X6 The observation of the trajectory at time $t = 6$.

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

Description

trajMeasures computes up to 20 measures for each longitudinal trajectory. See Details for the list of measures.

Usage

```
trajMeasures(
  Data,
  Time = NULL,
  ID = FALSE,
  measures = c(1:10, 12:20),
  midpoint = NULL,
  cap.outliers = FALSE
)

## 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. The default, c(1:10,12:20), excludes the measure which require specifying a midpoint.
midpoint	specifies which column of Time to use as the midpoint in measure 11 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.
cap.outliers	logical. If TRUE, extreme values of the measures will be capped. Defaults to FALSE.

x	object of class <code>trajMeasures</code> .
...	further arguments passed to or from other methods.
object	object of class <code>trajMeasures</code> .

Details

Each trajectory must have a minimum of 3 observations, otherwise it is omitted from the analysis. The 20 measures and their numerical identifiers are listed below. Please refer to the vignette for the specific formulas used to compute them.

1. Maximum
2. Minimum
3. Range
4. Mean value
5. Standard deviation
6. Slope of the affine approximation
7. Intercept of the affine approximation
8. Proportion of variance explained by the affine approximation
9. Rate of intersection with the best affine approximation
10. Net variation per unit of time
11. Late variation to early variation contrast
12. Total variation per unit time
13. Spikiness
14. Maximum of the first derivative
15. Minimum of the first derivative
16. Standard deviation of the first derivative
17. First derivative's net variation per unit of time

18. Maximum of the second derivative
19. Minimum of the second derivative
20. Standard deviation of the second derivative

If 'cap.outliers' is set to TRUE, Nishiyama's improved Chebychev bound for continuous distributions is used to determine extreme values for each measure, corresponding to a 0.3% probability threshold. Extreme values beyond the threshold are then capped to the 0.3% probability threshold (see vignette for more details).

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:
data("trajdata")
trajdata.noGrp <- trajdata[, -which(colnames(trajdata) == "Group")] #remove the Group column

m1 = trajMeasures(trajdata.noGrp, ID = TRUE, measures = 19, midpoint = NULL)
m2 = trajMeasures(trajdata.noGrp, ID = TRUE, measures = 19, midpoint = 3)

identical(m1$measures, m2$measures)

## End(Not run)
```

Description

This function examines the effect of reducing the number of measures on which the trajectories are clustered. Specifically, starting from a clustering C in the form of an object of class `trajClusters` and a choice of a similarity index to compare clusterings, this function finds the subset of measures which results in the clustering most similar to C .

Usage

```
trajReduce(Measures, Clusters, index = "ARI", keep = 3)
```

Arguments

Measures	object of class <code>trajMeasures</code> as returned by trajMeasures .
Clusters	object of class <code>trajClusters</code> as returned by trajClusters .
index	The similarity index. Either "ARI" for the Adjusted Rand Index of Hubert and Arabie (1985), "nVId" for the normalized variation of information distance (eg. Meila (2007)) or "nSJd" for the normalized split/joint distance of van Dongen (2000).
keep	The number of measures to keep. Defaults to 3.

Details

The Rand index ranges from 0 to 1 with 0 indicating identical clusters and 1 indicating maximally different clusters. The normalized variation of information distance (nVId) and normalized split-join distance (nSJd) and have the opposite interpretation with 0 indicating maximally different clusters and 1 indicating identical clusters. Therefore, to facilitate comparison, we plot $1 - \text{nVId}$ (resp. $1 - \text{nSJd}$) instead of nVId (resp. nSJd).

References

Hubert L, Arabie P. Comparing partitions. *Journal of Classification* 2:193-218, 1985.

Meila M. Comparing clusterings – an information based distance. *Journal of Multivariate Analysis*, 98, pp 873-895, 2007.

van Dongen S. Performance criteria for graph clustering and Markov cluster experiments. Technical Report INS-R0012, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam, May 2000.

Examples

```
## Not run:
data("trajdata")
trajdata.noGrp <- trajdata[, -which(colnames(trajdata) == "Group")] #remove the Group column

m = trajMeasures(trajdata.noGrp, ID = TRUE)
trajReduce(m)

## End(Not run)
```

Index

* datasets

trajdata, [5](#)

CVIplot (plot.trajClusters), [2](#)

plot.trajClusters, [2](#)

print.trajClusters (trajClusters), [3](#)

print.trajMeasures (trajMeasures), [6](#)

scatterplots (plot.trajClusters), [2](#)

summary.trajClusters (trajClusters), [3](#)

summary.trajMeasures (trajMeasures), [6](#)

trajClusters, [3](#), [9](#)

trajdata, [5](#)

trajMeasures, [6](#), [9](#)

trajReduce, [8](#)