Package 'PCAtools'

October 17, 2020

Type Package

Title PCAtools: Everything Principal Components Analysis

Version 2.0.0

Description Principal Component Analysis (PCA) is a very powerful technique that has wide applicability in data science, bioinformatics, and further afield. It was initially developed to analyse large volumes of data in order to tease out the differences/relationships between the logical entities being analysed. It extracts the fundamental structure of the data without the need to build any model to represent it. This 'summary' of the data is arrived at through a process of reduction that can transform the large number of variables into a lesser number that are uncorrelated (i.e. the 'principal components'), while at the same time being capable of easy interpretation on the original data. PCAtools provides functions for data exploration via PCA, and allows the user to generate publication-ready figures. PCA is performed via BiocSingular - users can also identify optimal number of principal components via different metrics, such as elbow method and Horn's parallel analysis, which has relevance for data reduction in single-cell RNA-seq (scRNA-seq) and high dimensional mass cytometry data.

License GPL-3

Depends ggplot2, ggrepel, reshape2, lattice, grDevices, cowplot

Imports methods, stats, utils, Matrix, DelayedMatrixStats, DelayedArray, BiocSingular, BiocParallel, Rcpp, dqrng

Suggests testthat, scran, BiocGenerics, knitr, Biobase, GEOquery, biomaRt, ggplotify, beachmat

LinkingTo Rcpp, beachmat, BH, dqrng

URL https://github.com/kevinblighe/PCAtools

biocViews RNASeq, GeneExpression, Transcription

VignetteBuilder knitr

SystemRequirements C++11

RoxygenNote 6.1.1

git_url https://git.bioconductor.org/packages/PCAtools

git_branch RELEASE_3_11

git_last_commit bb334c0

git_last_commit_date 2020-04-27

Date/Publication 2020-10-16

Author Kevin Blighe [aut, cre],
Anna-Leigh Brown [ctb],
Myles Lewis [ctb],
Aaron Lun [aut, ctb]

Maintainer Kevin Blighe <kevin@clinicalbioinformatics.co.uk>

R topics documented:

`PCAtools-package` *PCAtools: everything Principal Component Analysis*

Description

Principal Component Analysis (PCA) is a very powerful technique that has wide applicability in data science, bioinformatics, and further afield. It was initially developed to analyse large volumes of data in order to tease out the differences/relationships between the logical entities being analysed. It extracts the fundamental structure of the data without the need to build any model to represent it. This 'summary' of the data is arrived at through a process of reduction that can transform the large number of variables into a lesser number that are uncorrelated (i.e. the 'principal components'), while at the same time being capable of easy interpretation on the original data [@PCAtools] [@BligheK]. *PCAtools* provides functions for data exploration via PCA, and allows the user to generate publication-ready figures. PCA is performed via *BiocSingular* [@Lun] - users can also identify optimal number of principal components via different metrics, such as elbow method and Horn's parallel analysis [@Horn] [@Buja], which has relevance for data reduction in single-cell RNA-seq (scRNA-seq) and high dimensional mass cytometry data.

biplot *biplot*

Description

Draw a bi-plot, comparing 2 selected principal components / eigenvectors.

Usage

```
biplot(pcaobj,
x = 'PC1',
y = 'PC2',
colby = NULL,
colkey = NULL,
singlecol = NULL,
shape = NULL,
shapekey = NULL,
pointSize = 3.0,
legendPosition = 'none',
legendLabSize = 12,
legendIconSize = 5.0,
xlim = NULL,
ylim = NULL,
lab = rownames(pcaobj$metadata),
labSize = 3.0,
labhjust = 1.5,
labvjust = 0,
selectLab = NULL,
drawConnectors = TRUE,
widthConnectors = 0.5,
colConnectors = 'grey50',
xlab = paste0(x, ', ',
  round(pcaobj$variance[x], digits=2),
  '% variation'),
xlabAngle = 0,
xlabhjust = 0.5,
xlabvjust = 0.5,
ylab = paste0(y, ', ',
  round(pcaobj$variance[y], digits=2),
  '% variation'),
ylabAngle = 0,
ylabhjust = 0.5,
ylabvjust = 0.5,
axisLabSize = 16,
title = '',
subtitle = '',
caption = '',
titleLabSize = 16,
subtitleLabSize = 12,
captionLabSize = 12,
hline = NULL,
```

```
  hlineType = 'longdash',
  hlineCol = 'black',
  hlineWidth = 0.4,
  vline = NULL,
  vlineType = 'longdash',
  vlineCol = 'black',
  vlineWidth = 0.4,
  gridlines.major = TRUE,
  gridlines.minor = TRUE,
  borderWidth = 0.8,
  borderColour = 'black',
  returnPlot = TRUE)
```

Arguments

pcaobj	Object of class 'pca' created by pca(). REQUIRED.
x	A principal component to plot on x-axis. All principal component names are stored in pcaobj$label. DEFAULT = 'PC1'. REQUIRED.
y	A principal component to plot on y-axis. All principal component names are stored in pcaobj$label. DEFAULT = 'PC2'. REQUIRED.
colby	If NULL, all points will be coloured differently. If not NULL, value is assumed to be a column name in pcaobj$metadata relating to some grouping/categorical variable. DEFAULT = NULL. OPTIONAL.
colkey	Vector of name-value pairs relating to value passed to 'col', e.g., c(A='forestgreen', B='gold'). DEFAULT = NULL. OPTIONAL.
singlecol	If specified, all points will be shaded by this colour. Overrides 'col'. DEFAULT = NULL. OPTIONAL.
shape	If NULL, all points will be have the same shape. If not NULL, value is assumed to be a column name in pcaobj$metadata relating to some grouping/categorical variable. DEFAULT = NULL. OPTIONAL.
shapekey	Vector of name-value pairs relating to value passed to 'shape', e.g., c(A=10, B=21). DEFAULT = NULL. OPTIONAL.
pointSize	Size of plotted points. DEFAULT = 3.0. OPTIONAL.
legendPosition	Position of legend ('top', 'bottom', 'left', 'right', 'none'). DEFAULT = 'none'. OPTIONAL.
legendLabSize	Size of plot legend text. DEFAULT = 10. OPTIONAL.
legendIconSize	Size of plot legend icons / symbols. DEFAULT = 3.0. OPTIONAL.
xlim	Limits of the x-axis. DEFAULT = NULL. OPTIONAL.
ylim	Limits of the y-axis. DEFAULT = NULL. OPTIONAL.
lab	A vector containing labels to add to the plot. DEFAULT = rownames(pcaobj$metadata). OPTIONAL.
labSize	Size of labels. DEFAULT = 3.0. OPTIONAL.
labhjust	Horizontal adjustment of label. DEFAULT = 1.5. OPTIONAL.
labvjust	Vertical adjustment of label. DEFAULT = 0. OPTIONAL.
selectLab	A vector containing a subset of lab to plot. DEFAULT = NULL. OPTIONAL.
drawConnectors	Logical, indicating whether or not to connect plot labels to their corresponding points by line connectors. DEFAULT = TRUE. OPTIONAL.

widthConnectors
: Line width of connectors. DEFAULT = 0.5. OPTIONAL.

colConnectors
: Line colour of connectors. DEFAULT = 'grey50'. OPTIONAL.

xlab
: Label for x-axis. DEFAULT = paste0(x, ', ', round(pcaobj$variance[x], digits=2), '% variation'). OPTIONAL.

