

Package ‘LLAhclust’

February 14, 2012

Version 0.2-2

Date 2007-31-08

Title Hierarchical clustering of variables or objects based on the likelihood linkage analysis method

Author Ivan Kojadinovic, Israël-César Lerman, Philippe Peter

Maintainer Ivan Kojadinovic <ivan@stat.auckland.ac.nz>

Description The likelihood linkage analysis is a general agglomerative hierarchical clustering method developed in France by Lerman in a long series of research articles and books. Initially proposed in the framework of variable clustering, it has been progressively extended to allow the clustering of very general object descriptions. The approach mainly consists in replacing the value of the estimated similarity coefficient by the probability of finding a lower value under the hypothesis of ‘absence of link’. The package LLAhclust contains routines for computing various types of probabilistic similarity coefficients between variables or object descriptions. Once the similarity values between variables/objects are computed, a hierarchical clustering can be performed using several probabilistic and non-probabilistic aggregation criteria, and indices measuring the quality of the partitions compatible with the resulting hierarchy can be computed.

Depends R(>= 2.1.0)

Encoding latin1

License CeCILL-2

URL <http://www.stat.auckland.ac.nz/~ivan/LLAhclust>

Repository CRAN

Date/Publication 2009-02-27 07:40:31

R topics documented:

as.LLASim	2
as.matrix.LLASim	4
empcopula.simulate	5
LLAhclust	6
LLAparteval	8
LLAsimobj	10
LLAsimvar	11

Index	14
--------------	-----------

as.LLASim	<i>Converts the lower triangle of a square matrix into a LLASim object</i>
-----------	----------------------------------------------------------------------------

Description

Converts the lower triangle of a square matrix into a LLASim object. The LLASim object contains similarity coefficients among objects or variables of interest.

Usage

```
as.LLASim(m, upper = FALSE, probabilistic = FALSE)
```

Arguments

<code>m</code>	input square similarity matrix.
<code>upper</code>	logical value indicating whether the upper triangle of the similarity matrix should be printed by <code>print.LLASim</code> .
<code>probabilistic</code>	logical value indicating whether the coefficients in the input similarity matrix should be treated as probabilistic similarity values. If set to <code>FALSE</code> , the input similarity coefficients are scaled. See examples below.

Details

The following functions are also defined for objects of class LLASim: `names.LLASim`, `format.LLASim`, `as.matrix.LLASim` and `print.LLASim`.

Value

Returns an object of class LLASim whose attributes are very similar to those of objects of class `dist`. See [dist](#) for more details.

References

I.C. Lerman (1981), *Classification et analyse ordinaire de données*, Dunod, Paris.

I.C. Lerman (1991), *Foundations of the likelihood linkage analysis classification method*, Applied Stochastic Models and Data Analysis, 7, pages 63–76.

I.C. Lerman (1993), *Likelihood linkage analysis classification method: An example treated by hand*, Biochimie, 75, pages 379–397.

I.C. Lerman, Ph. Peter and H. Leredde (1993), *Principes et calculs de la méthode implantée dans le programme CHAVL (Classification Hiérarchique par Analyse de la Vraisemblance des Liens)*, Modulad, 12, pages 33-101.

See Also

[LLAsimvar](#),
[LLAsimobj](#),
[as.matrix.LLASim](#),
[dist](#).

Examples

```
## Assume that we have at hand a probabilistic similarity matrix
## between 5 objects (lower triangle only):
m <- matrix(runif(25), 5, 5)

## The corresponding LLAsim object is obtained as follows:
s <- as.LLASim(m, probabilistic=TRUE)

## Display the initial matrix and the LLAsim object:
m
s

## Assume now that we have at hand a non-probabilistic similiarity
## matrix:
m <- matrix(rnorm(25), 5, 5)

## The corresponding LLAsim object is obtained as follows:
s <- as.LLASim(m, probabilistic=FALSE)

## Display the initial matrix and the LLAsim object:
m
s
## Notice that the coefficients in s are scaled:
mean(s)
sd(s)
```

as.matrix.LLAsim *Useful functions for dealing with LLAsim objects*

Description

The function `as.matrix.LLAsim` converts a `LLAsim` object into a square symmetrical matrix. The usual R functions `format`, `print` and `names` have also been extended to deal with `LLAsim` objects.

