Package 'caviarpd'

August 13, 2021

Type Package

Title Cluster Analysis via Random Partition Distributions

Version 0.2.17

Description Cluster analysis is performed using pairwise distance information and a random partition distribution. The method is
implemented for two random partition distributions. It draws samples and then obtains and plots clustering estimates.
An implementation of a selection algorithm is provided for the mass parameter of the partition distribution. Since
pairwise distances are the principal input to this procedure, it is most comparable to the hierarchical and k-medoids
clustering methods. The method is currently under peer review at a journal.

License MIT + file LICENSE | Apache License 2.0

Depends R (>= 4.0.0), salso (>= 0.2.20)

Imports cluster (>= 2.1.2)

SystemRequirements Cargo (>= 1.51) for installation from sources: see INSTALL file

Encoding UTF-8

RoxygenNote 7.1.1

NeedsCompilation yes

Author David B. Dahl [aut, cre] (<https://orcid.org/0000-0002-8173-1547>),
Jacob Andros [aut] (<https://orcid.org/0000-0002-1289-385X>),
J. Brandon Carter [aut] (<https://orcid.org/0000-0003-1687-0564>)

Maintainer David B. Dahl <dahl@stat.byu.edu>

Repository CRAN

Date/Publication 2021-08-13 04:30:05 UTC

R topics documented:

`caviarpd` *Cluster Analysis via Random Partition Distributions*

Description

Returns a clustering estimate given pairwise distances using the CaviarPD method.

Usage

```
caviarpd(
  distance,
  nClusters,
  mass,
  nSamples = 1000,
  gridLength = 10,
  samplesOnly = FALSE,
  loss = "binder",
  distr = "EPA",
  temperature = 10,
  similarity = c("exponential", "reciprocal")[1],
  discount = 0,
  sd = 3,
  maxNClusters = 0,
  nCores = 0
)
```

Arguments

`distance`	An object of class 'dist' or a pairwise distance matrix.
`nClusters`	A numeric vector that specifies the range for the number of clusters to consider in the search for a clustering estimate. Should be missing if the `mass` argument is used. See `Details`.
`mass`	A numeric vector of mass values to consider in the search for a clustering estimate. Should be missing if the `nClusters` argument is used. See `Details`.
`nSamples`	The number of samples used to generate the clustering estimate.
`gridLength`	The length of the grid search for an optimal mass parameter. Only applicable if a range of values are provided for `nClusters`.
`samplesOnly`	If TRUE, the function only returns the samples generated for a given mass, temperature, and discount rather than an actual clustering estimate.
`loss`	The SALSO method (Dahl, Johnson, Müller, 2021) tries to minimize this expected loss when searching the partition space for an optimal estimate. This must be either "binder" or "VI".
`distr`	The random partition distribution used to generate samples. This must be specified as either "EPA" or "ddCRP".
`temperature`	A positive number that accentuates or dampens distance between observations.

similarity	Either "exponential" or "reciprocal" to indicate the desired similarity function.
discount	A number in $[0, 1)$ giving the discount parameter to control the distribution of subset sizes.
sd	Number of standard deviations away from the expectation to be considered in finding boundary mass values.
maxNClusters	The maximum number of clusters that can be considered by the SALSO method.
nCores	The number of CPU cores to use. A value of zero indicates to use all cores on the system.

Details

The mass argument is the main tuning parameter governing the number of clusters, with higher values tending toward more clusters. The mass is a real number bounded below by $-$discount. When a vector of mass values is supplied, a clustering estimate for each mass value is generated and the best clustering estimate is returned.

Alternatively, a range for the number of clusters to be considered can be supplied with the nClusters argument. Mass values that return a clustering estimate with the minimum and maximum value of the range will estimated. A grid of mass values (of length gridLength) between the estimated min and max cluster mass values will be considered in the search for a clustering estimate. If nClusters is a single integer, then a clustering estimate with nClusters clusters will be returned.

Value

A object of class salso.estimate, which provides a clustering estimate (a vector of cluster labels) that can be displayed and plotted.

References

D. B. Dahl, D. J. Johnson, and P. Müller (2021), Search Algorithms and Loss Functions for Bayesian Clustering, arXiv:2105.04451.

Examples

```
# To reduce load on CRAN servers, limit the number of samples, grid length, and CPU cores.
set.seed(34)
iris.dis <- dist(iris[,-5])
est <- caviarpd(distance=iris.dis, mass=c(1, 2), nSamples=50, nCores=1)
samples <- caviarpd(distance=iris.dis, mass=1, nSamples=50, samplesOnly=TRUE, nCores=1)
est <- caviarpd(distance=iris.dis, nClusters=3, nSamples=50, nCores=1)
est <- caviarpd(distance=iris.dis, nClusters=3:5, nSamples=50, gridLength=5, nCores=1)
summ <- summary(est, orderingMethod=2)
plot(summ, type="heatmap")
plot(summ, type="mds")
```

Index