xlabAngle
: Rotation angle of x-axis labels. DEFAULT = 0. OPTIONAL.

xlabhjust
: Horizontal adjustment of x-axis labels. DEFAULT = 0.5. OPTIONAL.

xlabvjust
: Vertical adjustment of x-axis labels. DEFAULT = 0.5. OPTIONAL.

ylab
: Label for y-axis. DEFAULT = paste0(y, ', ', round(pcaobj$variance[y], digits=2), '% variation'). OPTIONAL.

ylabAngle
: Rotation angle of y-axis labels. DEFAULT = 0. OPTIONAL.

ylabhjust
: Horizontal adjustment of y-axis labels. DEFAULT = 0.5. OPTIONAL.

ylabvjust
: Vertical adjustment of y-axis labels. DEFAULT = 0.5. OPTIONAL.

axisLabSize
: Size of x- and y-axis labels. DEFAULT = 16. OPTIONAL.

title
: Plot title. DEFAULT = ''. OPTIONAL.

subtitle
: Plot subtitle. DEFAULT = ''. OPTIONAL.

caption
: Plot caption. DEFAULT = ''. OPTIONAL.

titleLabSize
: Size of plot title. DEFAULT = 16. OPTIONAL.

subtitleLabSize
: Size of plot subtitle. DEFAULT = 12. OPTIONAL.

captionLabSize
: Size of plot caption. DEFAULT = 12. OPTIONAL.

hline
: Draw one or more horizontal lines passing through this/these values on y-axis. For single values, only a single numerical value is necessary. For multiple lines, pass these as a vector, e.g., c(60,90). DEFAULT = NULL. OPTIONAL.

hlineType
: Line type for hline ('blank', 'solid', 'dashed', 'dotted', 'dotdash', 'longdash', 'twodash'). DEFAULT = 'longdash'. OPTIONAL.

hlineCol
: Colour of hline. DEFAULT = 'black'. OPTIONAL.

hlineWidth
: Width of hline. DEFAULT = 0.4. OPTIONAL.

vline
: Draw one or more vertical lines passing through this/these values on x-axis. For single values, only a single numerical value is necessary. For multiple lines, pass these as a vector, e.g., c(60,90). DEFAULT = NULL. OPTIONAL.

vlineType
: Line type for vline ('blank', 'solid', 'dashed', 'dotted', 'dotdash', 'longdash', 'twodash'). DEFAULT = 'longdash'. OPTIONAL.

vlineCol
: Colour of vline. DEFAULT = 'black'. OPTIONAL.

vlineWidth
: Width of vline. DEFAULT = 0.4. OPTIONAL.

gridlines.major
: Logical, indicating whether or not to draw major gridlines. DEFAULT = TRUE. OPTIONAL.

gridlines.minor
: Logical, indicating whether or not to draw minor gridlines. DEFAULT = TRUE. OPTIONAL.

borderWidth
: Width of the border on the x and y axes. DEFAULT = 0.8. OPTIONAL.

borderColour
: Colour of the border on the x and y axes. DEFAULT = 'black'. OPTIONAL.

returnPlot
: Logical, indicating whether or not to return the plot object. DEFAULT = TRUE. OPTIONAL.

Value

A ggplot2 object.

Author(s)

Kevin Blighe <kevin@clinicalbioinformatics.co.uk>, Aaron Lun

Examples

```
options(scipen=10)
options(digits=6)

col <- 20
row <- 20000
mat1 <- matrix(
  rexp(col*row, rate = 0.1),
  ncol = col)
rownames(mat1) <- paste0('gene', 1:nrow(mat1))
colnames(mat1) <- paste0('sample', 1:ncol(mat1))

mat2 <- matrix(
  rexp(col*row, rate = 0.1),
  ncol = col)
rownames(mat2) <- paste0('gene', 1:nrow(mat2))
colnames(mat2) <- paste0('sample', (ncol(mat1)+1):(ncol(mat1)+ncol(mat2)))

mat <- cbind(mat1, mat2)

metadata <- data.frame(row.names = colnames(mat))
metadata$Group <- rep(NA, ncol(mat))
metadata$Group[seq(1,40,2)] <- 'A'
metadata$Group[seq(2,40,2)] <- 'B'
metadata$CRP <- sample.int(100, size=ncol(mat), replace=TRUE)
metadata$ESR <- sample.int(100, size=ncol(mat), replace=TRUE)

p <- pca(mat, metadata = metadata, removeVar = 0.1)

biplot(p)

biplot(p, colby = 'Group', shape = 'Group')

biplot(p, colby = 'Group', colkey = c(A = 'forestgreen', B = 'gold'),
  legendPosition = 'right')

biplot(p, colby = 'Group', colkey = c(A='forestgreen', B='gold'),
  shape = 'Group', shapekey = c(A=10, B=21), legendPosition = 'bottom')
```

`eigencorplot` *eigencorplot*

Description

Correlate principal components to continuous variable metadata and test significancies of these.

Usage

```
eigencorplot(pcaobj,
  components = getComponents(pcaobj, seq_len(10)),
  metavars,
  titleX = '',
  cexTitleX = 1.0,
  rotTitleX = 0,
  colTitleX = 'black',
  fontTitleX = 2,
  titleY = '',
  cexTitleY = 1.0,
  rotTitleY = 0,
  colTitleY = 'black',
  fontTitleY = 2,
  cexLabX = 1.0,
  rotLabX = 0,
  colLabX = 'black',
  fontLabX = 2,
  cexLabY = 1.0,
  rotLabY = 0,
  colLabY = 'black',
  fontLabY = 2,
  posLab = 'bottomleft',
  col = c('blue4', 'blue3', 'blue2', 'blue1', 'white',
    'red1', 'red2', 'red3', 'red4'),
  posColKey = 'right',
  cexLabColKey = 1.0,
  cexCorval = 1.0,
  colCorval = 'black',
  fontCorval = 1,
  scale = TRUE,
  main = '',
  cexMain = 2,
  rotMain = 0,
  colMain = 'black',
  fontMain = 2,
  corFUN = 'pearson',
  corUSE = 'pairwise.complete.obs',
  corMultipleTestCorrection = 'none',
  signifSymbols = c('***', '**', '*', ''),
  signifCutpoints = c(0, 0.001, 0.01, 0.05, 1),
  colFrame = 'white',
  plotRsquared = FALSE,
  returnPlot = TRUE)
```

Arguments

pcaobj	Object of class 'pca' created by pca(). REQUIRED.
components	The principal components to be included in the plot. DEFAULT = getComponents(pcaobj, seq_len(10)). OPTIONAL.
metavars	A vector of column names in metadata representing continuos variables. REQUIRED.

