Package 'dCUR'

October 8, 2020

Type Package

Title Dimension Reduction with Dynamic CUR

Version 1.0.0

Depends R (>= 3.5.0)

Maintainer Cesar Gamboa <info@cesargamboasanabria.com>

URL https://www.cesargamboasanabria.com

Author Cesar Gamboa-Sanabria [aut, mdc, cph, cre] (<https://orcid.org/0000-0001-6733-4759>),
Stefani Matarrita-Munoz [aut] (<https://orcid.org/0000-0002-1222-1981>),
Katherine Barquero-Mejias [aut] (<https://orcid.org/0000-0003-1760-9026>),
Greibin Villegas-Barahona [aut] (<https://orcid.org/0000-0002-4380-0812>),
Mercedes Sanchez-Barba [aqt] (<https://orcid.org/0000-0002-3324-5798>),
Maria Purificacion Galindo-Villardon [aqt] (<https://orcid.org/0000-0001-6977-7545>)

Description Dynamic CUR (dCUR) boosts the CUR decomposition (Mahoney MW., Drineas P. (2009) <doi:10.1073/pnas.0803205106>) varying the k, the number of columns and rows used, and its final purposes to help find the stage, which minimizes the relative error to reduce matrix dimension.
The goal of CUR Decomposition is to give a better interpretation of the matrix decomposition employing proper variable selection in the data matrix, in a way that yields a simplified structure. Its origins come from analysis in genetics.
The goal of this package is to show an alternative to variable selection (columns) or individuals (rows). The idea proposed consists of adjusting the probability distributions to the leverage scores and selecting the best columns and rows that minimize the reconstruction error of the matrix approximation ||A-CUR||. It also includes a method that recalibrates the relative importance of the leverage scores according to an external variable of the user's interest.

Imports parallel, magrittr, stackoverflow, mclust, MASS, ppcor, ggplot2, dplyr, Rdpack

License GPL-3

Encoding UTF-8

LazyData true

RdMacros Rdpack

RoxygenNote 7.1.1

Suggests testthat, snow

NeedsCompilation no

Repository CRAN

Date/Publication 2020-10-08 11:30:02 UTC

R topics documented:

AASP *Academic Achievement Score Projection -AASP-*

Description

Data from a Ph.D. thesis about Academic Achievement Score Projection, with 632 rows and 205 columns.

Usage

```
data("AASP")
```

Source

Further information about variables can be found in this link:

https://github.com/cgamboasanabria/dCUR/tree/master/inst/AASP_description_data.xlsx

Examples

```
data(AASP)
dim(AASP)
```

CUR *CUR*

Description

This function computes the canonical CUR decomposition using top scores as selection criteria to identify the most relevant columns and rows of a given data matrix. It also provides an option to use an extension of CUR decomposition, which reconfigures leverage scores by using the partial and semi partial correlations with an external variable of interest. Additionally, this function lets the user fit a probability distribution of leverage scores with Mixtures Gaussian Models.

Usage

```
CUR(
  data,
  variables,
  k = NULL,
  rows,
  columns,
  standardize = FALSE,
  cur_method = "sample_cur",
  correlation = NULL,
  correlation_type = c("partial", "semipartial"),
  ...
)
```

Arguments

`data`	a data frame containing the variables to be used in CUR decomposition and other externals variables with which you want to correlate.
`variables`	correspond to the variables used to compute the leverage scores in CUR analysis. The external variable's names must not be included. dplyr package notation can be used to specify the variables (see the example).
`k`	corresponds to the number of principal components used to compute the leverage scores. If NULL, it is considered the number of k main components that accumulate 80% of the variability explained. This argument can also be a proportion, in which case the function takes this value as the desired cumulative explained variance and automatically chooses the k.
`rows`	correspond to the proportion of rows to be selected from the total number of rows in the data matrix. When all the rows are needed and `mixture` is used as cur_method, a proportion of 0.999 must be used.
`columns`	correspond to the proportion of columns (variables) to be selected from the total number of variables in the data matrix.
`standardize`	If `TRUE` the data is standardized (by subtracting the average and dividing by the standard deviation)

`cur_method` character. If `sample_cur`, the selection of leverage scores is made according to the top score selection criteria set out by Mahoney & Drineas (2009). If `mixture` method is specified, the best Mixture Gaussian Model is fitted for the leverages, and the selection of the most relevant variables is based on a tabular value given the critical area specified in rows and columns arguments.

