Package 'depthTools'

October 16, 2020

Version 0.5

Date 2020-10-15

Title Depth Tools Package

Author Sara Lopez-Pintado <sl2929@columbia.edu> and Aurora Torrente <etorrent@est-econ.uc3m.es>.

Maintainer Aurora Torrente <etorrent@est-econ.uc3m.es>

Depends R (>= 2.8.0)

Description Implementation of different statistical tools for the description and analysis of gene expression data based on the concept of data depth, namely, the scale curves for visualizing the dispersion of one or various groups of samples (e.g. types of tumors), a rank test to decide whether two groups of samples come from a single distribution and two methods of supervised classification techniques, the DS and TAD methods. All these techniques are based on the Modified Band Depth, which is a recent notion of depth with a low computational cost, what renders it very appropriate for high dimensional data such as gene expression data.

License GPL (>= 2)

NeedsCompilation no

Repository CRAN

Date/Publication 2020-10-16 06:20:13 UTC

R topics documented:

`centralPlot` *Plot of the central curves*

Description

`centralPlot` plots distinctly the [np] most central observations, where [np] is the largest integer smaller than np, and the remaining most external ones, according to the modified band depth.

Usage

```
centralPlot(x, p=0.5,col.c='red',col.e='slategray',lty=c(1,3),gradient=FALSE,
            gradient.ramp=NULL,main=NULL,cex=1,...)
```

Arguments

`x`	a data matrix containing the observations (samples) by rows and the variables (genes) by columns.
`p`	proportion of most central samples to be displayed.
`col.c`	the color for the central samples, either as a character string or as a number. Ignored if gradient is TRUE.
`col.e`	the color for the external samples.
`lty`	a vector of two components with the line type of the central and external curves.
`gradient`	a logical value. If TRUE then the most central curves are plotted with colors according to the gradient.ramp parameter.
`gradient.ramp`	an optional vector of two components containing the first and last colors of the palette used to color the most central curves.
`main`	a character string for the plot title.
`cex`	the magnification to be used for the legend.
`...`	further graphical parameters to be passed to 'plot'.

Details

The `centralPlot` allows to visualise the most central curves within the dataset.

Author(s)

Sara Lopez-Pintado <sl2929@columbia.edu> and Aurora Torrente <etorrent@est-econ.uc3m.es>

References

Lopez-Pintado, S. *et al.* (2010). Robust depth-based tools for the analysis of gene expression data. *Biostatistics*, 11 (2), 254-264.

Examples

```
## simulated data
set.seed(0)
x <- matrix(rnorm(100),10,10)
centralPlot(x,p=0.2)

## real data
data(prostate)
prost.x<-prostate[,1:100]
prost.y<-prostate[,101]
centralPlot(prost.x[prost.y==0,], p=0.5)  ## 50 % most central normal samples
centralPlot(prost.x[prost.y==1,], p=0.5, gradient=TRUE, main='Tumor samples')
                                          ## 50 % most central tumoral samples
```

---

classDS *Distance to the Trimmed Mean Classification Method*

Description

Implementation of the classification technique based on assigning each observation to the group that minimizes the distance of the observation to the trimmed mean of the group.

Usage

```
classDS(xl,yl,xt,alpha=0.2)
```

Arguments

`xl`	an `nxp` data matrix containing the observations (samples) from the learning set by rows and the variables (genes) by columns
`yl`	a vector of length `n` containing the class each observations in `xl` belongs to
`xt`	an `mxp` data matrix containing the observations (samples) from the test set by rows and the variables (genes) by columns
`alpha`	the proportion of observations that are trimmed out when computing the mean. 0.2 by default.

Details

This classification method proceeds by first computing the alpha trimmed mean corresponding to each group from the learning set, then computing the distance from a new observation to each trimmed mean. The new sample will then be assigned to the group that minimizes such distance. At the moment, only the Euclidean distance is implemented.

Value

`pred`	the vector of length `m` containing the predicted class of observations in matrix `xt`

Author(s)

Sara Lopez-Pintado <sl2929@columbia.edu> and

Aurora Torrente <etorrent@est-econ.uc3m.es>

References

Lopez-Pintado, S. *et al.* (2010). Robust depth-based tools for the analysis of gene expression data. *Biostatistics*, 11 (2), 254-264.

See Also

classTAD

Examples

```
## simulated data
set.seed(10)
xl <- matrix(rnorm(100),10,10); xl[1:5,]<-xl[1:5,]+1
yl <- c(rep(0,5),rep(1,5))
xt <- matrix(rnorm(100),10,10)
classDS(xl,yl,xt)

## real data
data(prostate)
prost.x<-prostate[,1:100]
prost.y<-prostate[,101]
set.seed(1)
learning <- sample(50,40,replace=FALSE)
yl <- prost.y[learning]
xl <- prost.x[learning,]
training <- c(1:nrow(prost.x))[-learning]
yt.real <- prost.y[training]
xt <- prost.x[training,]
yt.estimated <- classDS(xl,yl,xt)
yt.real==yt.estimated
```

---

classTAD *Weighted Trimmed Mean Distance Classification Method*

Description

Implementation of the classification technique based on assigning each observation to the group that minimizes the trimmed average distance of the given observation to the deepest points of each group in the learning set, weighted by the depth of these points in their own group.

