Package ‘pvclust’

November 19, 2019

Version 2.2-0
Date 2019-11-19
Title Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling
Author Ryota Suzuki <suzuki@ef-prime.com>,
Yoshikazu Terada <terada@simgath.es.osaka-u.ac.jp>,
Hidetoshi Shimodaira <shimo@i.kyoto-u.ac.jp>
Maintainer Ryota Suzuki <suzuki@ef-prime.com>
Depends R (>= 2.10.0)
Suggests MASS, parallel
Description An implementation of multiscale bootstrap resampling for assessing the uncertainty in hierarchical cluster analysis.
It provides SI (selective inference) p-value, AU (approximately unbiased) p-value and BP (bootstrap probability) value for each cluster in a dendrogram.
License GPL (>= 2)
URL http://stat.sys.i.kyoto-u.ac.jp/prog/pvclust/
NeedsCompilation no
Repository CRAN
Date/Publication 2019-11-19 12:10:02 UTC

R topics documented:

lung ................................................................. 2
msfit ............................................................... 3
msplot ............................................................ 5
plot.pvclust ..................................................... 5
print.pvclust ................................................... 7
pvclust .......................................................... 8
pvpick ............................................................ 11
seplot ............................................................ 13

Index 14
DNA Microarray Data of Lung Tumors

Description
DNA Microarray data of 73 lung tissues including 67 lung tumors. There are 916 observations of genes for each lung tissue.

Usage
data(lung)

Format
data frame of size 916 × 73.

Details
This dataset has been modified from original data. Each one observation of duplicate genes has been removed. See source section in this help for original data source.

Source

References

Examples

## Reading the data
data(lung)

## Multiscale Bootstrap Resampling
lung.pv <- pvclust(lung, nboot=100)

## CAUTION: nboot=100 may be too small for actual use.
## We suggest nboot=1000 or larger.
## plot/print functions will be useful for diagnostics.

## Plot the result
plot(lung.pv, cex=0.8, cex.pv=0.7)

ask.bak <- par()$ask
par(ask=TRUE)
msfit

Curve Fitting for Multiscale Bootstrap Resampling

Description

msfit performs curve fitting for multiscale bootstrap resampling. It generates an object of class msfit. Several generic methods are available.

Usage

msfit(bp, r, nboot)

## S3 method for class 'msfit'
plot(x, curve=TRUE, main=NULL, sub=NULL, xlab=NULL, ylab=NULL, ...)

## S3 method for class 'msfit'
lines(x, col=2, lty=1, ...)

## S3 method for class 'msfit'
summary(object, digits=3, ...)

Arguments

bp numeric vector of bootstrap probability values.

r numeric vector of relative sample size of bootstrap samples defined as \( r = n'/n \)

for original sample size \( n \) and bootstrap sample size \( n' \).

nboot numeric value (vector) of the number of bootstrap replications.
x          object of class msfit.
curve      logical. If TRUE, the fitted curve is drawn.
main, sub, xlab, ylab, col, lty
          generic graphic parameters.
object     object of class msfit.
digits     integer indicating the precision to be used in rounding.
...        other parameters to be used in the functions.

Details

function msfit performs the curve fitting for multiscale bootstrap resampling. In package pvclust this function is only called from the function pvclust (or parPvclust), and may never be called from users. However one can access a list of msfit objects by x$msfit, where x is an object of class pvclust.

Value

msfit returns an object of class msfit. It contains the following objects:

p          numeric vector of p-values. au is AU (Approximately Unbiased) p-value computed by multiscale bootstrap resampling, which is more accurate than BP value (explained below) as unbiased p-value. bp is BP (Bootstrap Probability) value, which is simple but tends to be unbiased when the absolute value of c (a value in coef vector, explained below) is large.
se         numeric vector of estimated standard errors of p-values.
coef       numeric vector related to geometric aspects of hypotheses. v is signed distance and c is curvature of the boundary.
df         numeric value of the degree of freedom in curve fitting.
rss        residual sum of squares.
pchi       p-value of chi-square test based on asymptotic theory.

Author(s)

Ryota Suzuki <suzuki@ef-prime.com>

References


msplot

Drawing the Results of Curve Fitting for pvclust Object

Description
draws the results of curve fitting for pvclust object.

Usage
msplot(x, edges=NULL, ...)

Arguments
x object of class pvclust.
edges numeric vector of edge numbers to be plotted.
... other parameters to be used in the function.

Author(s)
Ryota Suzuki <suzuki@ef-prime.com>

See Also
plot.msfit

plot.pvclust

Draws Dendrogram with P-values for Pvclust Object

Description
plot dendrogram for a pvclust object and add p-values for clusters.