`correlation` character. It specifies the name of the external variable the computation of leverage must be adjusted with.

`correlation_type` character. It specifies if the computation of leverage must be adjusted by the `semipartial` or `partial` correlation with an external variable.

`...` additional arguments to be passed to `pcor` or `spcor`

Details

Extension of classic CUR descomposition with top scores selection criteria.

CUR decomposition chooses columns and rows that exhibit high leverage scores and exert a disproportionately large "influence" on the best low-rank fit of the data matrix. The main advantage of CUR Decomposition over SVD is that the original data matrix can be expressed as a reduced number of rows and columns instead of obtaining factorial axes resulting from a linear combination of all the original variables to facilitate interpretation.

The reconfiguration of the leverage scores according to the methodology of Villegas et al. (2018) dividing the leverage score by $(1 - \rho^2)$. Where ρ rho represents the partial or semi-partial correlation that the variables used in CUR decomposition have with an external variable, its purpose is recalibrating the relative importance of the leverage scores according to an external variable of interest.

The correlation type selection could be partial or semi-partial, according to Seongho (2015) of the package in R ppcor.

Value

`k` Number of principal components with which leverages scores are computed.

`CUR` CUR matrix.

`absolute_error` Absolute error computed as the Frobenius norm of the original data -detnoted as A- and CUR matrix: ‖A-CUR‖

`relative_error` Relative error $\frac{||A-CUR||}{||A||}$

`leverage_columns_sorted` a data frame which specifies the names of relevant columns and its leverages scores arranged downwardly.

`leverage_rows_sorted` a data frame which specifies the number of relevant rows and its leverages scores arranged downwardly.

`leverage_columns` a data frame which specifies the names of all columns and its leverages scores.

`leverage_rows` a data frame which specifies the number of all rows and its leverages scores.

Author(s)

Cesar Gamboa-Sanabria, Stefany Matarrita-Munoz, Katherine Barquero-Mejias, Greibin Villegas-Barahona, Mercedes Sanchez-Barba and Maria Purificacion Galindo-Villardon.

References

Mahoney MW, Drineas P (2009). "CUR matrix decompositions for improved data analysis." *Proceedings of the National Academy of Sciences*, **106**(3), 697–702. ISSN 0027-8424, doi: 10.1073/pnas.0803205106. Villegas G, others (2018). "Modelo estadistico pedagogico para la toma de decisiones administrativas y academicas con impacto en el mejoramiento continuo del rendimiento de los estudiantes universitarios, basado en los metodos de seleccion CUR." doi: 10.14201/gredos.139405. Villegas G, Martin-Barreiro C, Gonzalez-Garcia N, Hernandez-Gonzalez S, Sanchez-Barba M, Galindo-Villardon M (2019). "Dynamic CUR, an alternative to variable selection in CUR decomposition." *Revistas Investigacion Operacional*, **40**(3), 391–399. https://rev-inv-ope.univ-paris1.fr/fileadmin/rev-inv-ope/files/40319/40319-09.pdf. Drineas P, Mahoney MW, Muthukrishnan S (2008). "Relative-error cur matrix decompositions." *SIAM Journal on Matrix Analysis and Applications*, **30**(2), 844–881. https://doi.org/10.1137/07070471X.

Examples

```
#Classic CUR with top scores selection criteria.
result <- CUR(data=AASP, variables=hoessem:notabachillerato,
          k=20, rows = 1, columns = .2, standardize = TRUE,
          cur_method = "sample_cur")
result
#Extension of classic CUR: Recalibrating leverages scores
#and adjusting a mixtures Gaussian models to leverages.
result <- CUR(data=AASP, variables=hoessem:notabachillerato,
          k=20, rows = 1, columns = .2, standardize = TRUE,
          cur_method = "mixture",
          correlation = R1, correlation_type = "partial")
result
```

---

dCUR *dCUR*

Description

Dynamic CUR is a function that boosts the CUR decomposition varying the k, number of columns, and rows used. Its ultimate purpose is to find the stage which minimizes the relative error. The classic CUR and its extensions can be used in dCUR.

