Package ‘pvclust’

October 23, 2015

Version 2.0-0
Date 2015-10-23
Title Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling
Author Ryota Suzuki <suzuki@ef-prime.com>, Hidetoshi Shimodaira <shimo@sigmath.es.osaka-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 AU (approximately unbiased) p-value as well as BP (bootstrap probability) value for each cluster in a dendrogram.
License GPL (>= 2)
URL http://www.sigmath.es.osaka-u.ac.jp/shimo-lab/prog/pvclust/
NeedsCompilation no
Repository CRAN
Date/Publication 2015-10-23 16:23:16

R topics documented:

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>lung</td>
<td>2</td>
</tr>
<tr>
<td>msplit</td>
<td>3</td>
</tr>
<tr>
<td>msplot</td>
<td>5</td>
</tr>
<tr>
<td>plot.pvclust</td>
<td>5</td>
</tr>
<tr>
<td>print.pvclust</td>
<td>7</td>
</tr>
<tr>
<td>pvclust</td>
<td>8</td>
</tr>
<tr>
<td>pvpick</td>
<td>11</td>
</tr>
<tr>
<td>seplot</td>
<td>13</td>
</tr>
</tbody>
</table>

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**

```r
data(lung)
```

**Format**

data frame of size $916 \times 73$.

**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**

```r
## 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)
```
### Description

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

### Usage

```r
msfit(bp, r, nboot)
```

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

- **S3 method for class 'msfit'**
  ```r
  lines(x, col=2, lty=1, ...)
  ```

- **S3 method for class 'msfit'**
  ```r
  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(2,3,8), cex.pv=0.8, font.pv=NULL, col=NULL, cex=NULL,
     font=NULL, lty=NULL, lwd=NULL, main=NULL, sub=NULL, xlab=NULL, ...)

## S3 method for class 'pvclust'
text(x, col=c(2,3,8), 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 above the edges (clusters).

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
numeric vector of length three to specify the colors for \( p \)-values and edge numbers. From the beginning each value corresponds to the color of AU values, BP values and edge numbers, respectively.

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. AU \( p \)-value (printed in red color in default) is the abbreviation of "approximately unbiased" \( p \)-value, which is 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 AU values (e.g. 95\%) are strongly supported by data.

Author(s)

Ryota Suzuki <suzuki@ef-prime.com>

References


See Also
text.pvclust
Description

print clustering method and distance measure used in hierarchical clustering, \( p \)-values and related statistics for a \texttt{pvclust} object.

Usage

\[
\texttt{## S3 method for class 'pvclust'} \\
\texttt{print(x, which=NULL, digits=3, \ldots)}
\]

Arguments

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

Value

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

- \texttt{au} AU (Approximately Unbiased) \( p \)-value, which is more accurate than BP value as unbiased \( p \)-value. It is computed by multiscale bootstrap resampling.
- \texttt{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 \texttt{c} (explained below) is large.
- \texttt{se.au, se.bp} estimated standard errors for \texttt{au} and \texttt{bp}, respectively.
- \texttt{v, c} values related to geometric aspects of hypotheses. \texttt{v} is signed distance and \texttt{c} is curvature of the boundary.
- \texttt{pchi} \( p \)-values of chi-square test based on asymptotic theory.

Author(s)

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

```
qvclust(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)
```

Arguments

data          numeric data matrix or data frame.
method.hclust  the agglomerative method used in hierarchical clustering. This should be (an ab-
                breviation 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.
\( r \)
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 \( r = n'/n \).

\( \text{store} \)
logical. If \( \text{store} = \text{TRUE} \), all bootstrap replications are stored in the output object. The default is \( \text{FALSE} \).

\( \text{cl} \)
a cluster object created by package \texttt{parallel} or \texttt{snow}. If \( \text{NULL} \), use the registered default cluster.

\( \text{weight} \)
logical. If \( \text{weight} = \text{TRUE} \), resampling is made by weight vector instead of index vector. Useful for large \( r \) value (\( r > 1 \)). Currently, available only for distance "correlation" and "abscor".

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

\( \text{iseed} \)
An integer. If non-\( \text{NULL} \) value is supplied random number generators are initialized. It is passed to \texttt{set.seed} or \texttt{clusterSetRNGStream}.

\( \text{quiet} \)
logical. If \( \text{TRUE} \) it does not report the progress.

Details

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

\texttt{hclust} hierarchical clustering for original data generated by function \texttt{hclust}. See \texttt{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 msfit. See 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 store=TRUE was given for function pvclust or parPvclust.

Author(s)

Ryota Suzuki <suzuki@ef-prime.com>

References


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

See Also

lines.pvclust, print.pvclust, msfit, plot.pvclust, text.pvclust, pvrect and pvpick.

Examples

### 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=2, ...)

# 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 either of "au" or "bp", corresponding to AU p-value or 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

Diagnostic Plot for Standard Error of p-value

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

Usage
seplot(object, type=c("au", "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 "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**  
  mspplot, 5
  plot.pvclust, 5
  seplot, 13
*Topic **htest**  
  msfit, 3
*Topic **print**  
  print.pvclust, 7

cor, 8

dist, 8

cclust, 8, 9

cidentify, 13

clines.msfit (msfit), 3

clines.pvclust, 10

clines.pvclust (pvpick), 11

clung, 2

cmsfit, 3, 10

cmspplot, 5

cpar, 6, 13

cparPvclust (pvclust), 8

cplot.msfit, 5

cplot.msfit (msfit), 3

cplot.pvclust, 5, 10

cprint.pvclust, 7, 10

cpvclust, 6, 8

cpvpick, 10, 11

cpvrect, 10

cpvrect (pvpick), 11