Package 'mdendro'

October 13, 2022

Version 2.1.0

Date 2021-09-12

Title Extended Agglomerative Hierarchical Clustering

Description A comprehensive collection of linkage methods for agglomerative hierarchical clustering on a matrix of proximity data (distances or similarities), returning a multifurcated dendrogram or multidendrogram. Multidendrograms can group more than two clusters when ties in proximity data occur, and therefore they do not depend on the order of the input data. Descriptive measures to analyze the resulting dendrogram are additionally provided.

Encoding latin1

URL https://deim.urv.cat/~sergio.gomez/mdendro.php

License AGPL-3

Imports graphics, grDevices, Rcpp, stats, utils

LinkingTo Rcpp

Suggests ape, cluster, dendextend, knitr, rmarkdown

VignetteBuilder knitr, rmarkdown

NeedsCompilation yes

SystemRequirements C++11

Author Alberto Fernández [aut, cre] (<https://orcid.org/0000-0002-1241-1646>), Sergio Gómez [aut] (<https://orcid.org/0000-0003-1820-0062>)

Maintainer Alberto Fernández <alberto.fernandez@urv.cat>

Repository CRAN

Date/Publication 2021-09-12 21:40:07 UTC

R topics documented:

linkage 2
plot.linkage 7

Index **9**

linkage *Extended Agglomerative Hierarchical Clustering*

Description

Agglomerative hierarchical clustering on a dataset of distances or similarities, returning a multifurcated dendrogram or *multidendrogram*. Descriptive measures to analyze the resulting dendrogram are additionally provided.

Usage

```
linkage(prox, type.prox = "distance", digits = NULL,
        method = "arithmetic", par.method = 0, weighted = FALSE,
        group = "variable")

descplot(prox, ..., type.prox = "distance", digits = NULL,
         method = "versatile", par.method = c(-1,0,+1), weighted = FALSE,
         group = "variable", measure = "cor", slope = 10)
```

Arguments

prox	A structure of class "dist" containing non-negative proximity data (distances or similarities).
type.prox	A character string to indicate whether the proximity data represent "distance" (default) or "similarity" between objects.
digits	An integer value specifying the precision, i.e. the number of significant decimal digits to be used for the comparisons between proximity data. This is an important parameter, since equal proximity data at a certain precision may become different by increasing its value. Thus, it may be responsible of the existence of tied proximity data. If the value of this parameter is negative or NULL (default), then the precision is automatically set to that of the input proximity value with the largest number of significant decimal digits.
method	A character string specifying the linkage method to be used. For linkage(), this should be one of: "single", "complete", "arithmetic", "geometric", "harmonic", "versatile", "ward", "centroid" or "flexible". "ward" and "centroid" methods cannot be used with similarity data. "versatile" and "flexible" are the only two methods that can be used in descplot(). See the *Details* section.
par.method	A real value, in the case of linkage(), or a vector of real values, in the case of descplot(), required as parameter for the methods "versatile" and "flexible". The range of possible values is [-Inf, +Inf] for "versatile", and [-1, +1] for "flexible". See the *Details* section.
weighted	A logical value to choose between the weighted and the unweighted (default) versions of some linkage methods. Weighted linkage gives merging branches in a dendrogram equal weight regardless of the number of objects carried on

each branch. Such a procedure weights objects unequally, contrasting with unweighted linkage that gives equal weight to each object in the clusters. This parameter has no effect on the `"single"` and `"complete"` linkages.

`group` A character string to choose a grouping criterion between the `"variable"`-group approach (default) that returns a multifurcated dendrogram (m-ary tree), and the `"pair"`-group approach that returns a bifurcated dendrogram (binary tree). See the *Details* section.

`measure` A character string specifying the descriptive measure to be plotted. This should be one of: `"cor"`, for cophenetic correlation coefficient; `"sdr"`, for space distortion ratio; `"ac"`, for agglomerative coefficient; `"cc"`, for chaining coefficient; or `"tb"`, for tree balance.

`slope` A real value representing the slope of a sigmoid function to map the `"versatile"` linkage unbounded interval `(-Inf, +Inf)` onto the bounded interval `(-1, +1)`. It can be used to improve the distribution of points along the *x* axis.

`...` Graphical parameters (see `par`) may also be supplied and are passed to `plot.default`.