Dynamic CUR is an r package that boosts the CUR decomposition varying the k, the number of columns and rows used, and its final purposes to help find the stage, which minimizes the relative error to reduce matrix dimension. Mahoney & Drineas (2009) identified the singular vectors of the SVD as the PCs' interpretation problem and proposed another type of matrix factorization known

as CUR Decomposition (Mahoney & Drineas, 2009; Mahoney, Maggioni, & Drineas, 2008; Bodor, Csabai, Mahoney, & Solymosi, 2012). The goal of CUR Decomposition is to give a better interpretation of the matrix decomposition employing proper variable selection in the data matrix, in a way that yields a simplified structure. Its origins come from analysis in genetics. One example is the one showed in Mahoney & Drineas (2009), in which cancer microarrays highlighted to recognize, based on 5000 variables, genetic patterns in patients with soft tissue tumors analyzed with cDNA microarrays. The objective of this package is to show an alternative to variable selection (columns) or individuals (rows) to the ones developed by Mahoney & Drineas (2009). The idea proposed consists of adjusting the probability distributions to the leverage scores and selecting the best columns and rows that minimize the reconstruction error of the matrix approximation \|A-CUR\|. It also includes a method that recalibrates the relative importance of the leverage scores according to an external variable of the user's interest.

Usage

```
dCUR(
  data,
  variables,
  standardize = FALSE,
  dynamic_columns = FALSE,
  dynamic_rows = FALSE,
  parallelize = FALSE,
  skip = 0.05,
  ...
)
```

Arguments

`data`	a data frame that contains the variables to use in CUR decomposition and other externals variables with which you want to correlate.
`variables`	correspond to the variables used to compute the leverage scores in CUR analysis. The external variable's names must not be included. dplyr package notation can be used to specify the variables (see examples).
`standardize`	logical. If `TRUE` the data is standardized (by subtracting the average and dividing by the standard deviation)
`dynamic_columns`	logical. If `TRUE`, an iterative process begins where leverage scores are computed for the different values from 1 to k main components, as well as from 1 to c (the proportion of columns to be selected from the data matrix).
`dynamic_rows`	logical. If `TRUE`, an iterative process begins where leverage scores are computed for the different values from 1 to k main components, as well as from 1 to r (the proportion of rows to be selected from the data matrix).
`parallelize`	logical.If `TRUE` the CUR analysis is parallelized.
`skip`	numeric. It specifies the change ratio of columns and rows to be selected.
`...`	additional arguments to be passed to `CUR`.

Details

This function serves as a basis for selecting the best combination of k (principal components), c (number of columns) and r (number of rows), in other words, the stage that minimizes the relative error $\frac{||A-CUR||}{||A||}$, and thus optimizes the number of columns in the analysis, ensuring a percentage of explained variability of the data matrix and facilitating the interpretation of the data set by reducing the dimensionality of the original matrix.

If `skip = 0.1` for each k, it is tested with a column proportion of 0, 0.1, 0.11,0.22,...; the same applies for rows. Given the above, it is recommended not to choose a tiny skip, since this implies doing the CUR analysis for more stages.

Parallelizing the function improves its speed significantly.

Value

CUR returns a list of lists, each one represents a stage, and it contains:

`k`	Number of principal components with which leverages scores are computed.
`columns`	number of columns selected.
`rows`	number of rows selected.
`relative_error`	relative_error obtained: $\frac{||A-CUR||}{||A||}$

Author(s)

Cesar Gamboa-Sanabria, Stefany Matarrita-Munoz, Katherine Barquero-Mejias, Greibin Villegas-Barahona, Mercedes Sanchez-Barba and Maria Purificacion Galindo-Villardon.

