Package 'pald'

April 18, 2024

Title Partitioned Local Depth for Community Structure in Data

Version 0.0.4

Description Implementation of the Partitioned Local Depth (PaLD)

approach which provides a measure of local depth and the cohesion of a point to another which (together with a universal threshold for distinguishing strong and weak ties) may be used to reveal local and global structure in data, based on methods described in Berenhaut, Moore, and Melvin (2022) <doi:10.1073/pnas.2003634119>. No extraneous inputs, distributional assumptions, iterative procedures nor optimization criteria are employed. This package includes functions for computing local depths and cohesion as well as flexible functions for plotting community networks and displays of cohesion against distance.

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.2.3

Imports igraph, graphics, glue

Depends R (>= 2.10)

LazyData true

URL https://github.com/LucyMcGowan/pald

BugReports https://github.com/LucyMcGowan/pald/issues

Suggests testthat (>= 3.0.0)

Config/testthat/edition 3

NeedsCompilation no

Author Katherine Moore [aut] (<https://orcid.org/0000-0001-6943-2416>), Kenneth Berenhaut [aut], Lucy D'Agostino McGowan [aut, cre] (<https://orcid.org/0000-0001-7297-9359>)

Maintainer Lucy D'Agostino McGowan <lucydagostino@gmail.com>

Repository CRAN

Date/Publication 2024-04-18 19:12:58 UTC

R topics documented:

| aggregation | 2 3 |
|--------------------|--------|
| as_cohesion_matrix | 3 |
| cognate_dist | 4 |
| cohesion_matrix | 5 |
| cohesion_strong | 6 |
| community_clusters | 7 |
| community_graphs | 8 |
| cultures | 9 |
| dist_cohesion_plot | 9 |
| exdata1 | 10 |
| exdata2 | |
| | 11 |
| local_depths | |
| | |
| noisy_moons | |
| Para | |
| pald_colors | |
| F | |
| strong_threshold | |
| tissue_dist | 18 |
| | 20 |

Index

| aggregation | Aggregation |
|-------------|-------------|
|-------------|-------------|

Description

A synthetic data set of two-dimensional points created by Gionis et al. to demonstrate clustering aggregation.

Usage

aggregation

Format

A data frame with 788 rows and 2 columns, x1 and x2.

References

A. Gionis, H. Mannila, and P. Tsaparas, Clustering aggregation. ACM Transactions on Knowledge Discovery from Data (TKDD), 2007. 1(1): p. 1-30.

any_isolated Any isolated

Description

Checks for isolated points.

Usage

any_isolated(c)

Arguments

с

A cohesion_matrix object, a matrix of cohesion values (see cohesion_matrix).

Value

Logical, indicating whether any points are isolated.

Examples

d <- data.frame(
 x1 = c(1, 2, 3, 6),
 x2 = c(2, 1, 3, 10)
)
D <- dist(d)
C <- cohesion_matrix(D)
any_isolated(C)</pre>

as_cohesion_matrix Coerce a matrix to a cohesion matrix object

Description

as_cohesion_matrix() converts an existing matrix into an object of class cohesion_matrix.

Usage

as_cohesion_matrix(c)

Arguments

С

A matrix of cohesion values (see cohesion_matrix).

Value

Object of class cohesion_matrix

Examples

```
C <- matrix(
    c(0.25, 0.125, 0.125, 0,
    0.125, 0.25, 0, 0.125,
    0.125, 0, 0.25, 0.125,
    0, 0.125, 0.125, 0.25
), nrow = 4, byrow = TRUE)
class(C)
C <- as_cohesion_matrix(C)
class(C)
```

cognate_dist Cognate Data Distance Matrix

Description

A dist object describing distances between 87 Indo-European languages from the perspective of cognates.

Usage

cognate_dist

Format

A dist object for 87 Indo-European languages.

Details

Cognate relationships from a collection of essential words were collected from Dyen et al. and encoded in a 87x2665 binary matrix from which this distance matrix was derived (using Euclidean distance).

