Package ‘CluMP’

October 12, 2022

Title Clustering of Micro Panel Data
Version 0.8.1

Depends R (>= 3.4.0)
License GPL (>= 3)
Encoding UTF-8
LazyData true
RoxygenNote 7.1.1
Imports MASS, ggplot2 (>= 3.0.0), dplyr (>= 0.7.6), NbClust (>= 3.0),
amap (>= 0.8-16), tableone, stats, data.table, rlang
Suggests knitr, rmarkdown
NeedsCompilation no
Author Jan Fojtik [aut, cre],
Anna Grishko [aut],
Lukas Sobisek [aut, cph, rev]
Maintainer Jan Fojtik <9afojtik@gmail.com>
Repository CRAN
Date/Publication 2020-11-27 02:20:29 UTC

R topics documented:

CluMP ................................................................. 2
CluMP_profiles .................................................. 3
CluMP_view ....................................................... 4
GeneratePanel ..................................................... 5
OptiNum ........................................................... 7
PanelPlot ......................................................... 9
Cluster Micro-Panel (longitudinal) Data employing the CluMP algorithm

Description

This function clusters Micro-Panel (longitudinal) Data (or trajectories) to a pre-defined number of clusters by employing Feature-Based Clustering of Micro-Panel (longitudinal) Data algorithm called CluMP (see Reference). Currently, only univariate clustering analysis is available.

Usage

CluMP(formula, group, data, cl_numb = NA, base_val = FALSE, method = "ward.D")

Arguments

- formula: A two-sided formula object with a numeric clustering variable (Y) on the left of a ~ separator and the time (numeric) variable on the right. Time is measured from the start of the follow-up period (baseline). Any time units are possible.
- group: A grouping factor variable (vector), i.e. single identifier for each individual (trajectory).
- data: A data frame containing the variables named in the formula and group arguments.
- cl_numb: An integer, positive number (scalar) specifying the number of clusters. The OptiNum function can be used to determine the optimal number of clusters according to common evaluation criteria (indices).
- base_val: Indicates whether include a value at zero time point as an additional clustering variable. Default is FALSE and the standard number (7) of clustering parameters is used.
- method: A method which use in hierarhical clustering, same as in hclust function, namely "ward.D", "ward.D2", "single", "complete", "average", "mcquitty", "median", "centroid". Default is "ward.D".

Value

Cluster Micro-Panel data. The output is the list of 5 components which contain results from clustering.
source

Examples

data <- GeneratePanel(n = 100, Param = ParamLinear, NbVisit = 10)
CluMP(formula = Y ~ Time, group = "ID", data = data, cl_numb = 3,
      base_val = FALSE, method = "ward.D")

CluMP(formula = Y ~ Time, group = "ID", data = data, cl_numb = 3,
      base_val = TRUE, method = "ward.D")

CluMP_profiles

summary characteristics of identified clusters via CluMP

Description

The function CluMP_profiles provides a description (profile) for each cluster. The description is in the form of a summary list containing descriptive statistics of a cluster variable, time variable, cluster parameters and other variables (covariates), both continuous and categorical.

Usage

CluMP_profiles(CluMPoutput, cont_vars = NULL, cat_vars = NULL, show_NA = FALSE)

Arguments

CluMPoutput An object (output) from the CluMP function.
cont_vars An optional single character or a character vector of continuous variables’ names (from the original dataset).
cat_vars An optional single character or a character vector of categorical variables’ names (from the original dataset).
show_NA Logical scalar. Should be calculated and shown descriptive statistics for NA cluster if exists? Default is FALSE. NA cluster gathers improper individuals (trajectories with < 3 not missing observations) for longitudinal clustering.

Value

Returns a list with cluster variable (Y) summary, both baseline and changes; time and a summary of the number of observations (visits); clustering parameters summary and optional continuous variables summary (baseline and changes) and categorical variables summary (baseline and end).
Examples

set.seed(123)
dataMale <- GeneratePanel(n = 50, Param = ParamLinear, NbVisit = 10)
dataMale$Gender <- "M"
dataFemale <- GeneratePanel(n = 50, Param = ParamLinear, NbVisit = 10)
dataFemale$ID <- dataFemale$ID + 50
dataFemale$Gender <- "F"
data <- rbind(dataMale, dataFemale)

CluMPoutput <- CluMP(formula = Y ~ Time, group = "ID", data = data, cl_numb = 3)
CluMP_profiles(CluMPoutput, cat_vars = "Gender")

CluMP_view

Cluster profiles’ (CluMP results) visualisation

Description

This graphical function enables to visualise cluster profiles (mean representatives of each cluster). Available are three types of plots: non-parametric (LOESS method for small/medium or GAM method for complex data of large size. Both methods are applied from ggplot2 representatives (mean within-cluster trajectories) with/without all individual (original) trajectories, and nonparametric mean trajectories with error bars.