Cesar Gamboa-Sanabria <info@cesargamboasanabria.com>

See Also

`CUR` `optimal_stage`

Examples

```
results <- dCUR::dCUR(data=AASP, variables=hoessem:notabachillerato,
k=15, rows=0.25, columns=0.25,skip = 0.1, standardize=TRUE,
cur_method="sample_cur",
parallelize =TRUE, dynamic_columns  = TRUE,
dynamic_rows  = TRUE)
results
```

mixture_plots *mixture_plots*

Description

This function returns different plots associated with the fitting of leverages scores through Mixture Gaussian Models.

Usage

```
mixture_plots(data)
```

Arguments

data	An object resulting from a call to CUR when "mixture" is specified as cur_method.

Details

Gaussian Mixture Models Plots

Value

mixture_plots returns a list with the following plots:

BIC	BIC Plot of the Bayesian Information Criterion (BIC) for each number of mixture components. E and V stands for equal variance in mixture components or variable variance, respectively.
density	leverages score's density
Cumulative	cumulative density of leverages scores.
QQPlot	Plot the sample quantiles and controlled quantiles of the inverse of the cumulative distribution function.

Author(s)

Cesar Gamboa-Sanabria, Stefany Matarrita-Munoz, Katherine Barquero-Mejias, Greibin Villegas-Barahona, Mercedes Sanchez-Barba and Maria Purificacion Galindo-Villardon.

References

Mahoney MW, Drineas P (2009). "CUR matrix decompositions for improved data analysis." *Proceedings of the National Academy of Sciences*, **106**(3), 697–702. ISSN 0027-8424, doi: 10.1073/pnas.0803205106. Villegas G, others (2018). "Modelo estadistico pedagogico para la toma de decisiones administrativas y academicas con impacto en el mejoramiento continuo del rendimiento de los estudiantes universitarios, basado en los metodos de seleccion CUR." doi: 10.14201/gredos.139405. Villegas G, Martin-Barreiro C, Gonzalez-Garcia N, Hernandez-Gonzalez S, Sanchez-Barba M, Galindo-Villardon M (2019). "Dynamic CUR, an alternative to variable selection in CUR decomposition." *Revistas Investigacion Operacional*, **40**(3), 391–399. https://rev-inv-ope.univ-paris1.fr/fileadmin/rev-inv-ope/files/40319/40319-09.pdf.

See Also

dCUR CUR

Examples

```
results <- CUR(data=AASP, variables=hoessem:notabachillerato,
k=20, rows = .9999999, columns = .10, standardize = TRUE,
cur_method = "mixture")
mixture_plots(results)
```

---

optimal_stage *optimal_stage*

Description

optimal_stage is a function used to select the optimal k, the number of columns and rows of dynamic CUR object; it also produces a data frame and corresponding plots.

Usage

```
optimal_stage(data, limit = 80)
```

Arguments

data	An object resulting from a call to dCUR.
limit	Cumulative percentage average of relative error rate.

Details

Select the optimal stage of dynamic CUR descomposition

The objective of CUR decomposition is to find the most relevant variables and observations within a data matrix to reduce the dimensionality. It is well known that as more columns (variables) and rows are selected, the relative error will decrease; however, this is not true for k (number of components to compute leverages). Given the above, this function seeks to find the best-balanced stage of k, the number of relevant columns, and rows that have an error very close to the minimum, but at the same time maintain the low-rank fit of the data matrix.

Value

data	a data frame which specifies the relative error for each stage of CUR decomposition.
rows_plot	a plot where the average relative error is shown for each number of relevant rows selected.

columns_plot	a plot where the average relative error is shown for each number of relevant columns selected.
k_plot	a plot where the average relative error is shown for each k (number of components to compute leverage), given the optimal number of relevant columns and rows.
optimal	a data frame where the average relative error is shown for optimal k (number of components to compute leverage), given the optimal number of relevant columns and rows.

Author(s)

Cesar Gamboa-Sanabria, Stefany Matarrita-Munoz, Katherine Barquero-Mejias, Greibin Villegas-Barahona, Mercedes Sanchez-Barba and Maria Purificacion Galindo-Villardon.

References

Villegas G, Martin-Barreiro C, Gonzalez-Garcia N, Hernandez-Gonzalez S, Sanchez-Barba M, Galindo-Villardon M (2019). "Dynamic CUR, an alternative to variable selection in CUR decomposition." *Revistas Investigacion Operacional*, **40**(3), 391–399. https://rev-inv-ope.univ-paris1.fr/fileadmin/rev-inv-ope/files/40319/40319-09.pdf.