References

I. Dyen, J. B. Kruskal, P. Black, An Indoeuropean classification: A lexicostatistical experiment. Trans. Am. Phil. Soc. 82, iii-132 (1992).

4

cohesion_matrix Cohesion Matrix

Description

Creates a matrix of (pairwise) cohesion values from a matrix of pairwise distances or a dist object.

Usage

```
cohesion_matrix(d)
```

Arguments

d

A matrix of pairwise distances or a dist object.

Details

Computes the matrix of (pairwise) cohesion values, C_xw, from a matrix of pairwise distances or a dist object. Cohesion is an interpretable probability that reflects the strength of alignment of a point, w, to another point, x. The rows of the cohesion matrix can be seen as providing neighborhood weights. These values may be used for defining associated weighted graphs (for the purpose of community analysis) as in Berenhaut, Moore, and Melvin (2022).

Given an n x n distance matrix, the sum of the entries in the resulting cohesion matrix is always equal to n/2. Cohesion is partitioned local depth (see local_depths) and thus the row sums of the cohesion matrix provide a measure of local depth centrality.

If you have a matrix that is already a cohesion matrix and you would like to add the class, see as_cohesion_matrix().

Value

The matrix of cohesion values. An object of class cohesion_matrix.

References

K. S. Berenhaut, K. E. Moore, R. L. Melvin, A social perspective on perceived distances reveals deep community structure. Proc. Natl. Acad. Sci., 119(4), 2022.

Examples

```
plot(exdata1)
text(exdata1 + .08, lab = 1:8)
D <- dist(exdata1)
C <- cohesion_matrix(D)
C</pre>
```

neighbor weights (provided by cohesion) for the 8th point in exdata1

```
C[8, ]
localdepths <- rowSums(C)</pre>
```

cohesion_strong Cohesion Matrix: Strong Ties

Description

Provides the symmetrized and thresholded matrix of cohesion values.

Usage

```
cohesion_strong(c, symmetric = TRUE)
```

Arguments

| С | A cohesion_matrix object, a matrix of cohesion values (see cohesion_matrix). |
|-----------|---|
| symmetric | Logical. Whether the returned matrix should be made symmetric (using the minimum): the default is TRUE. |

Details

The threshold is that provided by strong_threshold (and is equal to half of the average of the diagonal of c). Values of the cohesion matrix which are less than the threshold are set to zero. The symmetrization, if desired, is computed using the entry-wise (parallel) minimum of C and its transpose (i.e., min(C_ij, C_ji)). The matrix provided by cohesion_strong (with default symmetric = TRUE) is the adjacency matrix for the graph of strong ties (the cluster graph), see community_graphs and pald.

Value

The symmetrized cohesion matrix in which all entries corresponding to weak ties are set to zero.

Examples

```
C <- cohesion_matrix(dist(exdata2))
strong_threshold(C)
cohesion_strong(C)
## To illustrate the calculation performed
C_strong <- C
## C_strong is equal to cohesion_strong(C, symmetric = FALSE)
C_strong[C < strong_threshold(C)] <- 0
## C_strong_sym is equal to cohesion_strong(C)
C_strong_sym <- pmin(C_strong, t(C_strong))</pre>
```

6

community_clusters

```
## The (cluster) graph whose adjacency matrix, CS,
## is the matrix of strong ties
CS <- cohesion_strong(C)
if (requireNamespace("igraph", quietly = TRUE)) {
G_strong <- igraph::simplify(
    igraph::graph.adjacency(CS, weighted = TRUE, mode = "undirected")
    )
plot(G_strong)
}
```

community_clusters Community clusters

Description

Community clusters

Usage

```
community_clusters(c)
```

Arguments

С

A cohesion_matrix object, a matrix of cohesion values (see cohesion_matrix).

