Package ‘lakhesis’

June 10, 2024

Title Consensus Seriation for Binary Data

Version 0.0.1

Description Determining consensus seriations for binary incidence matrices, using a two-step process of Procrustes-fit correspondence analysis for heuristic selection of partial seriations and iterative regression to establish a single consensus. Contains the Lakhesis Calculator, a graphical platform for identifying seriated sequences. Collins-Elliott (2024) <https://volweb.utk.edu/~scolli46/sceLakhesis.pdf>.

License GPL (>= 3)

Imports stats, readr, ca, ggplot2, Rdpack, shiny, shinydashboard, bslib

RdMacros Rdpack

Encoding UTF-8

RoxygenNote 7.3.1

Depends R (>= 2.10)

LazyData true

NeedsCompilation no

Author Stephen A. Collins-Elliott [aut, cre] (<https://orcid.org/0000-0002-5642-6903>)

Maintainer Stephen A. Collins-Elliott <sce@utk.edu>

Repository CRAN

Date/Publication 2024-06-10 17:10:05 UTC

Contents

`ca.procrustes` *Correspondence Analysis with Procrustes Fitting*

Description

Fit scores of correspondence analysis on an incidence matrix to those produced by reference matrix which contain an ideal seriation using a Procrustes method (on the reference matrix, see `im.ref`). Rotation is determined by minimizing Euclidean distance from each row score to the nearest reference row score. Correspondence analysis is performed using the `ca` package (Nenadic and Greenacre 2007).

Usage

```
ca.procrustes(obj)
```

Arguments

`obj`	An incidence matrix of size n x k.

Value

A list of the following:.

- `ref` The Procrustes-fit coordinates of the scores of the reference seriation.
- `x` The Procrustes-fit coordinates of the row scores of the data.
- `x.dat` A data frame containing the following information related to the fit of the row score after Procrustes fitting: `index`, the row name, `match`, the reference point nearest to the row score, and `dist`, the Euclidean distance between the row score and reference score point.
- `y` The Procrustes-fit coordinates of the column scores of the data.
- `y.dat` A data frame containing the same information as `x.dat`, but with respect to the column scores.

References

Nenadic O, Greenacre MJ (2007). “Correspondence Analysis in R, with Two- and Three-dimensional Graphics: The ca Package.” *Journal of Statistical Software*, **20**, 1–13. doi:10.18637/jss.v020.i03.

Examples

```
data("quattrofontanili")
ca.procrustes(quattrofontanili)
```

`ca.procrustes.curve` *Seriate Using Reference Curve*

Description

Obtain a ranking of row and column scores projected onto a reference curve of an ideal seriation (row and column scores are ranked separately). Scores of correspondence analysis have been fit to those produced by reference matrix contain an ideal seriation using a Procrustes method, projecting them. Rotation is determined by minimizing Euclidean distance from each row score to the nearest reference row score. Correspondence analysis is performed using the `ca` package (Nenadic and Greenacre 2007).

Usage

```
ca.procrustes.curve(obj, resolution = 10000)
```

Arguments

`obj`	An incidence matrix of size n x k.
`resolution`	Number of samples to use for plotting points along polynomial curve (default is 10000).

Value

A data frame of the following:.

- `Procrustes1,Procrustes2` The location of the point on the biplot after fitting.
- `CurveIndex` The orthogonal projection of the point onto the reference curve, given as the index of the point sampled along $y = \beta_2 x^2 + \beta_0$.
- `Distance` The squared Euclidean distance of the point to the nearest point on the reference curve.
- `Rank` The ranking of the row or column, a range of `1:nrow`` and `1:ncol“.
- `Type` Either `row` or `col`.
- `sel` Data frame column used in `shiny` app to indicate whether point is selected in biplot/curve projection.

References

Nenadic O, Greenacre MJ (2007). “Correspondence Analysis in R, with Two- and Three-dimensional Graphics: The ca Package.” *Journal of Statistical Software*, **20**, 1–13. doi:10.18637/jss.v020.i03.

Examples

```
data("quattrofontanili")
ca.procrustes.curve(quattrofontanili)
```

ca.procrustes.poly *Projection onto Reference Curve*

Description

Performs a polynomial regression on the row reference scores and orthogonally projects data points on to the reference curve. Sampling can be increased to refine ranking and avoid ties, but default is largely sufficient. Correspondence analysis is performed using the `ca` package (Nenadic and Greenacre 2007).