titleX	X-axis title. DEFAULT = ''. OPTIONAL.
cexTitleX	X-axis title cex. DEFAULT = 1.0. OPTIONAL.
rotTitleX	X-axis title rotation in degrees. DEFAULT = 0. OPTIONAL.
colTitleX	X-axis title colour. DEFAULT = 'black'. OPTIONAL.
fontTitleX	X-axis title font style. 1, plain; 2, bold; 3, italic; 4, bold-italic. DEFAULT = 2. OPTIONAL.
titleY	Y-axis title. DEFAULT = ''. OPTIONAL.
cexTitleY	Y-axis title cex. DEFAULT = 1.0. OPTIONAL.
rotTitleY	Y-axis title rotation in degrees. DEFAULT = 0. OPTIONAL.
colTitleY	Y-axis title colour. DEFAULT = 'black'. OPTIONAL.
fontTitleY	Y-axis title font style. 1, plain; 2, bold; 3, italic; 4, bold-italic. DEFAULT = 2. OPTIONAL.
cexLabX	X-axis labels cex. DEFAULT = 1.0. OPTIONAL.
rotLabX	X-axis labels rotation in degrees. DEFAULT = 0. OPTIONAL.
colLabX	X-axis labels colour. DEFAULT = 'black'. OPTIONAL.
fontLabX	X-axis labels font style. 1, plain; 2, bold; 3, italic; 4, bold-italic. DEFAULT = 2. OPTIONAL.
cexLabY	Y-axis labels cex. DEFAULT = 1.0. OPTIONAL.
rotLabY	Y-axis labels rotation in degrees. DEFAULT = 0. OPTIONAL.
colLabY	Y-axis labels colour. DEFAULT = 'black'. OPTIONAL.
fontLabY	Y-axis labels font style. 1, plain; 2, bold; 3, italic; 4, bold-italic. DEFAULT = 2. OPTIONAL.
posLab	Positioning of the X- and Y-axis labels. 'bottomleft', bottom and left; 'topright', top and right; 'all', bottom / top and left /right; 'none', no labels. DEFAULT = 'bottomleft'. OPTIONAL.
col	Colour shade gradient for RColorBrewer. DEFAULT = c('blue4', 'blue3', 'blue2', 'blue1', 'white', 'red1', 'red2', 'red3', 'red4'). OPTIONAL.
posColKey	Position of colour key. 'bottom', 'left', 'top', 'right'. DEFAULT = 'right'. OPTIONAL.
cexLabColKey	Colour key labels cex. DEFAULT = 1.0. OPTIONAL.
cexCorval	Correlation values cex. DEFAULT = 1.0. OPTIONAL.
colCorval	Correlation values colour. DEFAULT = 'black'. OPTIONAL.
fontCorval	Correlation values font style. 1, plain; 2, bold; 3, italic; 4, bold-italic. DEFAULT = 1. OPTIONAL.
scale	Logical, indicating whether or not to scale the colour range to max and min cor values. DEFAULT = TRUE. OPTIONAL.
main	Plot title. DEFAULT = ''. OPTIONAL.
cexMain	Plot title cex. DEFAULT = 2. OPTIONAL.
rotMain	Plot title rotation in degrees. DEFAULT = 0. OPTIONAL.
colMain	Plot title colour. DEFAULT = 'black'. OPTIONAL.
fontMain	Plot title font style. 1, plain; 2, bold; 3, italic; 4, bold-italic. DEFAULT = 2. OPTIONAL.

corFUN	Correlation method: 'pearson', 'spearman', or 'kendall'. DEFAULT = 'pearson'. OPTIONAL.
corUSE	Method for handling missing values (see documentation for cor function via ?cor). 'everything', 'all.obs', 'complete.obs', 'na.or.complete', or 'pairwise.complete.obs'. DEFAULT = 'pairwise.complete.obs'. OPTIONAL.
corMultipleTestCorrection	Multiple testing p-value adjustment method. Any method from stats::p.adjust() can be used. Activating this function means that signifSymbols and signifCutpoints then relate to adjusted (not nominal) p-values. DEFAULT = 'none'.
signifSymbols	Statistical significance symbols to display beside correlation values. DEFAULT = c('***', '**', '*', ''). OPTIONAL.
signifCutpoints	Cut-points for statistical significance. DEFAULT = c(0, 0.001, 0.01, 0.05, 1). OPTIONAL.
colFrame	Frame colour. DEFAULT = 'white'. OPTIONAL.
plotRsquared	Logical, indicating whether or not to plot R-squared values. DEFAULT = FALSE. OPTIONAL.
returnPlot	Logical, indicating whether or not to return the plot object. DEFAULT = TRUE. OPTIONAL.

Value

A lattice object.

Author(s)

Kevin Blighe <kevin@clinicalbioinformatics.co.uk>

Examples

```
options(scipen=10)
options(digits=6)

col <- 20
row <- 20000
mat1 <- matrix(
  rexp(col*row, rate = 0.1),
  ncol = col)
rownames(mat1) <- paste0('gene', 1:nrow(mat1))
colnames(mat1) <- paste0('sample', 1:ncol(mat1))

mat2 <- matrix(
  rexp(col*row, rate = 0.1),
  ncol = col)
rownames(mat2) <- paste0('gene', 1:nrow(mat2))
colnames(mat2) <- paste0('sample', (ncol(mat1)+1):(ncol(mat1)+ncol(mat2)))

mat <- cbind(mat1, mat2)

metadata <- data.frame(row.names = colnames(mat))
metadata$Group <- rep(NA, ncol(mat))
metadata$Group[seq(1,40,2)] <- 'A'
metadata$Group[seq(2,40,2)] <- 'B'
```

```
metadata$CRP <- sample.int(100, size=ncol(mat), replace=TRUE)
metadata$ESR <- sample.int(100, size=ncol(mat), replace=TRUE)

p <- pca(mat, metadata = metadata, removeVar = 0.1)

eigencorplot(p, components = getComponents(p, 1:10),
  metavars = c('ESR', 'CRP'))
```

`findElbowPoint` *Find the elbow point*

Description

Find the elbow point in the curve of variance explained by each successive PC. This can be used to determine the number of PCs to retain.

Usage

```
findElbowPoint(variance)
```

Arguments

`variance`	Numeric vector containing the variance explained by each PC. Should be monotonic decreasing.

Details

Our assumption is that each of the top PCs capturing real signal should explain much more variance than the remaining PCs. Thus, there should be a sharp drop in the percentage of variance explained when we move past the last "important" PC. This elbow point lies at the base of this drop in the curve, and is identified as the point that maximizes the distance from the diagonal.

Value

An integer scalar specifying the number of PCs at the elbow point.

Author(s)

Aaron Lun

Examples

```
col <- 20
row <- 1000
mat <- matrix(rexp(col*row, rate = 1), ncol = col)

# Adding some structure to make it more interesting.
mat[1:100,1:3] <- mat[1:100,1:3] + 5
mat[1:100+100,3:6] <- mat[1:100+100,3:6] + 5
mat[1:100+200,7:10] <- mat[1:100+200,7:10] + 5
mat[1:100+300,11:15] <- mat[1:100+300,11:15] + 5

p <- pca(mat)
chosen <- findElbowPoint(p$variance)
```

```
plot(p$variance)
abline(v=chosen, col="red")
```

---

getComponents *getComponents*

Description

Return the principal component labels for an object of class 'pca'.

Usage

```
getComponents(
  pcaobj,
  components = NULL)
```

Arguments

pcaobj	Object of class 'pca' created by pca(). REQUIRED.
components	Indices of the principal components whose names will be returned. If NULL, all PC names will be returned. DEFAULT = NULL. OPTIONAL.

Value

A character object.