Usage

```
classTAD(xl,yl,xt,alpha=0)
```

Arguments

xl	an nxp data matrix containing the observations (samples) from the learning set by rows and the variables (genes) by columns
yl	a vector of length n containing the class each observations in xl belongs to
xt	an mxp data matrix containing the observations (samples) from the test set by rows and the variables (genes) by columns
alpha	an optional value for the proportion of observations that are trimmed out when computing the mean. 0 by default.

Details

This method classifies a given observation x into one of g groups, of sizes n1,...,ng, but taking into account only the m=min{n1,...,ng} deepest elements of each group in the learning set. Additionally, this number can be reduced in a proportion alpha. The distance of x to these m elements is averaged and weighted with the depth of each element with respect to its own group.

Value

pred	the vector of length m containing the predicted class of observations in matrix xt

Author(s)

Sara Lopez-Pintado <sl2929@columbia.edu> and

Aurora Torrente <etorrent@est-econ.uc3m.es>

References

Lopez-Pintado, S. *et al.* (2010). Robust depth-based tools for the analysis of gene expression data. *Biostatistics*, 11 (2), 254-264.

See Also

classDS

Examples

```
## simulated data
set.seed(0)
xl <- matrix(rnorm(100),10,10); xl[1:5,]<-xl[1:5,]+1
yl <- c(rep(0,5),rep(1,5))
xt <- matrix(rnorm(100),10,10)
classTAD(xl,yl,xt)

## real data
data(prostate)
prost.x<-prostate[,1:100]
prost.y<-prostate[,101]
set.seed(0)
learning <- sample(50,40,replace=FALSE)
```

```
yl <- prost.y[learning]
xl <- prost.x[learning,]
training <- c(1:nrow(prost.x))[-learning]
yt.real <- prost.y[training]
xt <- prost.x[training,]
yt.estimated <- classTAD(xl,yl,xt)
yt.real==yt.estimated
```

MBD *Computation of the Modified Band Depth*

Description

MBD computes the modified band depth of each observation within a sample which either includes or not the given observation.

Usage

```
MBD(x, xRef=NULL, plotting=TRUE, grayscale=FALSE, band=FALSE, band.limits=NULL,
     lty=1, lwd=2, col=NULL, cold=NULL, colRef=NULL, ylim=NULL, cex=1,...)
```

Arguments

x	a data matrix containing the observations (samples) by rows and the variables (genes) by columns.
xRef	an optional data matrix containing the sample of observations with respect to the modified band depth is computed. If unprovided, then all elements in matrix x are taken into account to compute the depth.
plotting	logical value. If TRUE then the observations in the data matrix x are plotted, in parallel coordinates.
grayscale	logical value. If TRUE then a different color from a given color palette is assigned to each sample, according to its depth.
band	logical value. If TRUE then the convex hull (a polygon) of the bands formed by the percentage p of most internal samples are represented. Different values of p can be set with the argument band.limits.
band.limits	a vector of values in the range 0-1 giving the proportion p of most central curves to be considered to form a band.
lty	the line type for drawing both the data set and the reference set.
lwd	the line width for both the data set and the reference set. The thickness of the deepest point is increased by 0.5 with respect to the thickest line drawn.
col	the color specification for the data set, except for the deepest point. If grayscale is true and no color is specified, then the depth of each point is represented in grayscale colors, with higher intensities corresponding to smaller depths.
cold	the color used to plot the deepest point.
colRef	the color specification for the reference data set.

ylim	numeric vector giving the y coordinates range.
cex	the magnification used for the legend.
...	graphical parameters (see 'par') and any further arguments of 'plot'.

Details

The modified band depth is the average proportion of components of the considered observation that are between the corresponding components of all possible pairs of elements in the sample with respect to the depth is computed. The depth is efficiently obtained using the multiplicity of each value in the data matrix ordered by columns rather than exhaustively searching for all pairs of samples.

Value

a list containing:

ordering	vector giving the ordering of the samples according to their corresponding depths
MBD	vector of the computed depths

Author(s)

Sara Lopez-Pintado <sl2929@columbia.edu> and

Aurora Torrente <etorrent@est-econ.uc3m.es>

References

Lopez-Pintado, S. and Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, 104, 486-503.

Lopez-Pintado, S. *et al.* (2010). Robust depth-based tools for the analysis of gene expression data. *Biostatistics*, 11 (2), 254-264.