Usage

```
ca.procrustes.poly(obj, resolution = 10000)
```

Arguments

`obj`	An incidence matrix of size n x k.
`resolution`	Number of samples to use for plotting points along polynomial curve (default is 10000).

Value

A list of the following:.

- 'ref" The Procrustes-fit coordinates of the scores of the reference seriation.
- `x` The Procrustes-fit coordinates of the row scores of the data.
- `x.dat` A data frame containing the following information related to the fit of the row score after Procrustes fitting: `index`, the row name, `match`, the reference point nearest to the row score, and `dist`, the Euclidean distance between the row score and reference score point.
- `y` The Procrustes-fit coordinates of the column scores of the data.
- `y.dat` A data frame containing the same information as `x.dat`, but with respect to the column scores.

References

Nenadic O, Greenacre MJ (2007). "Correspondence Analysis in R, with Two- and Three-dimensional Graphics: The ca Package." *Journal of Statistical Software*, **20**, 1–13. doi:10.18637/jss.v020.i03.

Examples

```
data("quattrofontanili")
ca.procrustes.poly(quattrofontanili)
```

`concentration.col` *Kendall-Doran Concentration*

Description

The Kendall-Doran measure of concentration (Kendall 1963; Doran 1971). In a seriated matrix, this function computes the total number cells between the first and last non-zero value, column by column.

Usage

```
concentration.col(obj)
```

Arguments

`obj`	A seriated binary matrix.

Value

The measure of concentration.

References

Doran J (1971). "Computer Analysis of Data from the la Tène Cemetry at Münsingen-Rain." In Hodson FR, Kendall DG, Táutu P (eds.), *Mathematics in the Archaeological and Historical Sciences*, 422–431. Edinburgh University Press, Edinburgh.

Kendall DG (1963). "A Statistical Approach to Flinders Petrie's Sequence Dating." *Bulletin of the International Statistical Institute*, **40**, 657–680.

Examples

```
data("quattrofontanili")
concentration.col(quattrofontanili)
```

`element.eval` *Evaluating Element Fit*

Description

Performs a goodness-of-fit test on individual row and column elements using deviance, using a quadratic-logistic model to fit row and column occurrences. In the case of perfect separation of 0/1 values, an `NA` value is assigned. Results are reported as p values for each row and column.

Usage

```
element.eval(obj)
```

Arguments

obj	A seriated binary matrix.

Value

A `list` containing results in frames for row and column elements:

- `RowFit` a data frame containing
  - `id` Row element
  - `p.val` p values of the row elements
- `ColFit` a data frame containing
  - `id` Column element
  - `p.val` p values of the column elements

Examples

```
data("quattrofontanili")
element.eval(quattrofontanili)
```

im.csv.read *Read csv File to Incidence Matrix*

Description

Wrapper around the `read_csv` function from the `readr` package (Wickham et al. 2024). Read a `.csv` file in which the first column represents row elements and the second column represents column elements, and convert it into an incidence matrix.

Usage

```
im.csv.read(
  filename,
  header = FALSE,
  characterencoding = "iso-8859-1",
  remove.hapax = FALSE
)
```

Arguments

`filename` The filename to uploaded (must be in `.csv` format).

`header` If the `.csv` file contains a header. Default is `FALSE`.

`characterencoding`
File encoding as used by `locale`. Default is `"iso-8859-1"` to handle special characters.

`remove.hapax` Remove any row or column which has a sum of 1 (i.e., is only attested once), since they do not directly contribute to the result of the seriation. Default is `FALSE`.

Value

An incidence matrix of binary values (0 = row/column occurrence is absence; 1 = row/column occurrence is present).

References

Wickham H, Hester J, Bryan J (2024). *readr: Read Rectangular Text Data*. R package version 2.1.5, https://github.com/tidyverse/readr, `https://readr.tidyverse.org`.

`im.long` *Convert Incidence Matrix to Pairs (Long Format)*

Description

Take an incidence matrix and convert it to a data frame of two columns, where the first column represents the row elements of the incidence matrix and the second column represents the column elements of the incidence matrix. Each row pair represents the incidence (or occurrence) of that row and column element together.

Usage

```
im.long(obj)
```

Arguments

`obj` An incidence matrix.

Value

A data frame of two columns (row and column of the incidence matrix), in which row of the data frame represents a pair of an

Examples

```
data(quattrofontanili)
qf <- im.long(quattrofontanili)

# to export for uploading into the Lakhesis Calculator, use write.table() to
# remove both row and column names:

# write.table(qf, file = 'qf.csv', row.names = FALSE, col.names = FALSE, sep = ",")
```

im.merge *Merge Two Incidence Matrices*

Description

From two incidience matrices, create a single incidence matrix. Matrices may contain same row or column elements.