Author(s)

Kevin Blighe <kevin@clinicalbioinformatics.co.uk>

Examples

```
options(scipen=10)
options(digits=6)

col <- 20
row <- 20000
mat1 <- matrix(
  rexp(col*row, rate = 0.1),
  ncol = col)
rownames(mat1) <- paste0('gene', 1:nrow(mat1))
colnames(mat1) <- paste0('sample', 1:ncol(mat1))

mat2 <- matrix(
  rexp(col*row, rate = 0.1),
  ncol = col)
rownames(mat2) <- paste0('gene', 1:nrow(mat2))
colnames(mat2) <- paste0('sample', (ncol(mat1)+1):(ncol(mat1)+ncol(mat2)))

mat <- cbind(mat1, mat2)

metadata <- data.frame(row.names = colnames(mat))
```

```
metadata$Group <- rep(NA, ncol(mat))
metadata$Group[seq(1,40,2)] <- 'A'
metadata$Group[seq(2,40,2)] <- 'B'
metadata$CRP <- sample.int(100, size=ncol(mat), replace=TRUE)
metadata$ESR <- sample.int(100, size=ncol(mat), replace=TRUE)

p <- pca(mat, metadata = metadata, removeVar = 0.1)

getComponents(p)
```

`getLoadings` *getLoadings*

Description

Return component loadings for principal components from an object of class 'pca'.

Usage

```
getLoadings(
  pcaobj,
  components = NULL)
```

Arguments

`pcaobj`	Object of class 'pca' created by pca(). REQUIRED.
`components`	Indices of the principal components whose component loadings will be returned. If NULL, all PC names will be returned. DEFAULT = NULL. OPTIONAL.

Value

A `data.frame` object.

Author(s)

Kevin Blighe <kevin@clinicalbioinformatics.co.uk>

Examples

```
options(scipen=10)
options(digits=6)

col <- 20
row <- 20000
mat1 <- matrix(
  rexp(col*row, rate = 0.1),
  ncol = col)
rownames(mat1) <- paste0('gene', 1:nrow(mat1))
colnames(mat1) <- paste0('sample', 1:ncol(mat1))

mat2 <- matrix(
  rexp(col*row, rate = 0.1),
  ncol = col)
```

```
rownames(mat2) <- paste0('gene', 1:nrow(mat2))
colnames(mat2) <- paste0('sample', (ncol(mat1)+1):(ncol(mat1)+ncol(mat2)))

mat <- cbind(mat1, mat2)

metadata <- data.frame(row.names = colnames(mat))
metadata$Group <- rep(NA, ncol(mat))
metadata$Group[seq(1,40,2)] <- 'A'
metadata$Group[seq(2,40,2)] <- 'B'
metadata$CRP <- sample.int(100, size=ncol(mat), replace=TRUE)
metadata$ESR <- sample.int(100, size=ncol(mat), replace=TRUE)

p <- pca(mat, metadata = metadata, removeVar = 0.1)

getLoadings(p)
```

---

getVars *getVars*

Description

Return the explained variation for each principal component for an object of class 'pca'.

Usage

```
getVars(
  pcaobj,
  components = NULL)
```

Arguments

pcaobj	Object of class 'pca' created by pca(). REQUIRED.
components	Indices of the principal components whose explained variances will be returned. If NULL, all values will be returned. DEFAULT = NULL. OPTIONAL.

Value

A numeric object.

Author(s)

Kevin Blighe <kevin@clinicalbioinformatics.co.uk>

Examples

```
options(scipen=10)
options(digits=6)

col <- 20
row <- 20000
mat1 <- matrix(
  rexp(col*row, rate = 0.1),
  ncol = col)
```

```
rownames(mat1) <- paste0('gene', 1:nrow(mat1))
colnames(mat1) <- paste0('sample', 1:ncol(mat1))

mat2 <- matrix(
  rexp(col*row, rate = 0.1),
  ncol = col)
rownames(mat2) <- paste0('gene', 1:nrow(mat2))
colnames(mat2) <- paste0('sample', (ncol(mat1)+1):(ncol(mat1)+ncol(mat2)))

mat <- cbind(mat1, mat2)

metadata <- data.frame(row.names = colnames(mat))
metadata$Group <- rep(NA, ncol(mat))
metadata$Group[seq(1,40,2)] <- 'A'
metadata$Group[seq(2,40,2)] <- 'B'
metadata$CRP <- sample.int(100, size=ncol(mat), replace=TRUE)
metadata$ESR <- sample.int(100, size=ncol(mat), replace=TRUE)

p <- pca(mat, metadata = metadata, removeVar = 0.1)

getVars(p)
```

---

pairsplot *pairsplot*

Description

Draw multiple bi-plots.

Usage

```
pairsplot(pcaobj,
components = getComponents(pcaobj, seq_len(5)),
triangle = TRUE,
trianglelabSize = 18,
plotaxes = TRUE,
margingaps = unit(c(0.1, 0.1, 0.1, 0.1), 'cm'),
ncol = NULL,
nrow = NULL,
x = NULL,
y = NULL,
colby = NULL,
colkey = NULL,
singlecol = NULL,
shape = NULL,
shapekey = NULL,
pointSize = 1.0,
legendPosition = 'none',
legendLabSize = 6,
legendIconSize = 1.5,
xlim = NULL,
ylim = NULL,
lab = NULL,
```

```
  labSize = 1.5,
  labhjust = 1.5,
  labvjust = 0,
  selectLab = NULL,
  drawConnectors = FALSE,
  widthConnectors = 0.5,
  colConnectors = 'grey50',
  xlab = NULL,
  xlabAngle = 0,
  xlabhjust = 0.5,
  xlabvjust = 0.5,
  ylab = NULL,
  ylabAngle = 0,
  ylabhjust = 0.5,
  ylabvjust = 0.5,
  axisLabSize = 10,
  title = NULL,
  titleLabSize = 32,
  hline = NULL,
  hlineType = 'longdash',
  hlineCol = 'black',
  hlineWidth = 0.4,
  vline = NULL,
  vlineType = 'longdash',
  vlineCol = 'black',
  vlineWidth = 0.4,
  gridlines.major = TRUE,
  gridlines.minor = TRUE,
  borderWidth = 0.8,
  borderColour = 'black',
  returnPlot = TRUE)
```

Arguments

pcaobj	Object of class 'pca' created by pca(). REQUIRED.
components	The principal components to be included in the plot. These will be compared in a pairwise fashion via multiple calls to biplot(). DEFAULT = getComponents(pcaobj, seq_len(5)). OPTIONAL.
triangle	Logical, indicating whether or not to draw the plots in the upper panel in a triangular arrangement? Principal component names will be labeled along the diagonal. DEFAULT = TRUE. OPTIONAL.
trianglelabSize	Size of p rincipal component label (when triangle = TRUE). DEFAULT = 18. OPTIONAL.
plotaxes	Logical, indicating whether or not to draw the axis tick, labels, and titles. DEFAULT = TRUE. OPTIONAL.
margingaps	The margins between plots in the plot space. Takes the form of a 'unit()' variable. DEFAULT = unit(c(0.1, 0.1, 0.1, 0.1), 'cm'). OPTIONAL.
ncol	If triangle = FALSE, the number of columns in the final merged plot. DEFAULT = NULL. OPTIONAL.

