Package 'cellWise'

December 7, 2016

Type Package

Version 1.0.0

Date 2016-12-06

Title Analyzing Data with Cellwise Outliers

Depends R (>= 3.2.0)

Suggests knitr, robustHD, robustbase, rrcov, MASS

Imports reshape2, scales, ggplot2, matrixStats, gridExtra

Description
Tools for detecting cellwise outliers and robust methods to analyze data which may contain them.

License GPL (>= 2)

LazyLoad yes

Author Jakob Raymaekers [aut, cre],
Peter Rousseeuw [aut],
Wannes Van den Bossche [aut]

Maintainer Jakob Raymaekers <jakob.raymaekers@kuleuven.be>

VignetteBuilder knitr

NeedsCompilation no

Repository CRAN

Date/Publication 2016-12-07 02:29:23

R topics documented:

cellMap *Draw a cellmap*

Description

This function draws a cellmap, possibly of a subset of rows and columns of the data, and possibly combining cells into blocks. A cellmap shows which cells are missing and which one are outlying, marking them in red for unusually large cell values and in blue for unusually low cell values. When cells are combined into blocks, the final color is the average of the colors in the individual cells.

Usage

```
cellMap(D, R, indcells, indrows, showVals = FALSE, xlabels = "",
        ylabels = "", mTitle = "",xtitle = "", ytitle = "",
        xshowindex = NULL, yshowindex = NULL, xblocksize = 1,
        yblocksize = 1, autolabel = TRUE, anglex = 90, sizexy = 1.1,
        hjustXlabels = 1,hjustYlabels = 1)
```

Arguments

D	The data matrix (required input argument).
R	Matrix of cell residuals (required input argument).
indcells	Indices of outlying cells (required input argument).
indrows	Indices of outlying rows (required input argument).
showVals	Whether to show the entries of D in the cellmap. Defaults to FALSE.
xlabels	Labels for the x-axis.
ylabels	Labels for the y-axis.
mTitle	Main title of the cellMap.
xtitle	Title for the x-axis.
ytitle	Title for the y-axis.
xshowindex	Indices of the cells that will be shown, in the x direction.
yshowindex	Indices of the cells that will be shown, in the y direction.
xblocksize	Size of combination blocks in the x direction. Defaults to 1.
yblocksize	Size of combination blocks in the y direction. Defaults to 1.
autolabel	Automatically combines labels of cells in blocks. If FALSE, you must provide the final xlabels and/or ylabels. Defaults to TRUE.
anglex	Angle of the labels on the x-axis. Defaults to 90.
sizexy	Size of title for x-axis and y-axis. Defaults to 1.1.
hjustXlabels	Adjust x-labels: 0=left, 0.5=centered, 1=right. Defaults to 1.
hjustYlabels	Adjust y-labels: 0=left, 0.5=centered, 1=right. Defaults to 1.

Author(s)

Rousseeuw P.J., Van den Bossche W.

References

Rousseeuw P.J., Van den Bossche W. (2016). Detecting deviating data cells. *arXiv:1601.07251*

See Also

DetectDeviatingCells

Examples

```
# For examples of the cellmap, we refer to the vignette:
vignette("DDC_examples")
```

---

checkDataSet *Purify the dataset*

Description

This function checks the dataset X, and sets aside certain columns and rows that do not satisfy the conditions. It is used by the DetectDeviatingCells function but can be used by itself, to clean a dataset for a different type of analysis.

Usage

```
checkDataSet(X, fracNA = 0.5, numDiscrete = 3, precScale = 1e-12)
```

Arguments

X	X is the input data, and must be an n by d matrix or data frame.
fracNA	Only retain columns and rows with fewer NAs than this fraction. Defaults to 0.5.
numDiscrete	A column that takes on numDiscrete or fewer values will be considered discrete and not retained in the cleaned data. Defaults to 3.
precScale	Only consider columns whose scale is larger than precScale. Here scale is measured by the median absolute deviation. Defaults to $1e-12$.

Value

A list with components:

- colInAnalysis
  Column indices of the columns used in the analysis.

- `rowInAnalysis`
  Row indices of the rows used in the analysis.
- `namesNotNumeric`
  Names of the variables which are not numeric.
- `namesCaseNumber`
  Names of the cases.
- `namesNAcol`
  Names of the NA columns.
- `namesNArow`
  Names of the NA rows.
- `namesDiscrete`
  Names of the discrete variables.
- `namesZeroScale`
  Names of the variables with zero scale.
- `remX`
  Cleaned data after checkDataSet.

Author(s)

Rousseeuw P.J., Van den Bossche W.

References

Rousseeuw P.J., Van den Bossche W. (2016). Detecting deviating data cells. *arXiv:1601.07251*

See Also

`DetectDeviatingCells`

Examples

```
library(MASS)
set.seed(12345)
n = 100; d = 10
A = matrix(0.9, d, d); diag(A) = 1
xclean = mvrnorm(n, rep(0,d), A)
checkedx = checkDataSet(xclean)

# For more examples, we refer to the vignette:
vignette("DDC_examples")
```

DetectDeviatingCells *Detect Deviating Cells*

Description

This function aims to detect cellwise outliers in the data. These are entries in the data matrix which are substantially higher or lower than what could be expected based on the other cells in its column as well as the other cells in its row, taking the relations between the columns into account. (Note that this function first calls checkDataSet and analyzes the remaining cleaned data.)