Usage

```
as.matrix.LLAsim(x, ...)
```

Arguments

`x` the `LLAsim` object to be converted.
`...` nothing so far.

Value

An object of class `matrix`.

References

- I.C. Lerman (1981), *Classification et analyse ordinaire de données*, Dunod, Paris.
- I.C. Lerman (1991), *Foundations of the likelihood linkage analysis classification method*, Applied Stochastic Models and Data Analysis, 7, pages 63–76.
- I.C. Lerman (1993), *Likelihood linkage analysis classification method: An example treated by hand*, Biochimie, 75, pages 379–397.
- I.C. Lerman, Ph. Peter and H. Leredde (1993), *Principes et calculs de la méthode implantée dans le programme CHAVL (Classification Hiérarchique par Analyse de la Vraisemblance des Liens)*, Modulad, 12, pages 33-101.

See Also

[LLAsimvar](#),
[LLAsimobj](#),
[as.LLAsim](#).

Examples

```
data(USArrests)

## Compute similarities between objects based on
## a local Euclidean distance (see references above):
s <- LLAsimobj(USArrests)

## Convert to a matrix object:
```

```
as.matrix(s)

## Other useful functions:
print(s, upper=TRUE)
names(s)
## For the format function, see the R help.
```

empcopula.simulate	<i>Simulation step used in the independence test based on the empirical copula process implemented in the LLAsimvar function</i>
--------------------	----------------------------------------------------------------------------------------------------------------------------------

Description

Simulation step used in the independence test based on the empirical copula process as proposed by Christian Genest and Bruno Rémillard. To be used in conjunction with the LLAsimvar function (method="empirical.copula"). The simulation step consists in simulating the distribution of the test statistic under independence for the sample size under consideration. More details can be found in the articles cited in the reference section.

Usage

```
empcopula.simulate(n, N = 2000)
```

Arguments

n	Sample size when simulating the distribution of the test statistic under independence.
N	Number of repetitions when simulating under independence.

Details

See the references below for more details, especially the third one.

Value

The function empcopula.simulate returns an object of class empcop.simulation whose attributes are: sample.size, number.repetitions and dist.independence (a vector of length N containing the values of the test statistic for each each repetition).

References

- P. Deheuvels (1979), La fonction de dépendance empirique et ses propriétés: un test non paramétrique d'indépendance, Acad. Roy. Belg. Bull. Cl. Sci. 5th Ser. 65, 274-292.
- P. Deheuvels (1981), A non parametric test for independence, Publ. Inst. Statist. Univ. Paris 26, 29-50.
- C. Genest and B. Rémillard (2004). Tests of independence and randomness based on the empirical copula process. Test, 13, 335-369.

C. Genest, J.-F. Quessy and B. Rémillard (2006). *Local efficiency of a Cramer-von Mises test of independence*. Journal of Multivariate Analysis, 97, 274-294.

C. Genest, J.-F. Quessy and B. Rémillard (2007). *Asymptotic local efficiency of Cramer-von Mises tests for multivariate independence*. The Annals of Statistics, 35, in press.

I. Kojadinovic (2007), *Hierarchical clustering of continuous variables based on the empirical copula process*, submitted.

See Also

[LLAsimvar](#),
[LLAhclust](#).

Examples

```
data(USArrests)

## Compute similarities between variables using the test of
## independence a la Deheuvels based on the empirical copula
## process recently studied by Genest and Remillard:
s <- LLAsimvar(USArrests, method = "empirical.copula")
s

## The previous computation could have been done in two steps:
d <- empcopula.simulate(n=50,N=2000)
s <- LLAsimvar(USArrests, method = "empirical.copula",
               simulated.distribution = d)
s
```

LLAhclust

Likelihood linkage analysis hierichal clustering

Description

Builds a hierarchy from similarity coefficients among objects or variables as returned by `LLAsimvar`, `LLAsimobj` or `as.LLASim`. The default aggregation criteria, called `lla`, can be regarded as a probabilistic version of the single linkage.