nrow	If triangle = FALSE, the number of rows in the final merged plot. DEFAULT = NULL. OPTIONAL.
x	A principal component to plot on x-axis. All principal component names are stored in pcaobj$label. DEFAULT = NULL. OPTIONAL.
y	A principal component to plot on y-axis. All principal component names are stored in pcaobj$label. DEFAULT = NULL. OPTIONAL.
colby	If NULL, all points will be coloured differently. If not NULL, value is assumed to be a column name in pcaobj$metadata relating to some grouping/categorical variable. DEFAULT = NULL. OPTIONAL.
colkey	Vector of name-value pairs relating to value passed to 'col', e.g., c(A='forestgreen', B='gold'). DEFAULT = NULL. OPTIONAL.
singlecol	If specified, all points will be shaded by this colour. Overrides 'col'. DEFAULT = NULL. OPTIONAL.
shape	If NULL, all points will be have the same shape. If not NULL, value is assumed to be a column name in pcaobj$metadata relating to some grouping/categorical variable. DEFAULT = NULL. OPTIONAL.
shapekey	Vector of name-value pairs relating to value passed to 'shape', e.g., c(A=10, B=21). DEFAULT = NULL. OPTIONAL.
pointSize	Size of plotted points. DEFAULT = 1.0. OPTIONAL.
legendPosition	Position of legend ('top', 'bottom', 'left', 'right', 'none'). DEFAULT = 'none'. OPTIONAL.
legendLabSize	Size of plot legend text. DEFAULT = 6. OPTIONAL.
legendIconSize	Size of plot legend icons / symbols. DEFAULT = 1.5. OPTIONAL.
xlim	Limits of the x-axis. DEFAULT = NULL. OPTIONAL.
ylim	Limits of the y-axis. DEFAULT = NULL. OPTIONAL.
lab	A vector containing labels to add to the plot. DEFAULT = NULL. OPTIONAL.
labSize	Size of labels. DEFAULT = 1.5. OPTIONAL.
labhjust	Horizontal adjustment of label. DEFAULT = 1.5. OPTIONAL.
labvjust	Vertical adjustment of label. DEFAULT = 0. OPTIONAL.
selectLab	A vector containing a subset of lab to plot. DEFAULT = NULL. OPTIONAL.
drawConnectors	Logical, indicating whether or not to connect plot labels to their corresponding points by line connectors. DEFAULT = FALSE. OPTIONAL.
widthConnectors	Line width of connectors. DEFAULT = 0.5. OPTIONAL.
colConnectors	Line colour of connectors. DEFAULT = 'grey50'. OPTIONAL.
xlab	Label for x-axis. DEFAULT = NULL. OPTIONAL.
xlabAngle	Rotation angle of x-axis labels. DEFAULT = 0. OPTIONAL.
xlabhjust	Horizontal adjustment of x-axis labels. DEFAULT = 0.5. OPTIONAL.
xlabvjust	Vertical adjustment of x-axis labels. DEFAULT = 0.5. OPTIONAL.
ylab	Label for y-axis. DEFAULT = NULL. OPTIONAL.
ylabAngle	Rotation angle of y-axis labels. DEFAULT = 0. OPTIONAL.
ylabhjust	Horizontal adjustment of y-axis labels. DEFAULT = 0.5. OPTIONAL.
ylabvjust	Vertical adjustment of y-axis labels. DEFAULT = 0.5. OPTIONAL.

axisLabSize	Size of x- and y-axis labels. DEFAULT = 10. OPTIONAL.
title	Plot title. DEFAULT = NULL. OPTIONAL.
titleLabSize	Size of plot title. DEFAULT = 32. OPTIONAL.
hline	Draw one or more horizontal lines passing through this/these values on y-axis. For single values, only a single numerical value is necessary. For multiple lines, pass these as a vector, e.g., c(60,90). DEFAULT = NULL. OPTIONAL.
hlineType	Line type for hline ('blank', 'solid', 'dashed', 'dotted', 'dotdash', 'longdash', 'twodash'). DEFAULT = 'longdash'. OPTIONAL.
hlineCol	Colour of hline. DEFAULT = 'black'. OPTIONAL.
hlineWidth	Width of hline. DEFAULT = 0.4. OPTIONAL.
vline	Draw one or more vertical lines passing through this/these values on x-axis. For single values, only a single numerical value is necessary. For multiple lines, pass these as a vector, e.g., c(60,90). DEFAULT = NULL. OPTIONAL.
vlineType	Line type for vline ('blank', 'solid', 'dashed', 'dotted', 'dotdash', 'longdash', 'twodash'). DEFAULT = 'longdash'. OPTIONAL.
vlineCol	Colour of vline. DEFAULT = 'black'. OPTIONAL.
vlineWidth	Width of vline. DEFAULT = 0.4. OPTIONAL.
gridlines.major	Logical, indicating whether or not to draw major gridlines. DEFAULT = TRUE. OPTIONAL.
gridlines.minor	Logical, indicating whether or not to draw minor gridlines. DEFAULT = TRUE. OPTIONAL.
borderWidth	Width of the border on the x and y axes. DEFAULT = 0.8. OPTIONAL.
borderColour	Colour of the border on the x and y axes. DEFAULT = 'black'. OPTIONAL.
returnPlot	Logical, indicating whether or not to return the plot object. DEFAULT = TRUE. OPTIONAL.

Value

A cowplot object.

Author(s)

Kevin Blighe <kevin@clinicalbioinformatics.co.uk>

Examples

```
options(scipen=10)
options(digits=6)

col <- 20
row <- 20000
mat1 <- matrix(
  rexp(col*row, rate = 0.1),
  ncol = col)
rownames(mat1) <- paste0('gene', 1:nrow(mat1))
colnames(mat1) <- paste0('sample', 1:ncol(mat1))

mat2 <- matrix(
```

```
  rexp(col*row, rate = 0.1),
  ncol = col)
rownames(mat2) <- paste0('gene', 1:nrow(mat2))
colnames(mat2) <- paste0('sample', (ncol(mat1)+1):(ncol(mat1)+ncol(mat2)))

mat <- cbind(mat1, mat2)

metadata <- data.frame(row.names = colnames(mat))
metadata$Group <- rep(NA, ncol(mat))
metadata$Group[seq(1,40,2)] <- 'A'
metadata$Group[seq(2,40,2)] <- 'B'
metadata$CRP <- sample.int(100, size=ncol(mat), replace=TRUE)
metadata$ESR <- sample.int(100, size=ncol(mat), replace=TRUE)

p <- pca(mat, metadata = metadata, removeVar = 0.1)

pairsplot(p, triangle = TRUE)
```

`parallelPCA` *Horn's parallel analysis*

Description

Perform Horn's parallel analysis to choose the number of principal components to retain.

Usage

```
parallelPCA(mat, max.rank=100, ..., niters=50, threshold=0.1,
  transposed=FALSE, BSPARAM=ExactParam(), BPPARAM=SerialParam())
```

Arguments

`mat`	A numeric matrix where rows correspond to variables and columns correspond to samples.
`max.rank`	Integer scalar specifying the maximum number of PCs to retain.
`...`	Further arguments to pass to `pca`.
`niters`	Integer scalar specifying the number of iterations to use for the parallel analysis.
`threshold`	Numeric scalar representing the "p-value" threshold above which PCs are to be ignored.
`transposed`	Logical scalar indicating whether `mat` is transposed, i.e., rows are samples and columns are variables.
`BSPARAM`	A BiocSingularParam object specifying the algorithm to use for PCA.
`BPPARAM`	A BiocParallelParam object specifying how the iterations should be paralellized.

Details

Horn's parallel analysis involves shuffling observations within each row of `x` to create a permuted matrix. PCA is performed on the permuted matrix to obtain the percentage of variance explained under a random null hypothesis. This is repeated over several iterations to obtain a distribution of curves on the scree plot.