Usage

CluMP_view(
  CluMPoutput,
  type = "all",
  nb_intervals = NULL,
  return_table = FALSE,
  title = NULL,
  x_title = NULL,
  y_title = NULL,
  plot_NA = FALSE
)

Arguments

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CluMPoutput</td>
<td>An object (output) from the CluMP function.</td>
</tr>
<tr>
<td>type</td>
<td>String. Indicates which type of graph is required. Possible values for this argument are: &quot;all&quot; (plots all data with non-parametric mean trajectories), &quot;cont&quot; (only non-parametric mean trajectories) or &quot;breaks&quot; (mean trajectories with error bars).</td>
</tr>
<tr>
<td>nb_intervals</td>
<td>An integer, positive number (scalar) specifying the number of regular timepoints into which should be follow-up period split. This argument works only with graph type = &quot;breaks&quot;. In case of other graph types the argument is ignored.</td>
</tr>
</tbody>
</table>
The number of error bars is equal to the number of timepoints specified by this argument.

**return_table** Logical scalar indicating if the summary table of plotted values in the graph of type = "breaks" should be returned. Default is FALSE.

**title** String. Optional title for a plot. If undefined, no title will used.

**x_title** String. An optional title for x axis. If undefined, the variable name after ~ in formula will used.

**y_title** String. An optional title for y axis. If undefined, the variable name before ~ in formula will used.

**plot_NA** Plot NA cluster if exists. Default is FALSE. NA cluster gathers improper individuals (< 3 observations) for longitudinal clustering.

**Value**

Returns graph for type "all" and "cont" or (list with) graph and table of mean trajectories (if specified) for type = "breaks".

**Examples**

```r
set.seed(123)
dataMale <- GeneratePanel(n = 50, Param = ParamLinear, NbVisit = 10)
dataMale$Gender <- "M"
dataFemale <- GeneratePanel(n = 50, Param = ParamLinear, NbVisit = 10)
dataFemale$ID <- dataFemale$ID + 50
dataFemale$Gender <- "F"
data <- rbind(dataMale, dataFemale)

CluMPoutput <- CluMP(formula = Y ~ Time, group = "ID", data = data, cl_numb = 3)
title <- "Plotting clusters' representatives with error bars"
CluMP_view(CluMPoutput, type = "all", return_table = TRUE)
CluMP_view(CluMPoutput, type = "cont")
CluMP_view(CluMPoutput, type = "breaks", nb_intervals = 5, return_table=TRUE, title = title)
```

**GeneratePanel**

*Generate an artificial Micro-Panel (longitudinal) Data*

**Description**

This function creates artificial linear or non-linear micro-panel (longitudinal) data coming from generating process with a certain function (linear, quadratic, cubic, exponencial) set of parameters (fixed and random (intercept, slope) effects of time).
Usage

GeneratePanel(
  n,
  Param,
  NbVisit,
  VisitFreq = NULL,
  TimeVar = NULL,
  RegModel = NULL,
  ClusterProb = NULL,
  Rho = NULL,
  units = NULL
)

Arguments

n
An integer specifying the number of individuals (trajectories) being observed.

Param
Object of data.frame containing regression parameters for each cluster. The dimensions are the various number of generating clusters and the fixed number of parameters. The second dimension (the fixed number of parameters) is given by the type of regression model specified by the argument "RegModel". For more information about the parameters, see documentation of: ParamLinear for linear model, ParamQuadrat for quadratic, ParamCubic for cubic model and ParamExpon for exponential model.

NbVisit
A positive integer numeric input defining expected number of visits. Option is Fixed or Random. Number of visits given by the argument VisitFreq. If VisitFreq is Fixed, the NbVisits defines exact number of visits for all individuals. If VisitFreq is Random then each individual has different number of visits. The number of visits is then generated from the poisson distribution with the mean (lambda) equal to NbVisits.

VisitFreq
String that defines the frequency of visits for each individual. Option is Random or Fixed. If set to Fixed or not defined, each individual has the same number of visits given by NbVisits. If set as Random the number of visits is generated from poisson distribution for each individual with the mean equal to the argument NbVisits. For example if this parameter is set as 5 then the random integer from interval of -5 to 5 is drawn and added to the time variable. Make sure that TimeVar must be lower then the number of days in parameter units.

TimeVar
A positive integer representing daily, time variability of the occurrence of repeated measurement (timepoint) from the regular, fixed occurrence (visit) given by the argument units. For example, if this argument is set to 5 then the random integer from interval of -5 to 5 is drawn and added to the time variable. TimeVar must be lower than the regular frequency of repeat measurement given by the argument units.