See Also

scalecurve, R.test

Examples

```
## MBD of all elements within a sample

  ## simulated data
  set.seed(0)
  x <- matrix(rnorm(1000),100,10)
  x.depths.1<-MBD(x,plotting=TRUE)

  ## real data
  data(prostate)
  prost.x<-prostate[,1:100]
  prost.y<-prostate[,101]
  normal.depths<-MBD(prost.x[prost.y==0,],plotting=TRUE,
```

```
                    main="Normal samples")
  tumor.depths<-MBD(prost.x[prost.y==1,],plotting=TRUE, band=TRUE,
                    band.limits=c(.33,.67,1),grayscale=TRUE)



## MBD of a vector with respect to a set of observations

  ## simulated data
  set.seed(0)
  v <- matrix(c(2,1,0,3,-2,1,2,1,0,-2,rnorm(3)),3,5)
  xR <- matrix(rnorm(100),10,5)
  depths<-MBD(v,xR,plotting=TRUE)

  # MBD of normal prostate samples with respect to tumoral ones
  normal.depths<-MBD(prost.x[prost.y==0,],prost.x[prost.y==1,],
                     plotting=TRUE)
  normal.depths<-MBD(prost.x[prost.y==0,],prost.x[prost.y==1,],plotting=TRUE,
                    band=TRUE,band.limits=c(.33,.67,1),grayscale=TRUE)
```

`prostate` *Gene Expression Data from Tumoral and Normal Prostate Samples and Labels*

Description

Normalized subset from Singh et al. (2002) data included in the `prostate` dataset. The raw data comprise the expression of 52 tumoral and 50 non-tumoral prostate samples, obtained using the Affymetrix technology. The data were preprocessed by setting thresholds at 10 and 16000 units, excluding genes whose expression varied less than 5-fold relatively or less than 500 units absolutely between the sample, applying a base 10 logarithmic transformation, and standardising each experiment to zero mean and unit variance across the genes. The 100 most variable genes were selected following the B/W criterion (Dudoit et al. (2002)) and a random selection of 25 normal samples and 25 tumour samples was performed.

Usage

```
data(prostate)
```

Format

a 50x101 matrix containing in the first 100 columns the gene expression data of 25 plus 25 randomly selected tumor and normal prostate samples at the 100 most variable genes, selected by the B/W criterion; the last column contains the sample type: 0=normal, 1=tumor.

Source

The data are described in Singh et al. (2002).

References

Singh *et al*. (2002). Gene expression correlates of clinincal prostate cancer behavior, *Cancer cell*, 1 (2), 203-209.

Dudoit *et al*. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association*, 97 (457), 77-87.

`R.test` *Rank Test Based on the Modified Band Depth*

Description

`R.test` performs the rank test based on the modified band depth, to decide whether two samples come from a single parent distribution.

Usage

```
R.test(x,y,n,m,seed=0)
```

Arguments

`x`	a data matrix containing the observations (samples) by rows and the variables (genes) by columns from the first population
`y`	a data matrix containing the observations (samples) by rows and the variables (genes) by columns from the second population
`n`	size of the first sample (less or equal than the number of rows in `x`)
`m`	size of the second sample (less or equal than the number of rows in `y`)
`seed`	seed to inizialize the random number generation. 0 by default

Details

Given a population P from which a sample of `n` vectors is drawn, and another population P' from which a second sample of `m` vectors is obtained, assume there is a third reference sample (from the same population as the largest sample), whose size is also larger than `n` and `m`. `R.test` identifies the largest sample as the one to be split into test and reference samples and verifies if there are enough observations to run the test. Then, the rank test calculates the proportions R and R' of elements from the reference sample whose depths are less or equal than those from the other samples, relative to the reference one, respectively, and order these values from smallest to highest, giving them a rank from 1 to `n+m`. The statistic sum of the ranks of values R' (from the second population) has the distribution of a sum of m elements randomly drawn from 1 to `n+m` without replacement.

Value

a list containing:

`p.value`	the p-value of the rank test
`statistic`	the value of the statistic W of the rank test

Author(s)

Sara Lopez-Pintado <sl2929@columbia.edu> and

Aurora Torrente <etorrent@est-econ.uc3m.es>

References

Lopez-Pintado, S. *et al.* (2010). Robust depth-based tools for the analysis of gene expression data. *Biostatistics*, 11 (2), 254-264.