Usage
## S3 method for class 'pvclust'
plot(x, print.pv=TRUE, print.num=TRUE, float=0.01,
col.pv=c(si=4, au=2, bp=3, edge=8), cex.pv=0.8, font.pv=NULL,
col=NULL, cex=NULL, font=NULL, lty=NULL, lwd=NULL, main=NULL,
sub=NULL, xlab=NULL, ylab=NULL, ...)  
## S3 method for class 'pvclust'
text(x, col=c(au=2, bp=3, edge=8), print.num=TRUE,
float=0.01, cex=NULL, font=NULL, ...)
Arguments

- **x**: object of class pvclust, which is generated by function pvclust. See pvclust for details.
- **print.pv**: logical flag to specify whether print p-values around the edges (clusters), or character vector of length 0 to 3 which specifies the names of p-values to print (c("si","au","bp") for example).
- **print.num**: logical flag to specify whether print edge numbers below clusters.
- **float**: numeric value to adjust the height of p-values from edges.
- **col.pv**: named numeric vector to specify the colors for p-values and edge numbers. For back compatibility it can also be unnamed numeric vector of length 3, which corresponds to the color of AU, BP values and edge numbers.
- **cex.pv**: numeric value which specifies the size of characters for p-values and edge numbers. See cex argument for par.
- **font.pv**: numeric value which specifies the font of characters for p-values and edge numbers. See font argument for par.
- **col, cex, font**: in text function, they correspond to col.pv, cex.pv and font.pv in plot function, respectively. In plot function they are used as generic graphic parameters.
- **lty, lwd, main, sub, xlab, ...**: generic graphic parameters. See par for details.

Details

This function plots a dendrogram with p-values for given object of class pvclust. SI p-value (printed in blue color in default) is the approximately unbiased p-value for selective inference, and AU p-value (printed in red color in default) is also the approximately unbiased p-value but for non-selective inference. They are calculated by multiscale bootstrap resampling. BP value (printed in green color in default) is "bootstrap probability" value, which is less accurate than AU value as p-value. One can consider that clusters (edges) with high SI or AU values (e.g. 95%) are strongly supported by data. SI value is newly introduced in Terada and Shimodaira (2017) for selective inference, which is more appropriate for testing clusters identified by looking at the tree. AU value has been used since Shimodaira (2002), which is not designed for selective inference. AU is valid when you know the clusters before looking at the data. See also documentation (Multiscale Bootstrap using Scaleboot Package, version 0.4-0 or higher) in scaleboot package.

Author(s)

Ryota Suzuki <suzuki@ef-prime.com>

References

### Description

Print clustering method and distance measure used in hierarchical clustering, \( p \)-values and related statistics for a `pvclust` object.

### Usage

```r
## S3 method for class 'pvclust'
print(x, which=NULL, digits=3, ...)
```

### Arguments

- **x**: object of class `pvclust`.
- **which**: numeric vector which specifies the numbers of edges (clusters) of which the values are printed. If `NULL` is given, it prints the values of all edges. The default is `NULL`.
- **digits**: integer indicating the precision to be used in rounding.
- **...**: other parameters used in the function.

### Value

This function prints \( p \)-values and some related statistics.

- **au**: AU (Approximately Unbiased) \( p \)-value, which is more accurate than BP value as unbiased \( p \)-value. It is computed by multiscale bootstrap resampling.
- **bp**: BP (Bootstrap Probability) value, which is a simple statistic computed by bootstrap resampling. This value tends to be biased as \( p \)-value when the absolute value of \( c \) (explained below) is large.
- **se.au, se.bp**: estimated standard errors for `au` and `bp`, respectively.
- **v, c**: values related to geometric aspects of hypotheses. \( v \) is signed distance and \( c \) is curvature of the boundary.
- **pchi**: \( p \)-values of chi-square test based on asymptotic theory.

### Author(s)

Ryota Suzuki <suzuki@ef-prime.com>
Calculating P-values for Hierarchical Clustering

Description

calculates p-values for hierarchical clustering via multiscale bootstrap resampling. Hierarchical clustering is done for given data and p-values are computed for each of the clusters.

Usage

```
pvclust(data, method.hclust="average", method.dist="correlation", use.cor="pairwise.complete.obs", nboot=1000, parallel=FALSE, r=seq(.5,1.4,by=.1), store=FALSE, weight=FALSE, iseed=NULL, quiet=FALSE)
```

```
parPvclust(cl=NULL, data, method.hclust="average", method.dist="correlation", use.cor="pairwise.complete.obs", nboot=1000, r=seq(.5,1.4,by=.1), store=FALSE, weight=FALSE, init.rand=NULL, iseed=NULL, quiet=FALSE)
```

Arguments

data numeric data matrix or data frame.