RegModel
String specifying the mathematical function for generating trajectory for each of n individuals. Options are linear, quadratic, cubic or exponential. If set to linear or not defined, then each trajectory has a linear trend. If set to quadratic, then each trajectory has a quadratic development in time. If set to cubic then each
trajectory has cubic development. If set to exponential, then each trajectory has exponential development.

ClusterProb: Numeric scalar (for 2 clusters) or a vector of numbers (for >2 clusters) defining the probability of each cluster. If not defined, then each cluster has the same occurrence probability.

Rho: A numeric scalar specifying autocorrelation parameter with the values from range 0 to 1. If set as 0 or not define then there is no autocorrelation between the within-individual repeated observations.

units: String defining the units of time series. Options are day, week, month or year.

Value

Generates artificial panel data.

Examples

```r
set.seed(123)
#Simple Linear model where each individual has 10 observations.
data <- GeneratePanel(n = 100, Param = ParamLinear, NbVisit = 10)

#Exponential model where each individual has 10 observations.
data <- GeneratePanel(100, ParamExpon, NbVisit = 10, VisitFreq = "Fixed", RegModel = "exponential")
PanelPlot(data)

#Cubic model where each individual has random number of observations on daily basis. #Average number of observation is given by parameter NbVisit.
data <- GeneratePanel(n = 100, Param = ParamCubic, NbVisit = 100, RegModel = "cubic", units = "day")
PanelPlot(data)

#Quadratic model where each individual has random number of observations.
#Each object is observed weekly with variability 2 days.
data <- GeneratePanel(5, ParamQuadrat, NbVisit=50, RegModel="quadratic",units="week",TimeVar=2)
PanelPlot(data)

#Generate panel data with linear trend with 75% objects in first cluster and 25% in the second.
data <- GeneratePanel(n = 100, Param = ParamLinear, NbVisit = 10, ClusterProb = c(0.75, 0.25))
PanelPlot(data, colour = "Cluster")
```

Finding an optimal number of clusters

Description

This function finds optimal number of clusters based on evaluation criteria (indices) available from the NbClust package.
Usage

OptiNum(
    formula,
    group,
    data,
    index = c("silhouette", "ch", "db"),
    max_clust = 10,
    base_val = FALSE
)

Arguments

formula A two-sided formula object, with a numeric, clustering variable (Y) on the left of a ~ separator and the time (numeric) variable on the right. Time is measured from the start of the follow-up period (baseline).

group A grouping factor variable (vector), i.e. single identifier for each individual (trajectory).

data A data frame containing the variables named in formula and group arguments.

index String vector of indices to be computed. Default is c("silhouette", "ch", "db"). See NbClust package for available indices and their description.

max_clust An integer, positive number (scalar) defining the maximum number of clusters to check. Default value of this argument is 10 or maximum number of individuals.

base_val Indicates whether include a value at zero time point as an additional clustering variable. Default is FALSE and the standard number (7) of clustering parameters is used.

Value

Determine the optimal number of clusters, returns graphical output (red dot in plot indicates the recommended number of clusters according to that index) and table with indices.

Source


Examples

set.seed(123)
data <- GeneratePanel(n = 100, Param = ParamLinear, NbVisit = 10)
OptiNum(data = data, formula = Y ~ Time, group = "ID")
PanelPlot

Plot Micro-Panel (longitudinal) Data

Description

This function plots micro-panel (longitudinal) data from stored data.frame or randomly generated panel data from GeneratePanel function.

Usage

PanelPlot(
  data,
  formula = Y ~ Time,
  group = "ID",
  colour = NA,
  mean_traj_all = FALSE,
  mean_traj_group = FALSE,
  show_legend = TRUE,
  title = NULL,
  x_title = NULL,
  y_title = NULL
)

Arguments

data A data frame containing the variables named in formula and group arguments.

formula A two-sided formula object, with a numeric, clustering variable (Y) on the left of a ~ separator and the time (numeric) variable on the right. Time is measured from the start of the follow-up period (baseline).

group A grouping factor variable (vector), i.e. single identifier for each (trajectory).

colour Character, which is a variable’s name in data. The trajectories are distinguished by colour according to this variable.

mean_traj_all Logical scalar. It indicates whether to show mean overall trajectory. Default is FALSE.

mean_traj_group Logical scalar. It indicates whether to show mean trajectory by group. Default is FALSE.

show_legend Logical scalar. It indicates whether to show cluster legend. Default is TRUE.

title String. Is an optional title for a plot. Otherwise no title will used.

x_title String. Is an optional title for x axis. Otherwise variable name after ~ in formula will used.

y_title String. Is an optional title for y axis. Otherwise variable name before ~ in formula will used.
**Value**

Returns plot using package ggplot2.