Usage

```
im.merge(obj1, obj2)
```

Arguments

`obj1, obj2` Two incidence matrices of any size.

Value

A single incidence matrix.

Examples

```
data(quattrofontanili)
qf1 <- quattrofontanili[1:20, 1:40]
qf1 <- qf1[rowSums(qf1) != 0, colSums(qf1) != 0]

qf2 <- quattrofontanili[30:50, 20:60]
qf2 <- qf2[rowSums(qf2) != 0, colSums(qf2) != 0]

im.merge(qf1, qf2)
```

`im.ref` *Create Reference Matrix*

Description

Create an ideal reference matrix of well-seriated values of the same size as the input matrix.

Usage

```
im.ref(obj)
```

Arguments

`obj`	A matrix of size $n \times k$.

Value

A matrix of size $n \times k$ with 1s along the diagonal. If $n > k$, 1s are placed from cell (i, i) to $(i, i + k - n)$, with 0 in all other cells.

Examples

```
im.ref(matrix(NA, 5, 5))
im.ref(matrix(1, 7, 12))
```

`kappa.coef` *Kappa Concentration*

Description

The concentration coefficient κ, which extends the Kendall-Doran measure of concentration to include rows and then weights the total measure by the total sum of values in the matrix. See `concentration.col`.

Usage

```
## S3 method for class 'coef'
kappa(obj)
```

Arguments

`obj`	A seriated binary matrix.

Value

The κ coefficient of concentration.

Examples

```
data("quattrofontanili")
kappa.coef(quattrofontanili)
```

lakhesize *Lakhesize*

Description

This function returns the row and column consensus seriation for a `list` of strands, containing their rankings, the results of their PCA, and coefficients of association and concentration.

Usage

```
lakhesize(strands, obj)
```

Arguments

`strands`	A `list` of strands, which are data frames returned by `ca.procrustes.curve`.
`obj`	The intial incidence matrix.

Details

Consensus seriation is achieved by iterative, multi-step linear regression using simulation. On one iteration, strands are chosen at random, omitting incomplete or missing pairs, using PCA to determine the best-fitting line for their rankings. Both strands' rankings are then regressed onto that line to determine missing values, and then re-ranked, repeating until all strands have been regressed. PCA of the simulated rankings is then used to determine the final sequence of the row and column elements.

Value

A `list` of the following:

- `RowConsensus` Data frame of the consensus seriation of the row elements in the order of their projection on the first principal axis. Contains one column, `Row`.
- `ColConsensus` Data frame of the consensus seriation of the column elements in the order of their project onto the first principal axis. Contains one column, `Column`.
- `RowPCA` The results of `\link[stats]{prcomp}` performed on the row elements of strands.
- `ColPCA` The results of `\link[stats]{prcomp}` performed on the column elements of strands.
- `Coef` A data frame containing the coefficients of agreement and concentration:
  - `Strand` The number of the strand.
  - `Consensus.Spearman.Sq` the measure of agreement, i.e., how well each strand accords with the consensus seriation. Using the square of Spearman's rank correlation coefficient, ρ^2, between each strand and the consensus ranking, agreement is computed as the product of ρ^2 for their row and column rankings, $\rho_r^2 \rho_c^2$.

  - Concentration.Kappa the concentration coefficient κ, which provides a measure of the optimality of each strand (see kappa.coef).

Examples

```
data("quattrofontanili")
data("qfStrands")
lakhesize(qfStrands, quattrofontanili)
```

LC *Lakhesis Calculator*

Description

Launch Lakhesis Calculator, a graphical interface to explore binary matrices via correspondence analysis, select potentially well-seriated sequences, and perform consensus seriation. Interface is made with ggplot2, shiny, shinydashboard, and bslib (Wickham 2016; Chang et al. 2024; Chang and Borges Ribeiro 2021; Sievert et al. 2024).

Usage

```
LC()
```

Details

Input is done in the calculator, via a "long" format a two-column .csv file giving pairs of row and column incidences. See im.csv.read for details. Conversion of a pre-existing incidence matrix to long format can be performed with im.long.

Results can be downloaded from the calculator as an .rds file containing a list of the following:

- results The consensus seriation, PCA, and coefficients of agreement and concentration (lakhesize).
- strands The strands selected by the investigator.
- im.seriated The incidence matrix of the consensus seriation.

