Package 'PAMhm'

September 6, 2021

Type Package

Title Generate Heatmaps Based on Partitioning Around Medoids (PAM)

Version 0.1.2

Date 2021-08-24

Author Vidal Fey [aut, cre],
Henri Sara [aut]

Maintainer Vidal Fey <vidal.fey@gmail.com>

Description Data are partitioned (clustered) into k clusters ``around medoids'', which is a more robust version of K-means implemented in the function pam() in the 'cluster' package. The PAM algorithm is described in Kaufman and Rousseeuw (1990) <doi:10.1002/9780470316801>. Please refer to the pam() function documentation for more references. Clustered data is plotted as a split heatmap allowing visualisation of representative ``group-clusters'' (medoids) in the data as separated fractions of the graph while those ``sub-clusters'' are visualised as a traditional heatmap based on hierarchical clustering.

Depends heatmapFlex, cluster, grDevices, graphics, stats

Imports RColorBrewer, R.utils, readxl, readmoRe, utils, plyr, robustHD

Suggests rmarkdown, knitr

License GPL-3

Encoding UTF-8

RoxygenNote 7.1.1

VignetteBuilder knitr

NeedsCompilation no

Repository CRAN

Date/Publication 2021-09-06 07:50:02 UTC

R topics documented:

PAM.hm *Main function to produce a heatmap using PAM clustering.*

Description

This is the main wrapper function to be called by end users. It accepts a numeric matrix (or an object that can be coerced to a numeric matrix) or a number of data file formats and produces one or more PDFs with the plots.

Usage

```
PAM.hm(
  x,
  project.folder = ".",
  nsheets = 1,
  dec = ".",
  header = TRUE,
  symbolcol = 1,
  sample.names = NULL,
  cluster.number = 4,
  trim = NULL,
  winsorize.mat = TRUE,
  cols = "BlueWhiteRed",
  dendrograms = "Both",
  autoadj = TRUE,
  pdf.height = 10,
  pdf.width = 10,
  labelheight = 0.25,
  labelwidth = 0.2,
  r.cex = 0.5,
  c.cex = 1,
  medianCenter = NULL,
  log = FALSE,
  do.log = FALSE,
  log.base = 2,
  metric = "manhattan",
  na.strings = "NA",
  makeFolder = TRUE,
  do.pdf = FALSE,
  do.png = FALSE,
  save.objects = FALSE
)
```

Arguments

x	(character, data.frame, numeric). The name(s) of the input files(s) (character vector) or a data object such as a data.frame or numeric matrix. See 'Details'.

`project.folder` (`character`). Name of the root folder inside which the results will be created if any files are to be saved. See 'Details'.

`nsheets` (`integer`). Number of sheets to be read if file is of type ".xls" or ".xlsx". All sheets starting from 1 up to the given number in the respective data file will be read. If more than one file is read this must be be an integer vector with the numbers of sheets in exactly the same order as the files.

`dec` (`character`). The decimal separator for numbers.

`header` (`logical`). Does the input file have a header row?

`symbolcol` (`character`). The name of the column with identifiers used as labels.

`sample.names` (`character`). A vector of names used for plot titles and output files.

`cluster.number` (`character` or `integer`). A vector of numbers used for PAM clustering (corresponds to argument `k` in `pam`). If a character vector, this is broken down to a numeric vector accepting comma-separated strings in the form of, e.g, "4" and "2-5". The clustering algorithm then iterates through all given numbers. See 'Details'.

`trim` (`numeric`). Value to "cut off" data distribution. Values at both ends of the distribution, larger or smaller, respectively, will be made equal to `+/-trim`, i.e., data will be symmetrical around 0. `NULL` means no trimming which is the default. If `trim` is `-1` (or any negative value) and `winsorize.mat` is `TRUE` the matrix will be *winsorized* and then the smaller of the two largest absolute values at both ends of the distribution rounded to three digits will be used. If `winsorize.mat` is `FALSE` the largest possible absolute integer, i.e., the smaller of the to extreme integers is used. Trimming is disabled for only positive or only negative values.

`winsorize.mat` (`logical`). Should the matrix be *winsorized* (cleaned of outliers) before plotting? Defaults to `TRUE`. See 'Details'.

`cols` (`character`). Name of the colour palette.

`dendrograms` (`character`). Which dendrograms are to be plotted? One of "Vertical", "Horizontal", "None" or "Both". Defaults to "Both".

`autoadj` (`logical`). Should label sizes and pdf dimensions be adjusted automatically? See 'Details'.

`pdf.height` (`numeric`). Heigth of the PDF device.

`pdf.width` (`numeric`). Width of the PDF device.

`labelheight` (`numeric` or `lcm(numeric)`). Relative or absolute height (using `lcm`, see `layout`) of the labels.

`labelwidth` (`numeric` or `lcm(numeric)`). Relative or absolute width (using `lcm`, see `layout`) of the labels.