Examples

```
## Rank test for samples from the same population
  x <- matrix(rnorm(100),10,10)
  R.test(x,x,4,4)$p.value

  ## real data
  data(prostate)
  prost.x<-prostate[,1:100]
  prost.y<-prostate[,101]
  normal<-prost.x[prost.y==0,]
  R.test(normal,normal,10,10)$p.value


##  Rank test for samples from different populations
  x <- matrix(rnorm(100),10,10)
  y <- matrix(rnorm(100,5),10,10)
  R.test(x,y,4,4)$p.value

  ## real data
  tumor<-prost.x[prost.y==1,]
  R.test(normal,tumor,10,10)$p.value
```

---

scalecurve *Computation and Representation of the Scale Curve*

Description

`scalecurve` computes the scale curve of a given group, based on the modified band depth, at a given value p as the area of the band delimited by the [np] most central observations, where [np] is the largest integer smaller than np.

Usage

```
scalecurve(x,y=NULL,xlab="p",ylab="A(p)",main="Scale curve",lwd=2,
           ...)
```

Arguments

x	a data matrix containing the observations (samples) by rows and the variables (genes) by columns
y	an optional vector (numeric or factor) of length equal to the number of rows in x, containing the class of each observation. If unprovided, then all the elements in x are assumed to belong to a single class
xlab	label in the x axis
ylab	label in the y axis
main	plot title
lwd	line widths for the corresponding scale curve(s)
...	graphical parameters to be passed to 'plot'

Details

The scale curve measures the increase in the area of the band determined by the fraction p most central curves, where p moves from 0 to 1, thus providing a measure of the sample dispersion. If the data set is represented in parallel coordinates, then the area is computed using the trapezoid formula.

Value

r	the value of the scale curve at equidistant values of p, determined by the number of observation within each class. If y is not provided, then r is a vector, otherwise is a list with as many components as classes described by y.

Author(s)

Sara Lopez-Pintado <sl2929@columbia.edu> and

Aurora Torrente <etorrent@est-econ.uc3m.es>

References

Lopez-Pintado, S. *et al.* (2010). Robust depth-based tools for the analysis of gene expression data. *Biostatistics*, 11 (2), 254-264.

Examples

```
## scale curve of a single data set
  ## simulated data
  set.seed(0)
  x <- matrix(rnorm(100),10,10)
  scalecurve(x)

  ## real data
  data(prostate)
  prost.x<-prostate[,1:100]
  prost.y<-prostate[,101]
  scalecurve(prost.x[prost.y==0,])  ## scale curve of normal samples
```

```
  scalecurve(prost.x[prost.y==1,])  ## scale curve of tumoral samples

## scalecurve of different groups
  ## simulated data
  x <- matrix(rnorm(100),10,10)
  y <- c(rep("tumoral",5),rep("normal",5))
  scalecurve(x,y)

  ## real data
  labels<-prost.y
  labels[prost.y==0]<-"normal"; labels[prost.y==1]<-"tumoral"
  scalecurve(prost.x,labels)
```

---

tmean *Modified Band Depth-Based Alpha Trimmed Mean*

Description

tmean computes the mean of the deepest observations within the sample, their depths given by the Modified Band Depth, trimming out the proportion alpha of the outest observations.

Usage

```
tmean(x,alpha=0.2,plotting=FALSE,new=TRUE,cols=c(1,4,8),...)
```

Arguments

x	an nxd data matrix containing the observations (samples) by rows and the variables (genes) by columns
alpha	the proportion of observations that are trimmed out when computing the mean. 0.2 by default.
plotting	a logical value. If TRUE then a plot is built. If alpha has length 1, then the trimmed mean, the samples used for its computation and the discarded ones are plotted with different colors, according to the values of cols, below. If alpha has length greater than 1, then a plot with several trimmed means is constructed. The first element in cols is used to determine a color palette, from cols[1] (for the smallest value in alpha) to 'gray' (for the greatest value in alpha).
new	a logical value. If alpha has length 1 or plotting is FALSE, then it is ignored. If TRUE, a new plot is started; otherwise, the new trimmed means are added to the existing plot.
cols	a vector of length 3 containing, in the following ordering, the colors for depicting the trimmed mean, the trimmed collection of samples and the samples which are not taken into account in the computation of the trimmed mean.
...	graphical parameters (see 'par') and any further arguments of 'plot'.

Details

The rows of matrix x, corresponding to genes, are ordered from center outward, that is, starting with the deepest one(s) and ending with the less deep one(s), according to MBD. The alpha-trimmed mean is computed by first removing the proportion alpha of less deep points, and then computing the component-wise average of the remaining observations.

Value

tm	the alpha-trimmed mean vector of length p of matrix x
tm.x	the deespest points of x after removing the proportion alpha of less deep points

Author(s)

Sara Lopez-Pintado <sl2929@columbia.edu> and

Aurora Torrente <etorrent@est-econ.uc3m.es>

Examples

```
set.seed(50)
x <- matrix(rnorm(100),10,10)
m.x<-apply(x,2,mean)
t.x<-tmean(x,plotting=TRUE, lty=1)

t.x.seq <- tmean(x,alpha=c(0,0.25,0.5,0.75),plotting=TRUE, lty=1, cols=2)
```

Index