For each PC, the "p-value" (for want of a better word) is defined as the proportion of iterations where the variance explained at that PC is greater than that observed with the original matrix. The number of PCs to retain is defined as the last PC where the p-value is below `threshold`. This aims to retain all PCs that explain "significantly" more variance than expected by chance.

This function can be sped up by specifying `BSPARAM=IrlbaParam()` or similar, to use approximate strategies for performing the PCA. Another option is to set `BPPARAM` to perform the iterations in parallel.

Value

A list is returned, containing:

- `original`, the output from running `pca` on `mat` with the specified arguments.
- `permuted`, a matrix of variance explained from randomly permuted matrices. Each column corresponds to a single permutated matrix, while each row corresponds to successive principal components.
- `n`, the estimated number of principal components to retain.

Author(s)

Aaron Lun

References

Horn JL. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika* 30(2), 179-185.

Buja A and Eyuboglu N (1992). Remarks on Parallel Analysis. *Multivariate Behav. Res.* 27, 509-40.

Examples

```
# Mocking up some data.
ngenes <- 1000
means <- 2^runif(ngenes, 6, 10)
dispersions <- 10/means + 0.2
nsamples <- 50
counts <- matrix(rnbinom(ngenes*nsamples, mu=means,
    size=1/dispersions), ncol=nsamples)

# Choosing the number of PCs
lcounts <- log2(counts + 1)
output <- parallelPCA(lcounts)
output$n
```

Description

Principal Component Analysis (PCA) is a very powerful technique that has wide applicability in data science, bioinformatics, and further afield. It was initially developed to analyse large volumes of data in order to tease out the differences/relationships between the logical entities being analysed. It extracts the fundamental structure of the data without the need to build any model to represent it. This 'summary' of the data is arrived at through a process of reduction that can transform the large number of variables into a lesser number that are uncorrelated (i.e. the 'principal components'), whilst at the same time being capable of easy interpretation on the original data. PCAtools provides functions for data exploration via PCA, and allows the user to generate publication-ready figures. PCA is performed via BiocSingular - users can also identify optimal number of principal component via different metrics, such as elbow method and Horn's parallel analysis, which has relevance for data reduction in single-cell RNA-seq (scRNA-seq) and high dimensional mass cytometry data.

Usage

```
pca(
  mat,
  metadata = NULL,
  center = TRUE,
  scale = FALSE,
  rank = NULL,
  removeVar = NULL,
  transposed = FALSE,
  BSPARAM = ExactParam())
```

Arguments

mat	A data-matrix or data-frame containing numerical data only. Variables are expected to be in the rows and samples in the columns by default. REQUIRED.
metadata	A data-matrix or data-frame containing metadata. This will be stored in the resulting pca object. Strictly enforced that rownames(metadata) == colnames(mat). DEFAULT = NULL. OPTIONAL.
center	Center the data before performing PCA? Same as prcomp() 'center' parameter. DEFAULT = TRUE. OPTIONAL.
scale	Scale the data? Same as prcomp() 'scale' parameter. DEFAULT = FALSE. OPTIONAL.
rank	An integer scalar specifying the number of PCs to retain. OPTIONAL for an exact SVD, whereby it defaults to all PCs. Otherwise REQUIRED for approximate SVD methods.
removeVar	Remove this % of variables based on low variance. DEFAULT = NULL. OPTIONAL.
transposed	Is mat transposed? DEFAULT = FALSE. If set to TRUE, samples are in the rows and variables are in the columns. OPTIONAL.
BSPARAM	A BiocSingularParam object specifying the algorithm to use for the SVD. Defaults to an exact SVD. OPTIONAL.

Details

mat is If transposed=TRUE, \

Value

A pca object, containing:

- rotated, a data frame of the rotated data, i.e., the centred and scaled (if either or both are requested) input data multiplied by the variable loadings ('loadings'). This is the same as the 'x' variable returned by prcomp().
- loadings, a data frame of variable loadings ('rotation' variable returned by prcomp()).
- variance, a numeric vector of the explained variation for each principal component.
- sdev, the standard deviations of the principal components.
- metadata, the original metadata
- xvars, a character vector of rownames from the input data.
- yvars, a character vector of colnames from the input data.
- components, a character vector of principal component / eigenvector names.

Author(s)

Kevin Blighe <kevin@clinicalbioinformatics.co.uk>, Aaron Lun

Examples

```
options(scipen=10)
options(digits=6)

col <- 20
row <- 20000
mat1 <- matrix(
  rexp(col*row, rate = 0.1),
  ncol = col)
rownames(mat1) <- paste0('gene', 1:nrow(mat1))
colnames(mat1) <- paste0('sample', 1:ncol(mat1))

mat2 <- matrix(
  rexp(col*row, rate = 0.1),
  ncol = col)
rownames(mat2) <- paste0('gene', 1:nrow(mat2))
colnames(mat2) <- paste0('sample', (ncol(mat1)+1):(ncol(mat1)+ncol(mat2)))

mat <- cbind(mat1, mat2)

metadata <- data.frame(row.names = colnames(mat))
metadata$Group <- rep(NA, ncol(mat))
metadata$Group[seq(1,40,2)] <- 'A'
metadata$Group[seq(2,40,2)] <- 'B'
metadata$CRP <- sample.int(100, size=ncol(mat), replace=TRUE)
metadata$ESR <- sample.int(100, size=ncol(mat), replace=TRUE)

p <- pca(mat, metadata = metadata, removeVar = 0.1)

getComponents(p)

getVars(p)

getLoadings(p)
```

```
screeplot(p)

screeplot(p, hline = 80)

biplot(p)

biplot(p, colby = 'Group', shape = 'Group')

biplot(p, colby = 'Group', colkey = c(A = 'forestgreen', B = 'gold'),
  legendPosition = 'right')

biplot(p, colby = 'Group', colkey = c(A='forestgreen', B='gold'),
  shape = 'Group', shapekey = c(A=10, B=21), legendPosition = 'bottom')

pairsplot(p, triangle = TRUE)

plotloadings(p, drawConnectors=TRUE)

eigencorplot(p, components = getComponents(p, 1:10),
  metavars = c('ESR', 'CRP'))
```

`plotloadings` *plotloadings*

Description

Plot the component loadings for selected principal components / eigenvectors and label variables driving variation along these.