Value

A data frame with two columns:

- point: The points from cohesion matrix c
- community: The community cluster labels

Examples

```
D <- dist(exdata2)
C <- cohesion_matrix(D)
community_clusters(C)
```

Description

Provides the graphs whose edge weights are (mutual) cohesion, together with a graph layout.

Usage

```
community_graphs(c)
```

Arguments

с

A cohesion_matrix object, a matrix of cohesion values (see cohesion_matrix).

Details

Constructs the graphs whose edge weights are (mutual) cohesion (see cohesion_matrix), selfloops are removed. The graph G has adjacency matrix equal to the symmetrized cohesion matrix (using the entry-wise parallel minimum of C and its transpose). The graph G_strong has adjacency matrix equal to the thresholded and symmetrized cohesion matrix (see cohesion_strong). The threshold is equal to half of the average of the diagonal of the cohesion matrix (see strong_threshold).

A layout is also computed using the Fruchterman-Reingold (FR) force-directed graph drawing algorithm. As a result, it may provide a somewhat different layout each time it is run.

Value

A list consisting of:

- G: the weighted (community) graph whose edge weights are mutual cohesion
- G_strong: the weighted (community) graph consisting of edges for which mutual cohesion is greater than the threshold for strong ties (see strong_threshold)
- layout: the layout, using the Fruchterman Reingold (FR) force-directed graph drawing for the graph G

Examples

```
C <- cohesion_matrix(dist(exdata2))
plot(community_graphs(C)$G_strong)
plot(community_graphs(C)$G_strong, layout = community_graphs(C)$layout)</pre>
```

cultures

Description

Pairwise dissimilarities are given by the cultural fixation index obtained from World Values Survey responses.

Usage

cultures

Format

A 59x59 matrix of dissimilarities

References

M. Muthukrishna, et al., Beyond western, educated, industrial, rich, and democratic (WEIRD) psychology: measuring and mapping scales of cultural and psychological distance. Psychol. Sci. 1, 24 (2020).

R. Inglehart et al, World Values Survey: All Rounds-Country-Pooled Datafile 1981-2014, (JD Systems Institute, Madrid 2014).

dist_cohesion_plot Distance Cohesion Plot

Description

Provides a plot of cohesion against distance, with the threshold indicated by a horizontal line.

Usage

```
dist_cohesion_plot(
    d,
    mutual = FALSE,
    xlim_max = NULL,
    cex = 1,
    colors = NULL,
    weak_gray = FALSE
)
```

Arguments

| d | A matrix of pairwise distances or a dist object. |
|-----------|--|
| mutual | Set to TRUE to consider mutual cohesion (i.e., symmetrized using the minimum); the default is FALSE. |
| xlim_max | If desired, set the maximum value of distance which is displayed on the x-axis. |
| cex | Factor by which points should be scaled relative to the default. |
| colors | A vector of color names, if none is given a default is provided. |
| weak_gray | Set to TRUE to display the plot with all weak ties plotted in gray; the default is FALSE. |

Details

The plot of cohesion against distance provides a visualization for the manner in which distance is transformed. The threshold distinguishing strong and weak ties is indicated by a horizontal line. When there are separated regions with different density, one can often observe vertical bands of color, see example below and Berenhaut, Moore, and Melvin (2022). For each distance pair in d, the corresponding value of cohesion is computed. If the pair is within a single cluster, the point is colored (with the same color provided by the pald and plot_community_graphs functions). Weak ties appear below the threshold.

Note that cohesion is not symmetric, and so all n^2 points are plotted. A gray point above the threshold corresponds to a pair in which the value of cohesion is greater than the threshold in only one direction. If one only wants to observe mutual cohesion (i.e., cohesion made symmetric via the minimum), set mutual = TRUE.

Value

A plot of cohesion against distance with threshold indicated by a horizontal line.