`r.cex` (`numeric`). Font size for row labels.

`c.cex` (`numeric`). Font size for column labels.

`medianCenter` (`character`). If not `NULL`, how should data be median-centered? One of "grand", "row" or "column". Defaults to `NULL`, no median-centering.

`log` (`logical`). Is the data on log-scale. (The log-base is given in argument `log.base`).

`do.log` (`logical`). Should data be log-transformed? (The log-base is given in argument `log.base`).

log.base	(numeric). The log-base used for log and do.log.
metric	(character). The metric metric to be used for calculating dissimilarities between observations. The currently available options are "euclidean" and "manhattan". Euclidean distances are root sum-of-squares of differences, and manhattan distances are the sum of absolute differences. Defaults to "manhattan".
na.strings	(character). Character vector of strings to interpret as missing values when reading data files with read.table or readxlread_excel. By default, readxl treats blank cells as missing data.
makeFolder	(logical). Should the results folder be created?
do.pdf	(logical). Should images be saved to PDFs?
do.png	(logical). Should images be saved to PNGs?
save.objects	(logical). Should R objects be save to disk?

Details

Argument x can be a data.frame or numeric matrix to be used directly for plotting the heatmap. If it is a data.frame argument symbolcol sets the respective columns for symbols to be used as labels and the column where the numeric data starts.

Matrices will be coerced to data frames.

The read function accepts txt, tsv, csv and xls files.

If PDF, PNG or R object files are to be saved, i.e., if the corresponding arguments are TRUE, a results folder will be created using time and date to create a unique name. The folder will be created in the directory set by argument project.folder. The reasoning behind that behaviour is that during development the heatmap was used as data analysis tool testing various cluster.number values with numerous files and comparing the results.

The cluster.number argument defines the numbers of clusters when doing PAM. After processing it is passed one-by-one to argument k in pam. The numbers can be defined in the form c("2","4-7","9"), for example, depending on the experimental setup. An integer vector is coerced to character.

If autoadj is TRUE character expansion (cex) for rows annd columns, pdf width and height and label width and height are adjusted automatically based on the dimensions of the data matrix and length (number of characters) of the labels.

The default behavior regarding outliers is to *winsorize* the matrix before plotting, i.e., shrink outliers to the unscattered part of the data by replacing extreme values at both ends of the distribution with less extreme values. This is done for the same reason as trimming but the data will not be symmetrical around 0.

Value

A list: Invisibly returns the results object from the PAM clustering.

References

Kaufman, L., & Rousseeuw, P. J. (Eds.). (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley & Sons, Inc. doi: 10.1002/9780470316801

See Also

read.delim

read_excel

pam

Examples

```
# Generate a random 10x10 matrix and plot it using default values
set.seed(1234)                                  # for reproducibility
mat <- matrix(rnorm(120), nrow = 20)            # standard normal
PAM.hm(mat, cluster.number = 3)

## Plot with more than one cluster number
PAM.hm(mat, cluster.number = 2:4)               # integer vector
PAM.hm(mat, cluster.number = c("2", "4-5"))     # character vector

# Using the 'trim' argument
## Introduce outlier to the matrix and plot w/o trimming or winsorization
mat[1] <- mat[1] * 10
PAM.hm(mat, cluster.number = 3, trim = NULL, winsorize = FALSE)

## calculate a trim value by getting the largest possible absolute integer and
## plot again
tr <- min(abs(ceiling(c(min(mat, na.rm = TRUE), max(mat, na.rm = TRUE)))),
    na.rm = TRUE)
PAM.hm(mat, cluster.number = 3, trim = tr, winsorize = FALSE)
## Note that the outlier is still visible but since it is less extreme
## it does not distort the colour scheme.

# An example reading data from an Excel file
# The function readxl::read_excel is used internally to read Excel files.
# The example uses their example data.
readxl_datasets <- readxl::readxl_example("datasets.xlsx")
PAM.hm(readxl_datasets, cluster.number = 4, symbolcol = 5)
```

PAMhm *Generate Heatmaps Based on Partioning Around Medoids (PAM)*

Description

Data are partitioned (clustered) into k clusters "around medoids", which is a more robust version of K-means implemented in the function pam() in the 'cluster' package. The PAM algorithm is described in Kaufman and Rousseeuw (1990) <doi:10.1002/9780470316801>. Please refer to the pam() function documentation for more references. Clustered data is plotted as a split heatmap allowing visualisation of representative "group-clusters" (medoids) in the data as separated fractions of the graph while those "sub-clusters" are visualised as a traditional heatmap based on hierarchical clustering.

Details

Package:	PAMhm
Type:	Package
Initial version:	0.1-0
Created:	2011-01-07
License:	GPL-3
LazyLoad:	yes

Author(s)

Vidal Fey <vidal.fey@gmail.com>, Henri Sara <henri.sara@gmail.com> Maintainer: Vidal Fey <vidal.fey@gmail.com>

Index