method.hclust the agglomerative method used in hierarchical clustering. This should be (an abbreviation of) one of "average", "ward.D", "ward.D2", "single", "complete", "mcquitty", "median" or "centroid". The default is "average". See method argument in hclust.

method.dist the distance measure to be used. This should be a character string, or a function which returns a dist object. A character string should be (an abbreviation of) one of "correlation", "uncentered", "abs.cor" or those which are allowed for method argument in dist function. The default is "correlation". See details section in this help and method argument in dist.

use.cor character string which specifies the method for computing correlation with data including missing values. This should be (an abbreviation of) one of "all.obs", "complete.obs" or "pairwise.complete.obs". See the use argument in cor function.

nboot the number of bootstrap replications. The default is 1000.

parallel switch for parallel computation. If FALSE the computation is done in non-parallel mode. If TRUE or a positive integer is supplied, parallel computation is done with automatically generated PSOCK cluster. Use TRUE for default cluster size (parallel::detectCores() -1), or specify the size by an integer. If a cluster object is supplied the cluster is used for parallel computation. Note that NULL is currently not allowed for using the default cluster.
pvclust

numeric vector which specifies the relative sample sizes of bootstrap replications. For original sample size \( n \) and bootstrap sample size \( n' \), this is defined as \( r = n'/n \).

store
logical. If store=TRUE, all bootstrap replications are stored in the output object. The default is FALSE.

cl
a cluster object created by package parallel or snow. If NULL, use the registered default cluster.

weight
logical. If weight=TRUE, resampling is made by weight vector instead of index vector. Useful for large \( r \) value (\( r>10 \)). Currently, available only for distance \"correlation\" and \"abscor\".

init.rand
logical. If init.rand=TRUE, random number generators are initialized. Use iseed argument to achieve reproducible results. This argument is duplicated and will be unavailable in the future.

iseed
An integer. If non-NULL value is supplied random number generators are initialized. It is passed to set.seed or clusterSetRNGStream.

quiet
logical. If TRUE it does not report the progress.

Details
Function pvclust conducts multiscale bootstrap resampling to calculate \( p \)-values for each cluster in the result of hierarchical clustering. parPvclust is the parallel version of this procedure which depends on package parallel for parallel computation.

For data expressed as \((n \times p)\) matrix or data frame, we assume that the data is \( n \) observations of \( p \) objects, which are to be clustered. The \( i \)'th row vector corresponds to the \( i \)'th observation of these objects and the \( j \)'th column vector corresponds to a sample of \( j \)'th object with size \( n \).

There are several methods to measure the dissimilarities between objects. For data matrix \( X = \{x_{ij}\} \), "correlation" method takes

\[
1 - \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2} \sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}}
\]

for dissimilarity between \( j \)'th and \( k \)'th object, where \( \bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \) and \( \bar{x}_k = \frac{1}{n} \sum_{i=1}^{n} x_{ik} \).

"uncentered" takes uncentered sample correlation

\[
1 - \frac{\sum_{i=1}^{n} x_{ij}x_{ik}}{\sqrt{\sum_{i=1}^{n} x_{ij}^2} \sqrt{\sum_{i=1}^{n} x_{ik}^2}}
\]

and "abscor" takes the absolute value of sample correlation

\[
1 - \left| \frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2} \sqrt{\sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2}} \right| .
\]

Value
hclust hierarchical clustering for original data generated by function hclust. See hclust for details.
edges  data frame object which contains \( p \)-values and supporting informations such as standard errors.

count  data frame object which contains primitive information about the result of multiscale bootstrap resampling.

msfit  list whose elements are results of curve fitting for multiscale bootstrap resampling, of class \texttt{msfit}. See \texttt{msfit} for details.

nboot  numeric vector of number of bootstrap replications.

r  numeric vector of the relative sample size for bootstrap replications.

store  list contains bootstrap replications if \texttt{store=TRUE} was given for function \texttt{pvclust} or \texttt{parPvclust}.

version  \texttt{package\_version} of \texttt{pvclust} used to generate this object.

\textbf{Author(s)}

Ryota Suzuki <suzuki@ef-prime.com>

\textbf{References}


http://www.sigmath.es.osaka-u.ac.jp/shimo-lab/prog/pvclust/

\textbf{See Also}

\texttt{lines.pvclust}, \texttt{print.pvclust}, \texttt{msfit}, \texttt{plot.pvclust}, \texttt{text.pvclust}, \texttt{pvrect} and \texttt{pvpick}.