Examples

```
D <- dist(exdata2)
dist_cohesion_plot(D)
dist_cohesion_plot(D, mutual = TRUE)
C <- cohesion_matrix(D)
threshold <- strong_threshold(C) #the horizontal line
dist_cohesion_plot(D, mutual = TRUE, weak_gray = TRUE)</pre>
```

exdata1

Example Data 1

Description

A data set consisting of 8 points (in 2-dimensional Euclidean space) to provide a simple illustrative example. This data is displayed in Figure 1 in Berenhaut, Moore, and Melvin (2022).

exdata2

Usage

exdata1

Format

A data frame with 8 rows and 2 columns, x1 and x2

References

K. S. Berenhaut, K. E. Moore, R. L. Melvin, A social perspective on perceived distances reveals deep community structure. Proc. Natl. Acad. Sci., 119(4), 2022.

exdata2

Example Data 2

Description

A data set consisting of 16 points (in 2-dimensional Euclidean space) to provide an illustrative example. This data is displayed in Figure 2 in Berenhaut, Moore, and Melvin (2022).

Usage

exdata2

Format

A data frame with 16 rows and 2 columns, x1 and x2

References

K. S. Berenhaut, K. E. Moore, R. L. Melvin, A social perspective on perceived distances reveals deep community structure. Proc. Natl. Acad. Sci., 119(4), 2022.

exdata3

Example Data 3

Description

A data set consisting of 240 points (in 2-dimensional Euclidean space) to provide an illustrative example. Points were generated from bivariate normal distributions with varying mean and variance (with covariance matrix cI). This data is displayed in Figure 4D in Berenhaut, Moore, and Melvin (2022).

Usage

exdata3

Format

A data frame with 240 rows and 2 columns, x1 and x2

References

K. S. Berenhaut, K. E. Moore, R. L. Melvin, A social perspective on perceived distances reveals deep community structure. Proc. Natl. Acad. Sci., 119(4), 2022.

local_depths

Local (Community) Depths

Description

Creates a vector of local depths from a matrix of distances (or dist object).

Usage

local_depths(d)

Arguments

d

A matrix of pairwise distances, a dist object, or a cohesion_matrix object.

Details

Local depth is an interpretable probability which reflects aspects of relative position and centrality via distance comparisons (i.e., d(z, x) < d(z, y)).

The average of the local depth values is always 1/2. Cohesion is partitioned local depth (see cohesion_matrix); the row-sums of the cohesion matrix are the values of local depth.

Value

A vector of local depths.

Examples

```
D <- dist(exdata1)
local_depths(D)
C <- cohesion_matrix(D)
local_depths(C)</pre>
```

local depths are the row sums of the cohesion matrix
rowSums(C)

cognate distance data

ld_lang <- sort(local_depths(cognate_dist))</pre>

noisy_circles Noisy circles

Description

Noisy circles data generated from scikit-learn

Usage

noisy_circles

Format

A dataframe with 500 rows and 2 columns, x1 and x2.

Source

https://scikit-learn.org/stable/modules/clustering.html#clustering

noisy_moons Noisy moons

Description

Noisy moons data generated from scikit-learn

Usage

noisy_moons

Format

A dataframe with 500 rows and 2 columns, x1 and x2.

Source

https://scikit-learn.org/stable/modules/clustering.html#clustering

pald

Description

A wrapper function which computes the cohesion matrix, local depths, community graphs and provides a plot of the community graphs with connected components of the graph of strong ties colored by connected component.

Usage

```
pald(
    d,
    show_plot = TRUE,
    show_labels = TRUE,
    only_strong = FALSE,
    emph_strong = 2,
    edge_width_factor = 50,
    colors = NULL,
    layout = NULL,
    ...
)
```