Usage

```
LLAhclust(s, method = "lla", epsilon = 1, members = NULL)
```

Arguments

`s` Similarity coefficients as returned by `LLAsimvar`, `LLAsimobj` or `as.LLASim`.

method	Linkage method (i.e. aggregation criterion). Can be one of <code>11a</code> (default), <code>tippett</code> (Tippett's p-value combination method), <code>average</code> , <code>complete</code> , <code>fisher</code> (Fisher's p-value combination method), <code>uniform</code> (uniform p-value combination method; can be regarded as a probabilistic version of the average linkage), <code>normal</code> (normal p-value combination method) or <code>maximum</code> (maximum p-value combination method; can be regarded as a probabilistic version of the complete linkage). See the last reference for more details.
epsilon	Coefficient used in the <code>11a</code> linkage. Should lie in $[0,1]$: <code>epsilon=0</code> corresponds to the single linkage, <code>epsilon=1</code> (default) yields a probabilistic version of the single linkage.
members	"Weights" of the objects to be clustered if not of equal "weight". See <code>hclust</code> for more details.

Value

An object of class `hclust` with the corresponding attributes. See `hclust` for more details.

References

- I.C. Lerman (1981), *Classification et analyse ordinale de donnés*, Dunod, Paris.
- I.C. Lerman (1991), *Foundations of the likelihood linkage analysis classification method*, Applied Stochastic Models and Data Analysis, 7, pages 63–76.
- I.C. Lerman (1993), *Likelihood linkage analysis classification method: An example treated by hand*, Biochimie, 75, pages 379–397.
- I.C. Lerman, Ph. Peter and H. Leredde (1993), *Principes et calculs de la méthode implantée dans le programme CHAVL (Classification Hiérarchique par Analyse de la Vraisemblance des Liens)*, Modulad, 12, pages 33-101.
- I. Kojadinovic (2007), *Hierarchical clustering of continuous variables based on the empirical copula process*, submitted.

See Also

`LLAsimvar`,
`LLAsimobj`,
`as.LLASim`,
`LLAparteval`,
`hclust`.

Examples

```
data(USArrests)

## Compute similarities between variables based on
## the LLAnumerical method:
s <- LLAsimvar(USArrests)
s

## Perform the hierarchical clustering of the variables
```

```

## using the default aggregation criterion (lla):
h <- LLAhclust(s)
plot(h)

## Compute the quality of the partitions compatible
## with the hierarchy in terms of the statistics defined by Lerman:
LLAparteval(h,s)

## Compute similarities between variables using the classical
## bilateral test of independence based on Spearman's rho:
s <- LLAsimvar(USArrests, method = "spearman.abs")
s

## Perform the hierarchical clustering of the variables
## using Fisher's p-value combination method:
h <- LLAhclust(s,method="fisher")
plot(h)
## NB: the height in the dendrogram is a p-value
## and can be used to identify mutually independent classes of
## variables, if any.

## Compute the quality of the partitions compatible
## with the hierarchy in terms of the indices defined in the
## last reference:
LLAparteval(h,s)

```

LLAparteval

Evaluates the quality of each partition compatible with a hierarchy in terms of several indices

Description

Evaluates the quality of each partition compatible with the hierarchy returned by LLAhclust . If the hierarchy is obtained from similarity coefficients computed using LLA* methods, the global and local statistics proposed by Lerman are calculated. Otherwise, for similarity coefficients obtained from independence tests (see LLAsimvar), for each partition, the inter-class p-values are combined using Tippett's and Fisher's rules. Furthermore, the minimum inter-class p-value and the maximum intra-class p-value are given. See the last reference and the examples below for more details.

Usage

```
LLAparteval(tree, s, m=NULL)
```

Arguments

tree	An object of class hclust as returned by LLAhclust.
s	An object of class LLAsim as returned by LLAsimvar, LLAsimobj or as.LLAsim.
m	Integer. If set, the quality of the m coarsest partitions only is evaluated.

Value

Returns a data.frame whose columns are: `global.stat` and `local.stat` if the hierarchy is obtained from similarity coefficients computed using LLA* methods, and `tippett.inter`, `fisher.inter`, `min.inter` and `max.intra` in case of similarity coefficients obtained from independence tests.

References

- I.C. Lerman (1981), *Classification et analyse ordinale de donnés*, Dunod, Paris.
- I.C. Lerman (1991), *Foundations of the likelihood linkage analysis classification method*, Applied Stochastic Models and Data Analysis, 7, pages 63–76.
- I.C. Lerman (1993), *Likelihood linkage analysis classification method: An example treated by hand*, Biochimie, 75, pages 379–397.
- I.C. Lerman, Ph. Peter and H. Leredde (1993), *Principes et calculs de la méthode implantée dans le programme CHAVL (Classification Hiérarchique par Analyse de la Vraisemblance des Liens)*, Modulad, 12, pages 33-101.
- I. Kojadinovic (2007), *Hierarchical clustering of continuous variables based on the empirical copula process*, submitted.