**Examples**

```r
set.seed(123)
dataMale <- GeneratePanel(n = 50, Param = ParamLinear, NbVisit = 10)
dataMale$Gender <- "M"
dataFemale <- GeneratePanel(n = 50, Param = ParamLinear, NbVisit = 10)
dataFemale$ID <- dataFemale$ID + 50
dataFemale$Gender <- "F"
data <- rbind(dataMale, dataFemale)
PanelPlot(data = data, formula = Y ~ Time, group = "ID", colour = "Gender")
PanelPlot(data = data, formula = Y ~ Time, group = "ID", colour = "Gender", mean_traj_all = TRUE)
PanelPlot(data = data, formula = Y ~ Time, group = "ID", colour = "Gender", mean_traj_group = TRUE)
```

---

**ParamCubic**

**Parameters of cubic model**

**Description**

Default parameters to generate micro-panel (longitudinal) data with quadratic trend. The parameters may differ per each cluster. The parameters of each cluster are in rows. Number of rows denotes the number of clusters. Fixed effects are taken from Allen et al. (2005), and the source for random effects is Uher et al. (2017).

**Usage**

ParamCubic

**Format**

Its advised to keep parameters in `data.frame`. The Parameters structure is as follows:

- **b0** fixed parameter of intercept
- **b1** fixed parameter of slope
- **b2** fixed parameter of defining the quadraticity
- **b3** fixed parameter of defining the cubicity
- **varU0** variance of random factor U0 given to fixed parameter b0
- **varU1** variance of random factor U1 given to fixed parameter b1
- **corr** correlation between random factors U0 and U1
- **varE** the variability of the residuals
**ParamExpon**

### Source


---

### Description

Default parameters to generate micro-panel (longitudinal) data with exponential trend. The parameters may differ per each cluster. The parameters of each cluster are in rows. Number of rows denotes the number of clusters. Fixed effects are taken from Jones et al. (2013).

### Usage

ParamExpon

### Format

It is advised to keep parameters in data.frame. The Parameters structure is as follows:

- **b0** fixed parameter of intercept
- **b1** fixed parameter of slope
- **b2** fixed parameter of defining the decay
- **varU0** variance of random factor U0 given to fixed parameter b0
- **varU1** variance of random factor U1 given to fixed parameter b1
- **corr** correlation between random factors U0 and U1
- **varE** the variability of the residuals

### Source

**ParamLinear**  
*Parameters of linear model*

**Description**  
Default parameters to generate micro-panel (longitudinal) data with linear trend. The parameters may differ per each cluster. The parameters of each cluster are in rows. Number of rows denotes the number of clusters. Fixed and random effects are taken from Uher et al. (2017).

**Usage**

ParamLinear

**Format**

It is advised to keep parameters in **data.frame**. The Parameters structure is as follows:

- **b0**  fixed parameter of intercept
- **b1**  fixed parameter of slope
- **varU0** variance of random factor U0 given to fixed parameter b0
- **varU1** variance of random factor U1 given to fixed parameter b1
- **corr** correlation between random factors U0 and U1
- **varE** the variability of the residuals

**Source**


---

**ParamQuadrat**  
*Parameters of quadratic model*

**Description**  
Parameters to generate panel data with quadratic trend. The parameters may differ per each cluster. The parameters of each cluster are in rows. Number of rows denotes the number of clusters. Fixed effects are taken from Allen et al. (2005), and the source for random effects is Uher et al. (2017).

**Usage**

ParamQuadrat
**Format**

It is advised to keep parameters in `data.frame`. The Parameters structure is as follows:

- **b0** fixed parameter of intercept
- **b1** fixed parameter of slope
- **b2** fixed parameter of defining the quadraticity
- **varU0** variance of random factor U0 given to fixed parameter b0
- **varU1** variance of random factor U1 given to fixed parameter b1
- **corr** correlation between random factors U0 and U1
- **varE** the variability of the residuals

**Source**


Index

* CLUMP
  PanelPlot, 9
* CluMP
  CluMP, 2
  CluMP_profiles, 3
  CluMP_view, 4
  GeneratePanel, 5
  OptiNum, 7
* datasets
  ParamCubic, 10
  ParamExpon, 11
  ParamLinear, 12
  ParamQuadrat, 12

CluMP, 2, 3, 4
CluMP_profiles, 3
CluMP_view, 4

data.frame, 6, 9–13

formula, 2, 8, 9

GeneratePanel, 5, 9

hclust, 2

list, 2, 3

OptiNum, 2, 7

PanelPlot, 9
ParamCubic, 6, 10
ParamExpon, 6, 11
ParamLinear, 6, 12
ParamQuadrat, 6, 12