Arguments

| d | A matrix of pairwise distances or a dist object. | |
|-------------------|---|--|
| show_plot | Set to TRUE to display plot; the default is TRUE. | |
| show_labels | Set to FALSE to omit vertex labels (to display a subset of labels, use optional parameter vertex.label to modify the label list). Default: TRUE. | |
| only_strong | Set to TRUE if only strong ties, G_strong, should be displayed; the default FALSE will show both strong (colored by connected component) and weak ties (in gray). | |
| emph_strong | Numeric. The numeric factor by which the edge widths of strong ties are emphasized in the display; the default is 2. | |
| edge_width_factor | | |
| | Numeric. Modify to change displayed edge widths. Default: 50. | |
| colors | A vector of display colors, if none is given a default list (of length 24) is pro- vided. | |
| layout | A layout for the graph. If none is specified, FR-graph drawing algorithm is used. | |
| | Optional parameters to pass to the igraph::plot.igraph. function. Some commonly passed arguments include: | |
| | • vertex.label A vector containing label names. If none is given, the row- names of c are used | |
| | • vertex.size A numeric value for vertex size (default = 1) | |
| | • vertex.color.vec A vector of color names for coloring the vertices | |
| | • vertex.label.cex A numeric value for modifying the vertex label size. (default = 1) | |

pald

Details

This function re-computes the cohesion matrix each time it is run. To avoid unnecessary computation when creating visualizations, use the function cohesion_matrix to compute the cohesion matrix which may then be taken as input for local_depths, strong_threshold, cohesion_strong, community_graphs, and plot_community_graphs. For further details regarding each component, see the documentation for each of the above functions.

Value

A list consisting of:

- C: the matrix of cohesion values
- local_depths: a vector of local depths
- clusters: a vector of (community) cluster labels
- · threshold: the threshold above which cohesion is considered particularly strong
- C_strong: the thresholded matrix of cohesion values
- G: the graph whose edges weights are mutual cohesion
- G_strong: the weighted graph whose edges are those for which cohesion is particularly strong
- layout: a FR force-directed layout associated with G

References

K. S. Berenhaut, K. E. Moore, R. L. Melvin, A social perspective on perceived distances reveals deep community structure. Proc. Natl. Acad. Sci., 119(4), 2022.

Examples

```
D <- dist(exdata2)
pald_results <- pald(D)
pald_results$local_depths
pald(D, layout = as.matrix(exdata2), show_labels = FALSE)
C <- cohesion_matrix(D)
local_depths(C)
plot_community_graphs(C, layout = as.matrix(exdata2), show_labels = FALSE)
pald_languages <- pald(cognate_dist)
head(pald_languages$local_depths)</pre>
```

pald_colors

Description

A vector of colors to use if comparing other clustering methods. These are the default colors used in the plotting functions.

Usage

pald_colors

Format

A vector of 24 colors

plot_community_graphs Plot Community Graphs

Description

Provides a plot of the community graphs, with connected components of the graph of strong ties colored by connected component.

Usage

```
plot_community_graphs(
    c,
    show_labels = TRUE,
    only_strong = FALSE,
    emph_strong = 2,
    edge_width_factor = 50,
    colors = NULL,
    ...
)
```

Arguments

| с | A cohesion_matrix object, a matrix of cohesion values (see cohesion_matrix). |
|-------------|---|
| show_labels | Set to FALSE to omit vertex labels (to display a subset of labels, use optional parameter vertex.label to modify the label list). Default: TRUE. |
| only_strong | Set to TRUE if only strong ties, G_strong, should be displayed; the default FALSE will show both strong (colored by connected component) and weak ties (in gray). |
| emph_strong | Numeric. The numeric factor by which the edge widths of strong ties are emphasized in the display; the default is 2. |

edge_width_factor Numeric. Modify to change displayed edge widths. Default: 50. colors A vector of display colors, if none is given a default list (of length 24) is provided. Optional parameters to pass to the igraph::plot.igraph. function. Some . . . commonly passed arguments include: • layout A layout for the graph. If none is specified, FR-graph drawing algorithm is used. • vertex.label A vector containing label names. If none is given, the rownames of c are used • vertex.size A numeric value for vertex size (default = 1) • vertex.color.vec A vector of color names for coloring the vertices • vertex.label.cex A numeric value for modifying the vertex label size. (default = 1)

Details

Plots the community graph, G, with the sub-graph of strong ties emphasized and colored by connected component. If no layout is provided, the Fruchterman-Reingold (FR) graph drawing algorithm is used. Note that the FR graph drawing algorithm may provide a somewhat different layout each time it is run. You can also access and save a given graph layout using community_graphs(C)\$layout. The example below shows how to display only a subset of vertex labels.