See Also

[LLAsimvar](#),
[LLAsimobj](#),
[as.LLASim](#),
[LLAhclust](#).

Examples

```
data(USArrests)

## Compute similarities between variables based on
## the LLAnumerical method:
s <- LLAsimvar(USArrests)
s

## Perform the hierarchical clustering of the variables:
h <- LLAhclust(s)
plot(h)

## Compute the quality of the partitions compatible
## with the hierarchy in terms of the statistics defined by Lerman:
LLAparteval(h,s)

## Compute similarities between variables using the classical
## bilateral test of independence based on Spearman's rho:
s <- LLAsimvar(USArrests, method = "spearman.abs")
s

## Perform the hierarchical clustering of the variables
```

```
## using Fisher's p-value combination method:
h <- LLAhclust(s,method="fisher")
plot(h)
## NB: the height in the dendrogram is a p-value
## and can be used to identify mutually independent classes of
## variables, if any.

## Compute the quality of the partitions compatible
## with the hierarchy in terms of the indices defined in the
## last reference:
LLAparteval(h,s)
```

LLAsimobj

Computes similarities among objects

Description

Computes similarities among objects using the likelihood linkage analysis approach proposed by Lerman. The likelihood linkage analysis method mainly consists in replacing the value of the similarity coefficient between two objects by the probability of finding a lower value under the hypothesis of *absence of link*. See the references below for more details.

Usage

```
LLAsimobj(x, method = "LLAeuclidean", upper = FALSE)
```

Arguments

x	a numeric matrix or data frame.
method	Can be one of LLAeuclidean, LLAcosinus, LLAcategorical, LLAordinal, or LLAboolean. The two first methods can be used to compute similarity coefficients between objects described by numerical variables.
upper	logical value indicating whether the upper triangle of the similarity matrix should be printed by <code>print.LLASim</code> .

Details

The following functions are also defined for objects of class `LLAsim`: `names.LLASim`, `format.LLASim`, `as.matrix.LLASim` and `print.LLASim`.

Value

Returns an object of class `LLAsim` whose attributes are very similar to those of objects of class `dist`. See [dist](#) for more details.

References

- I.C. Lerman (1981), *Classification et analyse ordinaire de données*, Dunod, Paris.
- I.C. Lerman (1991), *Foundations of the likelihood linkage analysis classification method*, Applied Stochastic Models and Data Analysis, 7, pages 63–76.
- I.C. Lerman (1993), *Likelihood linkage analysis classification method: An example treated by hand*, Biochimie, 75, pages 379–397.
- I.C. Lerman, Ph. Peter and H. Leredde (1993), *Principes et calculs de la méthode implantée dans le programme CHAVL (Classification Hiérarchique par Analyse de la Vraisemblance des Liens)*, Modulad, 12, pages 33-101.

See Also

[LLAsimvar](#),
[as.LLAsim](#),
[LLAhclust](#),
[LLAparteval](#),
[dist](#).

Examples

```
data(USArrests)

## Compute similarities between objects based on
## a local Euclidean distance (see references above):
s <- LLAsimobj(USArrests)
s
```

LLAsimvar	<i>Computes similarities among variables using the likelihood linkage analysis approach</i>
-----------	---------------------------------------------------------------------------------------------

Description

Computes similarities among variables using the likelihood linkage analysis approach proposed by Lerman. The likelihood linkage analysis method mainly consists in replacing the value of the estimated similarity coefficient between two variables by the probability of finding a lower value under the hypothesis of stochastic independence, called *absence of link* in that context. Nine similarity coefficients can be computed using the LLAsimvar function.