Value

Opens the Lakhesis Calculator.

References

Chang W, Borges Ribeiro B (2021). *shinydashboard: Create Dashboards with 'Shiny'*. https://CRAN.R-project.org/package=shinydashboard.

Chang W, Cheng J, Allaire JJ, Sievert C, Schloerke B, Xie Y, Allen J, McPherson J, Dipert A, Borges B (2024). *shiny: Web Application Framework for R*. R package version 1.8.1.9001; https://github.com/rstudio/shiny, https://shiny.posit.co.

Sievert C, Cheng J, Aden-Buie G (2024). *bslib: Custom 'Bootstrap' 'Sass' Themes for 'shiny' and 'rmarkdown'*. R package version 0.7.0, https://github.com/rstudio/bslib, `https://rstudio.github.io/bslib/`.

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York.

`qfStrands` *Quattro Fontanili - Strands*

Description

Three seriated strands selected from `quattrofontanili` data, identified by the package author as an example for the documentation of functions.

Usage

```
data("qfStrands")
```

Format

A list containing data frames output by `ca.procrustes.curve`.

Examples

```
data("qfStrands")
print(qfStrands)
```

`quattrofontanili` *Quattro Fontanili*

Description

The seriation of tombs from necropoleis at Veii, primarily Quattro Fontanili, but also Valle la Fata, Vaccareccia, and Picazzano, in southern Etruria, established by Close-Brooks and Ridgway (1979).

Usage

```
data("quattrofontanili")
```

Format

A seriated incidence matrix of 81 rows (tombs) and 82 columns (types).

Data entered from Close-Brooks and Ridgway (1979), an English translation of the authors' original publication in *Notizie degli Scavi* (1963). Descriptions of types may be found in that paper.

References

Close-Brooks J, Ridgway D (1979). “Veii in the Iron Age.” In Ridgway D, Ridgway FR (eds.), *Italy Before the Romans*, 95–127. Academic Press, London.

Examples

```
data("quattrofontanili")
print(quattrofontanili)
```

spearman.sq *Spearman Correlation Squared*

Description

The square of Spearman’s rank correlation coefficient applied to two rankings (Spearman 1904). Rows with NA values are automatically removed.

Usage

```
spearman.sq(r1, r2)
```

Arguments

r1, r2	Two vectors of paired ranks.

Value

The square of Spearman’s rank correlation coefficient with NA values removed.

References

Spearman C (1904). “The Proof and Measurement of Association between Two Things.” *American Journal of Psychology*, **15**, 72–101. doi:10.2307/1412159.

Examples

```
# e.g., for two partial seriations:
x <- c(1, 2, 3, 4, NA, 5, 6, NA, 7.5, 7.5, 9)
y <- c(23, 17, 19, NA, 21, 22, 25, 26, 27, 36, 32)
spearman.sq(x, y)
```

strand.extract *Strand Extract*

Description

From a `list` of strands produced by `ca.procrustes.curve`, extract two matrices containing the ranks of the rows and columns. The row/column elements are contained in the rows, and the strands are contained in the columns. NA values are entered where a given row/column element is missing from that strand.

Usage

```
strand.extract(strands, obj)
```

Arguments

`strands`	A list of `strands`, which are data frames returned by `ca.procrustes.curve`.
`obj`	The intial incidence matrix.

Value

A list of two matrices:

- `Row` A matrix of the ranks of the row elements.
- `Col` A matrix of the ranks of the column elements.

Examples

```
data("quattrofontanili")
data("qfStrands")
strand.extract(qfStrands, quattrofontanili)
```

strand.suppress *Suppress Element from Strands*

Description

Given a list of strands, remove a row or column element and re-run seriation by correspondence analysis with Procrustes fitting (`ca.procrustes.curve`) to generate a new list of strands that exclude the specified elements. If the resulting strand lacks sufficient points to perform correspondence analysis, that strand is deleted in the output.

Usage

```
strand.suppress(strands, obj, elements)
```

Arguments

`strands`	A list of strands, which are data frames returned by `ca.procrustes.curve`.
`obj`	The intial incidence matrix.
`elements`	A vector of one or more row or column ids to suppress.

Value

A list of the strands.

Examples

```
data("quattrofontanili")
data("qfStrands")
strand.suppress(qfStrands, quattrofontanili, "QF II 15-16")

strand.suppress(qfStrands, quattrofontanili, c("QF II 15-16", "I", "XIV"))
```

Index