Package 'traj'

July 5, 2024

Title Clustering of Functional Data Based on Measures of Change

Version 2.2.0

Description Implements a three-step procedure in the spirit of Leffondre et al. (2004) to identify clusters of individual longitudinal trajectories. The procedure involves (1) computing 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 7.3.2

Suggests knitr, rmarkdown, testthat (>= 3.0.0)

Config/testthat/edition 3

Imports stats, cluster, psych

Depends R (>= 2.10)

LazyData true

VignetteBuilder knitr

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 2024-07-04 22:40:05 UTC

Contents

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

m = Step1Measures(trajdata.noGrp, 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` *Compute Measures for Identifying Patterns of Change in Longitudinal Data*

Description

`Step1Measures` computes up to 18 measures for each longitudinal trajectory. See Details for the list of measures.

Usage

```
Step1Measures(
  Data,
  Time = NULL,
  ID = FALSE,
  measures = c(1:17),
  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, 1:17, corresponds to measures 1-17 and thus excludes the measures which require specifying a midpoint.
midpoint	specifies which column of Time to use as the midpoint in measure 18. 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. If FALSE, only the infinite values will be capped. Defaults to FALSE.
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 18 measures and their numerical identifiers are listed below. Please refer to the vignette for the specific formulas used to compute them.

1. Maximum
2. Range (max - min)
3. Mean value
4. Standard deviation
5. Slope of the linear model
6. R^2: Proportion of variance explained by the linear model
7. Curve length (total variation)
8. Rate of intersection with the mean
9. Proportion of time spent above the mean
10. Minimum of the first derivative
11. Maximum of the first derivative

12. Mean of the first derivative
13. Standard deviation of the first derivative
14. Minimum of the second derivative
15. Maximum of the second derivative
16. Mean of the second derivative
17. Standard deviation of the second derivative
18. Later change/Early change

If 'cap.outliers' is set to TRUE, or if some measures are infinite as a result of division by 0, 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). If applicable, the values which would be of the form 0/0 are set to 1.

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 = Step1Measures(trajdata.noGrp, ID = TRUE, measures = 18, midpoint = NULL)
m2 = Step1Measures(trajdata.noGrp, ID = TRUE, measures = 18, midpoint = 3)

identical(m1$measures, m2$measures)

## End(Not run)
```

Step2Selection *Select a Subset of the Measures Using Factor Analysis*

Description

This function applies the following dimension reduction algorithm to the measures computed by Step1Measures:

1. Drop the measures whose values are constant across the trajectories;
2. Whenever two measures are highly correlated (absolute value of Pearson correlation > 0.98), keep the highest-ranking measure on the list (see Step1Measures) and drop the other;
3. Use principal component analysis (PCA) on the measures to form factors summarizing the variability in the measures;
4. Drop the factors whose variance is smaller than any one of the standardized measures;
5. Perform a varimax rotation on the remaining factors;
6. For each rotated factor, select the measure that has the highest correlation (aka factor loading) with it and that hasn't yet been selected;
7. 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.
x	object of class trajSelection.
...	further arguments passed to or from other methods.
object	object of class trajSelection.

Details

Whenever two measures are highly correlated (Pearson correlation > 0.98), the highest-ranking measure on the list (see `Step1Measures`) is kept and the other is discarded and discards the others. 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:
data("trajdata")
trajdata.noGrp <- trajdata[, -which(colnames(trajdata) == "Group")] #remove the Group column

m = Step1Measures(trajdata.noGrp, measure = c(1:18), ID = TRUE)
s = Step2Selection(m)

print(s)

s2 = Step2Selection(m, select = c(13, 3, 12, 9))

## End(Not run)
```

`Step3Clusters` *Classify the Longitudinal Data Based on the Selected Measures.*

Description

Classifies the trajectories by applying the k-means clustering algorithm to the measures selected by `Step2Selection`.

Usage

```
Step3Clusters(
  trajSelection,
  algorithm = "k-medoids",
  metric = "euclidean",
  nstart = 200,
  iter.max = 100,
  nclusters = NULL,
  criterion = "Calinski-Harabasz",
  K.max = min(15, nrow(trajSelection$selection) - 1),
  boot = FALSE,
  R = 100,
  B = 500
)

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

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

Arguments

trajSelection	object of class trajSelection as returned by Step2Selection.
algorithm	either "k-medoids" or "k-means". Determines the clustering algorithm to use. Defaults to "k-medoids".
metric	to be passed to the metric argument of pam if "k-medoids" is the chosen algorithm. Defaults to "euclidean".
nstart	to be passed to the nstart argument of kmeans if "k-means" is the chosen algorithm. Defaults to 200.
iter.max	to be passed to the iter.max argument of kmeans if "k-means" is the chosen algorithm. Defaults to 100.
nclusters	either NULL or the desired number of clusters. If NULL, the number of clusters is determined using the criterion chosen in criterion. Defaults to NULL.
criterion	criterion to determine the optimal number of clusters if nclusters is NULL. Either "GAP" or "Calinski-Harabasz". Defaults to "Calinski-Harabasz".
K.max	maximum number of clusters to be considered if nclusters is set to NULL. Defaults to 15.
boot	logical. If TRUE, and if "Calinski-Harabasz" is the chosen criterion, the optimal number of clusters will be the first mode of sampling distribution of the optimal number of clusters obtained by bootstrap. Defaults to FALSE.
R	the number of bootstrap replicate if boot is set to TRUE. Defaults to 100.
B	to be passed to the B argument of clusGap if "GAP" is the chosen criterion.
x	object of class trajClusters.
...	further arguments passed to or from other methods.
object	object of class trajClusters.

Details

If `"GAP"` is the chosen `criterion` for determining the optimal number of clusters, the method described by Tibshirani et al. is implemented by the `clusGap` function.

Instead, if `"Calinski-Harabasz"` is the chosen `criterion`, the Calinski-Harabasz index is computed for each possible number of clusters between 2 and `K.max` and the optimal number of clusters is the maximizer of the Calinski-Harabasz index. Moreover, if `boot` is set to `TRUE`, then, following the guidelines suggested by Mesidor et al., a sampling distribution of the optimal number of clusters is obtained by bootstrap and the optimal number of clusters is chosen to be the (first) mode of this sampling distribution.

Value

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

References

Miceline Mésidor, Caroline Sirois, Marc Simard, Denis Talbot, A Bootstrap Approach for Evaluating Uncertainty in the Number of Groups Identified by Latent Class Growth Models, American Journal of Epidemiology, Volume 192, Issue 11, November 2023, Pages 1896–1903, https://doi.org/10.1093/aje/kwad148

Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of data clusters via the Gap statistic. Journal of the Royal Statistical Society B, 63, 411–423.

Tibshirani, R., Walther, G. and Hastie, T. (2000). Estimating the number of clusters in a dataset via the Gap statistic. Technical Report. Stanford.

See Also

`Step2Selection`

Examples

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

m = Step1Measures(trajdata.noGrp, ID = TRUE, measures = 1:18)
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` *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.

Index