Usage

```
LLAsimvar(x, method = "LLAnumerical", upper = FALSE,  
          simulated.distribution = NULL)
```

Arguments

<code>x</code>	a numeric matrix or data frame.
<code>method</code>	Can be one of <code>LLAnumerical</code> , <code>LLAcategorical</code> , <code>LLAordinal</code> , <code>LLAboolean</code> , <code>chi.square</code> , <code>pearson.abs</code> , <code>spearman.abs</code> , <code>kendall.abs</code> or <code>empirical.copula</code> . The methods <code>LLA*</code> were initially defined by Lerman (see references below). The four remaining methods compute the similarity between two variables as one minus the p-value obtained from a test of independence. See the last reference and the example section below for more details.
<code>upper</code>	logical value indicating whether the upper triangle of the similarity matrix should be printed by <code>print.LLASim</code> .
<code>simulated.distribution</code>	Object of class <code>empcopula.simulation</code> . Should be set only if the method <code>empirical.copula</code> is selected. See function empcopula.simulate and the example section below for more details.

Details

The following functions are also defined for objects of class `LLAsim`: `names.LLASim`, `format.LLASim`, `as.matrix.LLASim` and `print.LLASim`.

Value

Returns an object of class `LLAsim` whose attributes are very similar to those of objects of class `dist`. See [dist](#) for more details.

References

- I.C. Lerman (1981), *Classification et analyse ordinaire de donnés*, Dunod, Paris.
- I.C. Lerman (1991), *Foundations of the likelihood linkage analysis classification method*, Applied Stochastic Models and Data Analysis, 7, pages 63–76.
- I.C. Lerman (1993), *Likelihood linkage analysis classification method: An example treated by hand*, Biochimie, 75, pages 379–397.
- I.C. Lerman, Ph. Peter and H. Leredde (1993), *Principes et calculs de la méthode implantée dans le programme CHAVL (Classification Hiérarchique par Analyse de la Vraisemblance des Liens)*, Modulad, 12, pages 33-101.
- P. Deheuvels (1979), La fonction de dépendance empirique et ses propriétés: un test non paramétrique d'indépendance, Acad. Roy. Belg. Bull. Cl. Sci. 5th Ser. 65, 274-292.
- C. Genest and B. Rémillard (2004). *Tests of independence and randomness based on the empirical copula process*. Test, 13, 335-369.
- I. Kojadinovic (2007), *Hierarchical clustering of continuous variables based on the empirical copula process*, submitted.

See Also

[as.LLASim](#),
[empcopula.simulate](#),
[LLAsimobj](#),
[LLAhclust](#),
[LLAparteval](#),
[dist](#).

Examples

```
data(USArrests)

## Compute similarities between variables using the
## LLAnumerical method:
s <- LLAsimvar(USArrests)
s

## Compute similarities between variables using the classical
## bilateral test of independence based on Spearman's rho:
s <- LLAsimvar(USArrests, method = "spearman.abs")
s

## Compute similarities between variables using the classical
## bilateral test of independence based on Kendall's tau:
s <- LLAsimvar(USArrests, method = "kendall.abs")
s

## Compute similarities between variables using the test of
## independence e la Deheuvels based on the empirical copula
## process recently studied by Genest and Remillard:
s <- LLAsimvar(USArrests, method = "empirical.copula")
s

## The previous computation could have been done in two steps:
d <- empcopula.simulate(n=50,N=2000)
s <- LLAsimvar(USArrests, method = "empirical.copula",
              simulated.distribution = d)
s
```

Index

*Topic **cluster**

- as.LLASim, [2](#)
- as.matrix.LLASim, [4](#)
- empcopula.simulate, [5](#)
- LLAhclust, [6](#)
- LLAparteval, [8](#)
- LLAsimobj, [10](#)
- LLAsimvar, [11](#)

as.LLASim, [2](#), [4](#), [7](#), [9](#), [11](#), [13](#)

as.matrix.LLASim, [3](#), [4](#)

dist, [2](#), [3](#), [10–13](#)

empcopula.simulate, [5](#), [12](#), [13](#)

format.LLASim(as.matrix.LLASim), [4](#)

hclust, [7](#)

LLAhclust, [6](#), [6](#), [9](#), [11](#), [13](#)

LLAparteval, [7](#), [8](#), [11](#), [13](#)

LLAsimobj, [3](#), [4](#), [7](#), [9](#), [10](#), [13](#)

LLAsimvar, [3](#), [4](#), [6–9](#), [11](#), [11](#)

names.LLASim(as.matrix.LLASim), [4](#)

print.LLASim(as.matrix.LLASim), [4](#)