Usage

```
plotloadings(pcaobj,
components = getComponents(pcaobj, seq_len(5)),
rangeRetain = 0.05,
absolute = FALSE,
col = c('gold', 'white', 'royalblue'),
colMidpoint = 0,
shape = 21,
shapeSizeRange = c(10, 10),
legendPosition = 'top',
legendLabSize = 10,
legendIconSize = 3.0,
xlim = NULL,
ylim = NULL,
labSize = 2.0,
labhjust = 1.5,
labvjust = 0,
drawConnectors = TRUE,
positionConnectors = 'right',
widthConnectors = 0.5,
typeConnectors = 'closed',
endsConnectors = 'first',
```

```
  lengthConnectors = unit(0.01, 'npc'),
  colConnectors = 'grey50',
  xlab = 'Principal component',
  xlabAngle = 0,
  xlabhjust = 0.5,
  xlabvjust = 0.5,
  ylab = 'Component loading',
  ylabAngle = 0,
  ylabhjust = 0.5,
  ylabvjust = 0.5,
  axisLabSize = 16,
  title = '',
  subtitle = '',
  caption = '',
  titleLabSize = 16,
  subtitleLabSize = 12,
  captionLabSize = 12,
  hline = c(0),
  hlineType = 'longdash',
  hlineCol = 'black',
  hlineWidth = 0.4,
  vline = NULL,
  vlineType = 'longdash',
  vlineCol = 'black',
  vlineWidth = 0.4,
  gridlines.major = TRUE,
  gridlines.minor = TRUE,
  borderWidth = 0.8,
  borderColour = 'black',
  returnPlot = TRUE)
```

Arguments

pcaobj	Object of class 'pca' created by pca(). REQUIRED.
components	The principal components to be included in the plot. DEFAULT = getComponents(pcaobj, seq_len(5)). OPTIONAL.
rangeRetain	Cut-off value for retaining variables. The function will look across each specified principal component and retain the variables that fall within this top/bottom fraction of the loadings range. DEFAULT = 0.05. OPTIONAL.
absolute	Logical, indicating whether or not to plot absolute loadings. DEFAULT = FALSE. OPTIONAL.
col	Colours used for generation of fill gradient according to loadings values. Can be 2 or 3 colours. DEFAULT = c('gold', 'white', 'royalblue'). OPTIONAL.
colMidpoint	Mid-point (loading) for the colour range. DEFAULT = 0. OPTIONAL.
shape	Shape of the plotted points. DEFAULT = 21. OPTIONAL.
shapeSizeRange	Size range for the plotted points (min, max). DEFAULT = c(10, 10). OPTIONAL.
legendPosition	Position of legend ('top', 'bottom', 'left', 'right', 'none'). DEFAULT = 'top'. OPTIONAL.
legendLabSize	Size of plot legend text. DEFAULT = 10. OPTIONAL.

legendIconSize	Size of plot legend icons / symbols. DEFAULT = 3.0. OPTIONAL.
xlim	Limits of the x-axis. DEFAULT = NULL. OPTIONAL.
ylim	Limits of the y-axis. DEFAULT = NULL. OPTIONAL.
labSize	Size of labels. DEFAULT = 2.0. OPTIONAL.
labhjust	Horizontal adjustment of label. DEFAULT = 1.5. OPTIONAL.
labvjust	Vertical adjustment of label. DEFAULT = 0. OPTIONAL.
drawConnectors	Logical, indicating whether or not to connect plot labels to their corresponding points by line connectors. DEFAULT = TRUE. OPTIONAL.
positionConnectors	Position of the connectors and their labels with respect to the plotted points ('left', 'right'). DEFAULT = 'right'. OPTIONAL.
widthConnectors	Line width of connectors. DEFAULT = 0.5. OPTIONAL.
typeConnectors	Have the arrow head open or filled ('closed')? ('open', 'closed'). DEFAULT = 'closed'. OPTIONAL.
endsConnectors	Which end of connectors to draw arrow head? ('last', 'first', 'both'). DEFAULT = 'first'. OPTIONAL.
lengthConnectors	Length of the connectors. DEFAULT = unit(0.01, 'npc'). OPTIONAL
colConnectors	Line colour of connectors. DEFAULT = 'grey50'. OPTIONAL.
xlab	Label for x-axis. DEFAULT = 'Principal component'. OPTIONAL.
xlabAngle	Rotation angle of x-axis labels. DEFAULT = 0. OPTIONAL.
xlabhjust	Horizontal adjustment of x-axis labels. DEFAULT = 0.5. OPTIONAL.
xlabvjust	Vertical adjustment of x-axis labels. DEFAULT = 0.5. OPTIONAL.
ylab	Label for y-axis. DEFAULT = 'Component loading'. OPTIONAL.
ylabAngle	Rotation angle of y-axis labels. DEFAULT = 0. OPTIONAL.
ylabhjust	Horizontal adjustment of y-axis labels. DEFAULT = 0.5. OPTIONAL.
ylabvjust	Vertical adjustment of y-axis labels. DEFAULT = 0.5. OPTIONAL.
axisLabSize	Size of x- and y-axis labels. DEFAULT = 16. OPTIONAL.
title	Plot title. DEFAULT = ''. OPTIONAL.
subtitle	Plot subtitle. DEFAULT = ''. OPTIONAL.
caption	Plot caption. DEFAULT = ''. OPTIONAL.
titleLabSize	Size of plot title. DEFAULT = 16. OPTIONAL.
subtitleLabSize	Size of plot subtitle. DEFAULT = 12. OPTIONAL.
captionLabSize	Size of plot caption. DEFAULT = 12. OPTIONAL.
hline	Draw one or more horizontal lines passing through this/these values on y-axis. For single values, only a single numerical value is necessary. For multiple lines, pass these as a vector, e.g., c(60,90). DEFAULT = c(0). OPTIONAL.
hlineType	Line type for hline ('blank', 'solid', 'dashed', 'dotted', 'dotdash', 'longdash', 'twodash'). DEFAULT = 'longdash'. OPTIONAL.
hlineCol	Colour of hline. DEFAULT = 'black'. OPTIONAL.
hlineWidth	Width of hline. DEFAULT = 0.4. OPTIONAL.

vline	Draw one or more vertical lines passing through this/these values on x-axis. For single values, only a single numerical value is necessary. For multiple lines, pass these as a vector, e.g., c(60,90). DEFAULT = NULL. OPTIONAL.
vlineType	Line type for vline ('blank', 'solid', 'dashed', 'dotted', 'dotdash', 'longdash', 'twodash'). DEFAULT = 'longdash'. OPTIONAL.
vlineCol	Colour of vline. DEFAULT = 'black'. OPTIONAL.
vlineWidth	Width of vline. DEFAULT = 0.4. OPTIONAL.
gridlines.major	Logical, indicating whether or not to draw major gridlines. DEFAULT = TRUE. OPTIONAL.
gridlines.minor	Logical, indicating whether or not to draw minor gridlines. DEFAULT = TRUE. OPTIONAL.
borderWidth	Width of the border on the x and y axes. DEFAULT = 0.8. OPTIONAL.
borderColour	Colour of the border on the x and y axes. DEFAULT = 'black'. OPTIONAL.
returnPlot	Logical, indicating whether or not to return the plot object. DEFAULT = TRUE. OPTIONAL.

Value

A ggplot2 object.

Author(s)

Kevin Blighe <kevin@clinicalbioinformatics.co.uk>

Examples

```
options(scipen=10)
options(digits=6)

col <- 20
row <- 20000
mat1 <- matrix(
  rexp(col*row, rate = 0.1),
  ncol = col)
rownames(mat1) <- paste0('gene', 1:nrow(mat1))
colnames(mat1) <- paste0('sample', 1:ncol(mat1))

mat2 <- matrix(
  rexp(col*row, rate = 0.1),
  ncol = col)
rownames(mat2) <- paste0('gene', 1:nrow(mat2))
colnames(mat2) <- paste0('sample', (ncol(mat1)+1):(ncol(mat1)+ncol(mat2)))

mat <- cbind(mat1, mat2)

metadata <- data.frame(row.names = colnames(mat))
metadata$Group <- rep(NA, ncol(mat))
metadata$Group[seq(1,40,2)] <- 'A'
metadata$Group[seq(2,40,2)] <- 'B'
metadata$CRP <- sample.int(100, size=ncol(mat), replace=TRUE)
metadata$ESR <- sample.int(100, size=ncol(mat), replace=TRUE)
```

```
p <- pca(mat, metadata = metadata, removeVar = 0.1)

plotloadings(p, drawConnectors=TRUE)
```

---

screeplot *screeplot*

Description

Draw a SCREE plot, showing the distribution of explained variance across all or select principal components / eigenvectors.