See Also

dCUR CUR

Examples

```
results <- dCUR(data=AASP, variables=hoessem:notabachillerato,
k=15, rows=0.25, columns=0.25,skip = 0.1, standardize=TRUE,
cur_method="sample_cur",
parallelize =TRUE, dynamic_columns  = TRUE,
dynamic_rows  = TRUE)
result <- optimal_stage(results, limit = 80)
result
result$k_plot
result$columns_plot
result$data
result$optimal
```

`relevant_variables_plot`

relevant_variables_plot

Description

relevant_variables_plot returns a bar graph which contains the leverages of the most relevant variable of data matrix according to CUR decomposition.

Usage

```
relevant_variables_plot(data)
```

Arguments

data	An object resulting from a call to CUR.

Details

Relevant Variables Plot

Author(s)

Cesar Gamboa-Sanabria, Stefany Matarrita-Munoz, Katherine Barquero-Mejias, Greibin Villegas-Barahona, Mercedes Sanchez-Barba and Maria Purificacion Galindo-Villardon.

References

Villegas G, Martin-Barreiro C, Gonzalez-Garcia N, Hernandez-Gonzalez S, Sanchez-Barba M, Galindo-Villardon M (2019). "Dynamic CUR, an alternative to variable selection in CUR decomposition." *Revistas Investigacion Operacional*, **40**(3), 391–399. https://rev-inv-ope.univ-paris1.fr/fileadmin/rev-inv-ope/files/40319/40319-09.pdf.

See Also

dCUR CUR

Examples

```
result <- CUR(data=AASP, variables=hoessem:notabachillerato,
k=20, rows = 1, columns = .2, standardize = TRUE,
cur_method = "sample_cur")
relevant_variables_plot(result)
```

var_exp *var_exp*

Description

var_exp is used to compute the proportion of the fraction of variance explained by a principal component analysis.

Usage

```
var_exp(data, standardize = FALSE, ...)
```

Arguments

data	a data frame that contains the variables to be used in CUR decomposition.
standardize	logical. If TRUE rescale an original data frame to have a mean of zero and a standard deviation of one.
...	Additional arguments to be passed to dplyr::select

Details

The objective of CUR decomposition is to find the most relevant variables and observations within a data matrix and to reduce the dimensionality. It is well known that as more columns (variables) and rows are selected, the relative error will be lower; however, this is not true for k (number of components to calculate leverages). Given the above, this function seeks to find the best-balanced scenario of k, the number of relevant columns, and rows that have an error very close to the minimum, and that, in turn, uses a smaller amount of information.

Value

var_exp	a data frame with the proportion of explained variance for each principal component.

Author(s)

Cesar Gamboa-Sanabria, Stefany Matarrita-Munoz, Katherine Barquero-Mejias, Greibin Villegas-Barahona, Mercedes Sanchez-Barba and Maria Purificacion Galindo-Villardon.

References

Mahoney MW, Drineas P (2009). "CUR matrix decompositions for improved data analysis." *Proceedings of the National Academy of Sciences*, **106**(3), 697–702. ISSN 0027-8424, doi: 10.1073/pnas.0803205106. Villegas G, others (2018). "Modelo estadistico pedagogico para la toma de decisiones administrativas y academicas con impacto en el mejoramiento continuo del rendimiento de los estudiantes universitarios, basado en los metodos de seleccion CUR." doi: 10.14201/gredos.139405. Villegas G, Martin-Barreiro C, Gonzalez-Garcia N, Hernandez-Gonzalez S, Sanchez-Barba M, Galindo-Villardon M (2019). "Dynamic CUR, an alternative to variable selection in CUR decomposition." *Revistas Investigacion Operacional*, **40**(3), 391–399. https://rev-inv-ope.univ-paris1.fr/fileadmin/rev-inv-ope/files/40319/40319-09.pdf.

See Also

dCUR CUR

Examples

```
var_exp(AASP, standardize = TRUE, hoessem:notabachillerato)
```

Index