Usage

```
DetectDeviatingCells(X, DDCpars = list())
```

Arguments

X	X is the input data, and must be an n by d matrix or a data frame.
DDCpars	A list of available options:

- fracNA
  Only consider columns and rows with fewer NAs (missing values) than this fraction (percentage). Defaults to 0.5.
- numDiscrete
  A column that takes on numDiscrete or fewer values will be considered discrete and not used in the analysis. Defaults to 3.
- precScale
  Only consider columns whose scale is larger than precScale. Here scale is measured by the median absolute deviation. Defaults to $1e-12$.
- tolProb
  Tolerance probability, with default 0.99, which determines the cutoff values for flagging outliers. Used in several the steps of the algorithm.
- corrlim
  When trying to estimate z_{ij} from other variables h, we will only use variables h with $|\rho_{j,h}| \geq corrlim$. Variables j without any correlated variables h satisfying this are considered standalone, and treated on their own. Defaults to 0.5.
- combinRule
  The operation to combine estimates of z_{ij} coming from other variables h: can be wmean, wmedian, mean, median. Defaults to wmean.
- includeSelf
  Whether or not the combination rule will include the variable j itself. Defaults to TRUE.
- rowdetect
  Whether the rule for flagging rows is to be applied. Defaults to TRUE.
- returnBigXimp
  If TRUE, the imputed data matrix Ximp in the output will include the rows and columns that were not part of the analysis (and can still contain NAs). Defaults to FALSE.

Value

A list with components:

- `colInAnalysis`
  The column indices of the columns used in the analysis.
- `rowInAnalysis`
  The row indices of the rows used in the analysis.
- `namesNotNumeric`
  The names of the variables which are not numeric.
- `namesCaseNumber`
  The names of the cases which are used in the analysis.
- `namesNAcol`
  Names of the NA columns.
- `namesNArow`
  Names of the NA rows.
- `namesDiscrete`
  Names of the discrete variables.
- `namesZeroScale`
  Names of the variables with zero scale.
- `remX`
  Cleaned data after `checkDataSet`.
- `Z`
  Standardized `X`.
- `k`
  Number of columns used in estimation.
- `ngbrs`
  Indicates highest correlated columns.
- `robcors`
  Robust correlations.
- `robslopes`
  Robust slopes.
- `Xest`
  Estimated `X`.
- `stdResid`
  Standardized residuals of orginal `X` minus the estimated `Xest`.
- `indcells`
  Indices of the cells which were flagged in the analysis.
- `Ti`
  Testvalues (outlyingness) of each row.
- `indrows`
  Indices of the rows which were flagged in the analysis.

- `indall`
  Indices of both the cells and rows which were flagged in the analysis.
- `Ximp`
  Imputed `X`.

Author(s)

Rousseeuw P.J., Van den Bossche W.

References

Rousseeuw P.J., Van den Bossche W. (2016). Detecting deviating data cells. *arXiv:1601.07251*

See Also

`checkDataSet`,`cellMap`

Examples

```
library(MASS)
set.seed(12345)
n = 100; d = 10
A = matrix(0.9, d, d); diag(A) = 1
xclean = mvrnorm(n, rep(0,d), A)
DDCxclean = DetectDeviatingCells(xclean)

# For more examples, we refer to the vignette:
vignette("DDC_examples")
```

---

glass *The glass dataset*

Description

A dataset containing spectra with $d = 750$ wavelengths collected on $n = 180$ archeological glass samples.

Usage

```
data("glass")
```

Format

A data frame with 180 observations of 750 wavelengths.

Source

Lemberge, P., De Raedt, I., Janssens, K.H., Wei, F., and Van Espen, P.J. (2000). Quantitative Z-analysis of 16th-17th century archaeological glass vessels using PLS regression of EPXMA and μ-XRF data. *Journal of Chemometrics*, **14**, 751–763.

Examples

```
data(glass)
```

`mortality` *The mortality dataset*

Description

This dataset contains the mortality by age for males in France, from 1816 to 2010 as obtained from the Human Mortality Database.

Usage

```
data("mortality")
```

Format

A data frame with 198 calendar years (rows) and 91 age brackets (columns).

Source

Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org (data downloaded in November 2015).

References

Hyndman, R.J., and Shang, H.L. (2010), Rainbow plots, bagplots, and boxplots for functional data, *Journal of Computational and Graphical Statistics*, **19**, 29–45.

Examples

```
data(mortality)
```

`philips` *The philips dataset*

Description

A dataset containing measurements of $d = 9$ characteristics of $n = 677$ diaphragm parts, used in the production of TV sets.

Usage

```
data("philips")
```

Format

A matrix with 677 rows and 9 columns.

Source

The data were provided in 1997 by Gertjan Otten and permission to analyze them was given by Herman Veraa and Frans Van Dommelen at Philips Mecoma in The Netherlands.

References

Rousseeuw, P.J., and Van Driessen, K. (1999). A fast algorithm for the Minimum Covariance Determinant estimator. *Technometrics*, **41**, 212–223.

Examples

```
data(philips)
```

Index