Details

Starting from a matrix of proximity data (distances or similarities), `linkage()` calculates its dendrogram with the most commonly used agglomerative hierarchical clustering methods, i.e. single linkage, complete linkage, arithmetic linkage (also known as average linkage) and Ward's method. Importantly, it contains a new parameterized method named versatile linkage (Fernández and Gómez, 2020), which includes single linkage, complete linkage and average linkage as particular cases, and which naturally defines two new methods, geometric linkage and harmonic linkage.

The difference between the available hierarchical clustering methods rests in the way the proximity between two clusters is defined from the proximity between their constituent objects:

- `"single"`: the proximity between clusters equals the minimum distance or the maximum similarity between objects.
- `"complete"`: the proximity between clusters equals the maximum distance or the minimum similarity between objects.
- `"arithmetic"`: the proximity between clusters equals the arithmetic mean proximity between objects. Also known as average linkage, WPGMA (weighted version) or UPGMA (unweighted version).
- `"geometric"`: the proximity between clusters equals the geometric mean proximity between objects.
- `"harmonic"`: the proximity between clusters equals the harmonic mean proximity between objects.
- `"versatile"`: the proximity between clusters equals the generalized power mean proximity between objects. It depends on the value of `par.method`, with the following linkage methods as particular cases: `"complete"` (`par.method=+Inf`), `"arithmetic"` (`par.method=+1`), `"geometric"` (`par.method=0`), `"harmonic"` (`par.method=-1`) and `"single"` (`par.method=-Inf`).
- `"ward"`: the distance between clusters is a weighted squared Euclidean distance between the centroids of each cluster. This method is available only for distance data.

- "centroid": the distance between clusters equals the square of the Euclidean distance between the centroids of each cluster. Also known as WPGMC (weighted version) or UPGMC (unweighted version). This method is available only for distance data.
- "flexible": the proximity between clusters is a weighted sum of the proximity between clusters in the previous iteration. It depends on the value of par.method, in the range [-1, +1], and it is equivalent to "arithmetic" linkage when par.method=0.

With the argument group, users can choose between a variable-group approach (default) that returns a multifurcated dendrogram or multidendrogram, and a pair-group approach that returns a bifurcated dendrogram. Multidendrograms were introduced (Fernández and Gómez, 2008) to solve the non-uniqueness problem that arises when two or more minimum proximity values between different clusters are equal during the agglomerative process. Multidendrograms group more than two clusters when tied proximity values occur, what produces a uniquely determined solution that does not depend on the order of the input data. When there are no tied proximity values, the variable-group approach gives the same result as the pair-group one.

descplot() can be used with methods "versatile" and "flexible" to analyze graphically the variation of any descriptive measure as a function of the corresponding method parameter.

Value

An object of class "linkage" that describes the multifurcated dendrogram obtained. The object is a list with the following components:

call	The call that produced the result.
digits	Number of significant decimal digits used as precision. It is given by the user or automatically set to that of the input proximity value with the largest number of significant decimal digits.
merger	A list of vectors of integer that describes the merging of clusters at each step of the clustering. If a number j in a vector is negative, then singleton cluster $-j$ was merged at this stage. If j is positive, then the merge was with the cluster formed at stage j of the algorithm.
height	A vector with the proximity values between merging clusters (for the particular agglomeration) at the successive stages.
range	A vector with the range (the maximum minus the minimum) of proximity values between merging clusters. It is equal to 0 for binary clusters.
order	A vector giving a permutation of the original observations to allow for plotting, in the sense that the branches of a clustering tree will not cross.
coph	Object of class "dist" containing the cophenetic (or ultrametric) proximity data in the output dendrogram, sorted in the same order as the input proximity data in prox.
binary	A logical value indicating whether the output dendrogram is a binary tree or, on the contrary, it contains an agglomeration of more than two clusters due to the existence of tied proximity data. Its value is always TRUE when the "pair" grouping criterion is used.
cor	Cophenetic correlation coefficient (Sokal and Rohlf, 1962), defined as the Pearson correlation coefficient between the output cophenetic proximity data and the input proximity data. It is a measure of how faithfully the dendrogram preserves the pairwise proximity between objects.

sdr	Space distortion ratio (Fernández and Gómez, 2020), calculated as the difference between the maximum and minimum cophenetic proximity data, divided by the difference between the maximum and minimum initial proximity data. Space dilation occurs when the space distortion ratio is greater than 1.
ac	Agglomerative coefficient (Rousseeuw, 1986), a number between 0 and 1 measuring the strength of the clustering structure obtained.
cc	Chaining coefficient (Williams *et al.*, 1966), a number between 0 and 1 measuring the tendency for clusters to grow by the addition of clusters much smaller rather than by fusion with other clusters of comparable size.
tb	Tree balance (Fernández and Gómez, 2020), a number between 0 and 1 measuring the equality in the number of leaves in the branches concerned at each fusion in the hierarchical tree.