Usage

```
screeplot(pcaobj,
components = getComponents(pcaobj),
xlim = NULL,
ylim = c(0, 100),
xlab = 'Principal component',
xlabAngle = 90,
xlabhjust = 0.5,
xlabvjust = 0.5,
ylab = 'Explained variation (%)',
ylabAngle = 0,
ylabhjust = 0.5,
ylabvjust = 0.5,
axisLabSize = 16,
title = 'SCREE plot',
subtitle = '',
caption = '',
titleLabSize = 16,
subtitleLabSize = 12,
captionLabSize = 12,
colBar = 'dodgerblue',
drawCumulativeSumLine = TRUE,
colCumulativeSumLine = 'red2',
sizeCumulativeSumLine = 1.5,
drawCumulativeSumPoints = TRUE,
colCumulativeSumPoints = 'red2',
sizeCumulativeSumPoints = 2.0,
hline = NULL,
hlineType = 'longdash',
hlineCol = 'black',
hlineWidth = 0.4,
vline = NULL,
vlineType = 'longdash',
vlineCol = 'black',
vlineWidth = 0.4,
gridlines.major = TRUE,
gridlines.minor = TRUE,
```

```
  borderWidth = 0.8,
  borderColour = 'black',
  returnPlot = TRUE)
```

Arguments

pcaobj	Object of class 'pca' created by pca(). REQUIRED.
components	The principal components to be included in the plot. DEFAULT = getComponents(pcaobj). OPTIONAL.
xlim	Limits of the x-axis. DEFAULT = NULL. OPTIONAL.
ylim	Limits of the y-axis. DEFAULT = c(0, 100). OPTIONAL.
xlab	Label for x-axis. DEFAULT = 'Principal component'. OPTIONAL.
xlabAngle	Rotation angle of x-axis labels. DEFAULT = 90. OPTIONAL.
xlabhjust	Horizontal adjustment of x-axis labels. DEFAULT = 0.5. OPTIONAL.
xlabvjust	Vertical adjustment of x-axis labels. DEFAULT = 0.5. OPTIONAL.
ylab	Label for y-axis. DEFAULT = 'Explained variation (%)'. OPTIONAL.
ylabAngle	Rotation angle of y-axis labels. DEFAULT = 0. OPTIONAL.
ylabhjust	Horizontal adjustment of y-axis labels. DEFAULT = 0.5. OPTIONAL.
ylabvjust	Vertical adjustment of y-axis labels. DEFAULT = 0.5. OPTIONAL.
axisLabSize	Size of x- and y-axis labels. DEFAULT = 16. OPTIONAL.
title	Plot title. DEFAULT = 'SCREE plot'. OPTIONAL.
subtitle	Plot subtitle. DEFAULT = ''. OPTIONAL.
caption	Plot caption. DEFAULT = ''. OPTIONAL.
titleLabSize	Size of plot title. DEFAULT = 16. OPTIONAL.
subtitleLabSize	Size of plot subtitle. DEFAULT = 12. OPTIONAL.
captionLabSize	Size of plot caption. DEFAULT = 12. OPTIONAL.
colBar	DEFAULT = 'dodgerblue'. OPTIONAL.
drawCumulativeSumLine	Logical, indicating whether or not to overlay plot with a cumulative explained variance line. DEFAULT = TRUE. OPTIONAL.
colCumulativeSumLine	Colour of cumulative explained variance line. DEFAULT = 'red2'. OPTIONAL.
sizeCumulativeSumLine	Size of cumulative explained variance line. DEFAULT = 1.5. OPTIONAL.
drawCumulativeSumPoints	Logical, indicating whether or not to draw the cumulative explained variance points. DEFAULT = TRUE. OPTIONAL.
colCumulativeSumPoints	Colour of cumulative explained variance points. DEFAULT = 'red2'. OPTIONAL.
sizeCumulativeSumPoints	Size of cumulative explained variance points. DEFAULT = 2.0. OPTIONAL.
hline	Draw one or more horizontal lines passing through this/these values on y-axis. For single values, only a single numerical value is necessary. For multiple lines, pass these as a vector, e.g., c(60,90). DEFAULT = NULL. OPTIONAL.

hlineType	Line type for hline ('blank', 'solid', 'dashed', 'dotted', 'dotdash', 'longdash', 'twodash'). DEFAULT = 'longdash'. OPTIONAL.
hlineCol	Colour of hline. DEFAULT = 'black'. OPTIONAL.
hlineWidth	Width of hline. DEFAULT = 0.4. OPTIONAL.
vline	Draw one or more vertical lines passing through this/these values on x-axis. For single values, only a single numerical value is necessary. For multiple lines, pass these as a vector, e.g., c(60,90). DEFAULT = NULL. OPTIONAL.
vlineType	Line type for vline ('blank', 'solid', 'dashed', 'dotted', 'dotdash', 'longdash', 'twodash'). DEFAULT = 'longdash'. OPTIONAL.
vlineCol	Colour of vline. DEFAULT = 'black'. OPTIONAL.
vlineWidth	Width of vline. DEFAULT = 0.4. OPTIONAL.
gridlines.major	Logical, indicating whether or not to draw major gridlines. DEFAULT = TRUE. OPTIONAL.
gridlines.minor	Logical, indicating whether or not to draw minor gridlines. DEFAULT = TRUE. OPTIONAL.
borderWidth	Width of the border on the x and y axes. DEFAULT = 0.8. OPTIONAL.
borderColour	Colour of the border on the x and y axes. DEFAULT = 'black'. OPTIONAL.
returnPlot	Logical, indicating whether or not to return the plot object. DEFAULT = TRUE. OPTIONAL.

Value

A ggplot2 object.

Author(s)

Kevin Blighe <kevin@clinicalbioinformatics.co.uk>

Examples

```
options(scipen=10)
options(digits=6)

col <- 20
row <- 20000
mat1 <- matrix(
  rexp(col*row, rate = 0.1),
  ncol = col)
rownames(mat1) <- paste0('gene', 1:nrow(mat1))
colnames(mat1) <- paste0('sample', 1:ncol(mat1))

mat2 <- matrix(
  rexp(col*row, rate = 0.1),
  ncol = col)
rownames(mat2) <- paste0('gene', 1:nrow(mat2))
colnames(mat2) <- paste0('sample', (ncol(mat1)+1):(ncol(mat1)+ncol(mat2)))

mat <- cbind(mat1, mat2)

metadata <- data.frame(row.names = colnames(mat))
```

```
metadata$Group <- rep(NA, ncol(mat))
metadata$Group[seq(1,40,2)] <- 'A'
metadata$Group[seq(2,40,2)] <- 'B'
metadata$CRP <- sample.int(100, size=ncol(mat), replace=TRUE)
metadata$ESR <- sample.int(100, size=ncol(mat), replace=TRUE)

p <- pca(mat, metadata = metadata, removeVar = 0.1)

screeplot(p)

screeplot(p, hline = 80)
```

Index