\textbf{Examples}

```r
### example using Boston data in package MASS
data(Boston, package = "MASS")

## multiscale bootstrap resampling (non-parallel)
boston.pv <- pvclust(Boston, nboot=100, parallel=FALSE)

## CAUTION: nboot=100 may be too small for actual use.
## We suggest nboot=1000 or larger.
## plot/print functions will be useful for diagnostics.

## plot dendrogram with p-values
plot(boston.pv)
```
ask.bak <- par()$ask
par(ask=TRUE)

## highlight clusters with high au p-values
pvrect(boston.pv)

## print the result of multiscale bootstrap resampling
print(boston.pv, digits=3)

## plot diagnostic for curve fitting
msplot(boston.pv, edges=c(2,4,6,7))

par(ask=ask.bak)

## print clusters with high p-values
boston.pp <- pvpick(boston.pv)
boston.pp

### Using a custom distance measure

## Define a distance function which returns an object of class "dist".
## The function must have only one argument "x" (data matrix or data.frame).
cosine <- function(x) {
  x <- as.matrix(x)
  y <- t(x) %*% x
  res <- 1 - y / (sqrt(diag(y)) %*% t(sqrt(diag(y))))
  res <- as.dist(res)
  attr(res, "method") <- "cosine"
  return(res)
}

result <- pvclust(Boston, method.dist=cosine, nboot=100)
plot(result)

## Not run:
### parallel computation
result.par <- pvclust(Boston, nboot=1000, parallel=TRUE)
plot(result.par)

## End(Not run)

---

**pvpick**  
*Find Clusters with High/Low P-values*

**Description**

find clusters with relatively high/low $p$-values. *pvrect* and *lines* (S3 method for class *pvclust*) highlight such clusters in existing plot, and *pvpick* returns a list of such clusters.
Usage

pvpick(x, alpha=0.95, pv="au", type="geq", max.only=TRUE)

pvrect(x, alpha=0.95, pv="au", type="geq", max.only=TRUE, border=NULL, ...)

## S3 method for class 'pvclust'
lines(x, alpha=0.95, pv="au", type="geq", col=2, lwd=2, ...)

Arguments

  x    object of class pvclust.

  alpha threshold value for p-values.

  pv    character string which specifies the p-value to be used. It should be one of "si", "au" and "bp", corresponding to SI, AU p-value and BP value, respectively. See plot.pvclust for details.

  type one of "geq", "leq", "gt" or "lt". If "geq" is specified, clusters with p-value greater than or equals the threshold given by "alpha" are returned or displayed. Likewise "leq" stands for lower than or equals, "gt" for greater than and "lt" for lower than the threshold value. The default is "geq".

  max.only logical. If some of clusters with high/low p-values have inclusion relation, only the largest cluster is returned (or displayed) when max.only=TRUE.

  border numeric value which specifies the color of borders of rectangles.

  col numeric value which specifies the color of lines.

  lwd numeric value which specifies the width of lines.

  ... other graphic parameters to be used.

Value

pvpick returns a list which contains the following values.

  clusters a list of character string vectors. Each vector corresponds to the names of objects in each cluster.

  edges numeric vector of edge numbers. The i'th element (number) corresponds to the i'th name vector in clusters.

Author(s)

Ryota Suzuki <suzuki@ef-prime.com>
seplot

---

**seplot**  
*Diagnostic Plot for Standard Error of p-value*

**Description**

draws diagnostic plot for standard error of p-value for pvclust object.

**Usage**

```r
seplot(object, type=c("au", "si", "bp"), identify=FALSE, main=NULL, xlab=NULL, ylab=NULL, ...)
```

**Arguments**

- **object**: object of class `pvclust`.
- **type**: the type of p-value to be plotted, one of "si", "au" or "bp".
- **identify**: logical. If TRUE, edge numbers can be identified interactively. See `identify` for basic usage.
- **main, xlab, ylab**: generic graphic parameters. See `par` for details.
- **...**: other graphical parameters to be passed to generic `plot` or `identify` function.

**Author(s)**

Ryota Suzuki <suzuki@ef-prime.com>
Index

*Topic aplot
  pvpick, 11
*Topic cluster
  pvclust, 8
*Topic datasets
  lung, 2
*Topic hplot
  msplot, 5
  plot.pvclust, 5
  seplot, 13
*Topic htest
  msfit, 3
*Topic print
  print.pvclust, 7

  cor, 8
  dist, 8
  hclust, 8, 9

  identify, 13

  lines.msfit (msfit), 3
  lines.pvclust, 10
  lines.pvclust (pvpick), 11
  lung, 2

  msfit, 3, 10
  msplot, 5

  package_version, 10
  par, 6, 13
  parPvclust (pvclust), 8
  plot.msfit, 5
  plot.msfit (msfit), 3
  plot.pvclust, 5, 10
  print.pvclust, 7, 10
  pvclust, 6, 8
  pvpick, 10, 11
  pvrect, 10

  pvrect (pvpick), 11
  seplot, 13
  summary.msfit (msfit), 3
  text.pvclust, 7, 10
  text.pvclust (plot.pvclust), 5

  14