Class `"linkage"` has methods for the following generic functions: `summary`, `plot` (see `plot.linkage`), `as.dendrogram`, `as.hclust` and `cophenetic`.

Note

Except for the cases containing tied proximity data, the following equivalences hold between function `linkage()` in package *mdendro*, function `hclust()` in package *stats*, and function `agnes()` in package *cluster*. When relevant, weighted (`W`) or unweighted (`U`) versions of the linkage methods and the value for `par.method` (β) are indicated:

linkage()	hclust()	agnes()
"single"	"single"	"single"
"complete"	"complete"	"complete"
"arithmetic", U	"average"	"average"
"arithmetic", W	"mcquitty"	"weighted"
"ward"	"ward.D2"	"ward"
"centroid", U	"centroid"	--------
"centroid", W	"median"	--------
"flexible", U, β	--------	"gaverage", β
"flexible", W, β	--------	"flexible", $(1-\beta)/2$

Author(s)

Alberto Fernández <alberto.fernandez@urv.cat> and Sergio Gómez <sergio.gomez@urv.cat>.

References

Fernández, A.; Gómez, S. (2008). Solving non-uniqueness in agglomerative hierarchical clustering using multidendrograms. *Journal of Classification*, **25**, 43–65.

Fernández, A.; Gómez, S. (2020). Versatile linkage: a family of space-conserving strategies for agglomerative hierarchical clustering. *Journal of Classification*, **37**, 584–597.

Rousseeuw, P.J. (1986). A visual display for hierarchical classification. In E. Diday *et al.* (eds.) *Data Analysis and Informatics 4*, pp. 743–748. Amsterdam: North-Holland.

Sokal, R.R.; Rohlf, F.J. (1962). The comparison of dendrograms by objective methods. *Taxon*, **11**, 33–40.

Williams, W.T.; Lambert, J.M.; Lance, G.N. (1966). Multivariate methods in plant ecology: V. Similarity analyses and information-analysis. *Journal of Ecology*, **54**, 427–445.

See Also

plot.linkage, dist, dendrogram, hclust, agnes.

Examples

```
## Plot and summary of unweighted arithmetic linkage (UPGMA) dendrogram
lnk1 <- linkage(UScitiesD)
plot(lnk1)
summary(lnk1)

## Linkage of similarity data (non-negative correlations)
sim <- as.dist(cor(EuStockMarkets))
lnk2 <- linkage(sim, type.prox = "similarity")
plot(lnk2)

## Use function as.dendrogram to plot with package dendextend
d <- dist(scale(mtcars))  # distances of standardized data
lnk <- linkage(d, digits = 1, method = "complete")
lnk.dend <- as.dendrogram(lnk)
plot(dendextend::set(lnk.dend, "branches_k_color", k = 4),
     nodePar = list(cex = 0.4, lab.cex = 0.5))

## Plot heatmap containing multidendrograms
heatmap(scale(mtcars), hclustfun = linkage)

## Plot of different versatile linkages as we increase the method parameter
d = as.dist(matrix(c( 0,  7, 16, 12,
                      7,  0,  9, 19,
                     16,  9,  0, 12,
                     12, 19, 12, 0), nrow = 4))
par(mfrow = c(2, 3))
vals <- c(-Inf, -1, 0, +1, +Inf)
names <- c("single", "harmonic", "geometric", "arithmetic", "complete")
for (i in 1:length(vals)) {
  lnk <- linkage(d, digits = 1, method = "versatile", par.method = vals[i])
  plot(lnk, main = paste0("versatile (", vals[i], ") = ", names[i]),
       ylim = c(0, 20), cex = 0.6)
}

## Analyze how descriptive measures depend on versatile linkage parameter
par(mfrow = c(2, 3))
measures <- c("cor", "sdr", "ac", "cc", "tb")
vals <- c(-Inf, (-20:+20), +Inf)
for (measure in measures) {
  descplot(UScitiesD, method = "versatile", par.method = vals,
           measure = measure,  main = measure, type = "o", col = "blue")
```

```
}
```

---

plot.linkage *Plots for Extended Agglomerative Hierarchical Clustering*

Description

Creates plots for visualizing an object of class "linkage".