Note that the parameter emph_strong is for visualization purposes only and does not influence the network layout.

Value

A plot of the community graphs.

Examples

```
C <- cohesion_matrix(dist(exdata1))
plot_community_graphs(C, emph_strong = 1, layout = as.matrix(exdata1))
plot_community_graphs(C, only_strong = TRUE)
C2 <- cohesion_matrix(cognate_dist)
subset_lang_names <- rownames(C2)
subset_lang_names[sample(1:87, 60)] <- ""
plot_community_graphs(C2, vertex.label = subset_lang_names, vertex.size = 3)</pre>
```

strong_threshold Cohesion Threshold for Strong Ties

Description

Given a cohesion matrix, provides the value of the threshold above which values of cohesion are considered "particularly strong".

Usage

strong_threshold(c)

Arguments

С

A cohesion_matrix object, a matrix of cohesion values (see cohesion_matrix).

Details

The threshold considered in Berenhaut, Moore, and Melvin (2022) which may be used for distinguishing between strong and weak ties. The threshold is equal to half the average of the diagonal of the cohesion matrix, see Berenhaut, Moore, and Melvin (2022).

Value

The value of the threshold.

References

K. S. Berenhaut, K. E. Moore, R. L. Melvin, A social perspective on perceived distances reveals deep community structure. Proc. Natl. Acad. Sci., 119(4), 2022.

Examples

```
C <- cohesion_matrix(dist(exdata1))
strong_threshold(C)
mean(diag(C)) / 2
## points whose cohesion are greater than the threshold may be considered
## (strong) neighbors
which(C[3, ] > strong_threshold(C))
## note that the number of (strongly-cohesive) neighbors varies across the
## space
which(C[4, ] > strong_threshold(C))
```

C[4, c(2, 3, 4, 6)] # cohesion values can provide neighbor weights

tissue_dist Tissue Data Distance Matrix

Description

A dist object describing distances from a subset of tissue gene expression data from the following papers:

- http://www.ncbi.nlm.nih.gov/pubmed/17906632
- http://www.ncbi.nlm.nih.gov/pubmed/21177656
- http://www.ncbi.nlm.nih.gov/pubmed/24271388 obtained from the tissuesGeneExpression bioconductor package.

tissue_dist

Usage

tissue_dist

Format

A dist object of 189 tissue types

Details

The original data frame had 189 rows, each with a corresponding tissue, such as colon, kidney or cerebellum. There were 22,215 columns corresponding to gene expression data from each of these rows. This was then converted into a distance matrix.

References

M. Love and R. Irizarry. tissueGeneExpression. Bioconductor Package

Index

* datasets aggregation, 2 cognate_dist, 4 cultures, 9 exdata1, 10 exdata2, 11 exdata3, 11 noisy_circles, 13 noisy_moons, 13 pald_colors, 16 tissue_dist, 18 aggregation, 2any_isolated, 3 as_cohesion_matrix, 3 as_cohesion_matrix(), 5 $cognate_dist, 4$ cohesion_matrix, 3, 5, 6-8, 12, 15, 16, 18 cohesion_strong, 6, 8, 15 community_clusters, 7 community_graphs, 6, 8, 15 cultures, 9 dist, 4, 5, 10, 12, 14, 18, 19 dist_cohesion_plot, 9 exdata1, 10 exdata2, 11 exdata3, 11 igraph::plot.igraph, 14, 17 local_depths, 5, 12, 15 noisy_circles, 13 noisy_moons, 13 pald, 6, 10, 14 pald_colors, 16 plot_community_graphs, 10, 15, 16

strong_threshold, 8, 15, 17

tissue_dist, 18