Usage

```
## S3 method for class 'linkage'
plot(x, type = c("rectangle", "triangle"),
     center = FALSE, edge.root = FALSE,
     nodePar = NULL, edgePar = list(),
     leaflab = c("perpendicular", "textlike", "none"),
     dLeaf = NULL, xlab = "", ylab = "", xaxt = "n", yaxt = "s",
     horiz = FALSE, frame.plot = FALSE, xlim, ylim,
     col.rng = "lightgray", ...)
```

Arguments

x	An object of class "linkage", typically created by linkage().
type	Type of plot.
center	Logical; if TRUE, nodes are plotted centered with respect to the leaves in the branch. Otherwise (default), plot them in the middle of all direct child nodes.
edge.root	Logical; if true, draw an edge to the root node.
nodePar	A list of plotting parameters to use for the nodes (see points) or NULL by default which does not draw symbols at the nodes. The list may contain components named pch, cex, col, xpd, and/or bg each of which can have length two for specifying separate attributes for inner nodes and leaves. Note that the default of pch is 1:2, so you may want to use pch = NA if you specify nodePar.
edgePar	A list of plotting parameters to use for the edge segments. The list may contain components named col, lty and lwd. As with nodePar, each can have length two for differentiating leaves and inner nodes.
leaflab	A string specifying how leaves are labeled. The default "perpendicular" writes text vertically (by default), "textlike" writes text horizontally (in a rectangle), and "none" suppresses leaf labels.
dLeaf	A number specifying the distance in user coordinates between the tip of a leaf and its label. If NULL as per default, 3/4 of a letter width or height is used.
xlab, ylab	A label for the axis.
xaxt, yaxt	A character which specifies the axis type. Specifying "n" suppresses plotting, while any value other than "n" implies plotting.
horiz	Logical indicating if the dendrogram should be drawn horizontally or not.

frame.plot	Logical indicating if a box around the plot should be drawn, see plot.default.
xlim, ylim	Optional x- and y-limits of the plot, passed to plot.default. The defaults for these show the full dendrogram.
col.rng	Color ("lightgray" by default) to be used for plotting range rectangles in case of tied heights. If NULL, range rectangles are not plotted.
...	Graphical parameters (see par) may also be supplied and are passed to plot.default.

Details

Based on the plot function for objects of class "dendrogram" (see plot.dendrogram), the plot function for objects of class "linkage" adds the possibility of visualizing the existence of tied heights in a dendrogram thanks to the col.rng parameter.

See Also

linkage, dendrogram.

Examples

```
## Plot complete linkage of mtcars distances, showing and hiding ranges
mtcars.dist <- dist(scale(mtcars))  # distances of standardized data
lnk <- linkage(mtcars.dist, digits = 1, method = "complete")
par(mfrow = c(1, 2))
nodePar <- list(cex = 0, lab.cex = 0.4)
plot(lnk, col.rng = "bisque", main = "show ranges", nodePar = nodePar)
plot(lnk, col.rng = NULL, main = "hide ranges", nodePar = nodePar)
```

Index

∗ **agglomerative coefficient**
    linkage, 2
∗ **chaining coefficient**
    linkage, 2
∗ **cluster**
    linkage, 2
    plot.linkage, 7
∗ **cophenetic correlation coefficient**
    linkage, 2
∗ **hplot**
    plot.linkage, 7
∗ **space distortion ratio**
    linkage, 2
∗ **tree balance**
    linkage, 2

agnes, *5, 6*
as.dendrogram, *5*
as.hclust, *5*

cophenetic, *5*

dendrogram, *6*, *8*
descplot *(*linkage*)*, 2
dist, *6*

hclust, *5, 6*

linkage, 2, *7, 8*

par, *3*, *8*
plot.default, *3*, *8*
plot.dendrogram, *8*
plot.linkage, *5, 6*, 7
points, *7*

segments, *7*
summary, *5*