

Package ‘pvclust’

August 9, 2006

Version 1.2-0

Date 2006-07-11

Title Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling

Author Ryota Suzuki <ryota.suzuki@is.titech.ac.jp>, Hidetoshi Shimodaira <shimo@is.titech.ac.jp>

Maintainer Ryota Suzuki <ryota.suzuki@is.titech.ac.jp>

Depends R (>= 1.8.0)

Suggests MASS, snow, Rmpi

Description pvclust is a package for assessing the uncertainty in hierarchical cluster analysis. It provides AU (approximately unbiased) p-values as well as BP (bootstrap probability) values computed via multiscale bootstrap resampling.

License GPL version 2 or newer

URL <http://www.is.titech.ac.jp/~shimo/prog/pvclust/>

R topics documented:

lung	2
msfit	3
msplot	4
plot.pvclust	5
print.pvclust	6
pvclust	7
pvpick	10
seplot	11
Index	12

lung

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

http://genome-www.stanford.edu/lung_cancer/adeno/

References

Garber, M. E. et al. (2001) "Diversity of gene expression in adenocarcinoma of the lung", *Proceedings of the National Academy of Sciences*, 98, 13784-13789.

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)

pvrect(lung.pv, alpha=0.9)
msplot(lung.pv, edges=c(51,62,68,71))

par(ask=ask.bak)

## Print a cluster with high p-value
```

```

lung.pp <- pvpick(lung.pv, alpha=0.9)
lung.pp$clusters[[2]]

## Print its edge number
lung.pp$edges[2]

## We recommend parallel computing for large dataset as this one
## Not run:
library(snow)
cl <- makeCluster(10, type="MPI")
lung.pv <- parPvclust(cl, lung, nboot=1000)
## End(Not run)

```

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:

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

Author(s)

Ryota Suzuki <ryota.suzuki@is.titech.ac.jp>

References

- Shimodaira, H. (2004) "Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling", *Annals of Statistics*, 32, 2616-2641.
- Shimodaira, H. (2002) "An approximately unbiased test of phylogenetic tree selection", *Systematic Biology*, 51, 492-508.

`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

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

Author(s)

Ryota Suzuki <ryota.suzuki@is.titech.ac.jp>

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

<code>x</code>	object of class <code>pvclust</code> , which is generated by function <code>pvclust</code> . See pvclust for details.
<code>print.pv</code>	logical flag to specify whether print p -values above the edges (clusters).
<code>print.num</code>	logical flag to specify whether print edge numbers below clusters.
<code>float</code>	numeric value to adjust the height of p -values from edges.
<code>col.pv</code>	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.
<code>cex.pv</code>	numeric value which specifies the size of characters for p -values and edge numbers. See <code>codecex</code> argument for par .
<code>font.pv</code>	numeric value which specifies the font of characters for p -values and edge numbers. See <code>codefont</code> argument for par .
<code>col, cex, font</code>	in <code>text</code> function, they correspond to <code>col.pv</code> , <code>cex.pv</code> and <code>font.pv</code> in <code>plot</code> function, respectively. In <code>plot</code> function they are used as generic graphic parameters.
<code>lty, lwd, main, sub, xlab, ...</code>	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 <ryota.suzuki@is.titech.ac.jp>

References

Shimodaira, H. (2004) "Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling", *Annals of Statistics*, 32, 2616-2641.

Shimodaira, H. (2002) "An approximately unbiased test of phylogenetic tree selection", *Systematic Biology*, 51, 492-508.

See Also

[text.pvclust](#)

`print.pvclust`

Print Function for Pvclust Object

Description

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

Usage

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

Arguments

<code>x</code>	object of class <code>pvclust</code> .
<code>which</code>	numeric vector which specifies the numbers of edges (clusters) of which the values are printed. If <code>NULL</code> is given, it prints the values of all edges. The default is <code>NULL</code> .
<code>digits</code>	integer indicating the precision to be used in rounding.
<code>...</code>	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 <ryota.suzuki@is.titech.ac.jp>

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

```
pvclust(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)

parPvclust(cl, 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=TRUE, seed=NULL)
```

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", "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 (an abbreviation of) one of "correlation", "uncentered", "abscor" or those which are allowed for method argument in dist function. The default is "correlation". See <i>details</i> section in this help and method argument in dist .

<code>use.cor</code>	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 <code>use</code> argument in <code>cor</code> function.
<code>nboot</code>	the number of bootstrap replications. The default is 1000.
<code>r</code>	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 $r = n'/n$.
<code>store</code>	logical. If <code>store=TRUE</code> , all bootstrap replications are stored in the output object. The default is <code>FALSE</code> .
<code>cl</code>	snow cluster object which may be generated by function <code>makeCluster</code> . See <code>snow-startstop</code> in snow package.
<code>weight</code>	logical. If <code>weight=TRUE</code> , 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".
<code>init.rand</code>	logical. If <code>init.rand=TRUE</code> , random number generators are initialized at child processes. Random seeds can be set by <code>seed</code> argument.
<code>seed</code>	integer vector of random seeds. It should have the same length as <code>cl</code> . If <code>NULL</code> is specified, <code>1:length(cl)</code> is used as seed vector. The default is <code>NULL</code> .

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 **snow** package 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 <code>msfit</code> . See <code>msfit</code> 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 <code>store=TRUE</code> was given for function <code>pvclust</code> or <code>parPvclust</code> .

Author(s)

Ryota Suzuki <ryota.suzuki@is.titech.ac.jp>

References

Shimodaira, H. (2004) "Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling", *Annals of Statistics*, 32, 2616-2641.

Shimodaira, H. (2002) "An approximately unbiased test of phylogenetic tree selection", *Systematic Biology*, 51, 492-508.

Suzuki, R. and Shimodaira, H. (2004) "An application of multiscale bootstrap resampling to hierarchical clustering of microarray data: How accurate are these clusters?", *The Fifteenth International Conference on Genome Informatics 2004*, P034.

<http://www.is.titech.ac.jp/~shimo/prog/pvclust/>

See Also

[lines.pvclust](#), [print.pvclust](#), [msfit](#), [plot.pvclust](#), [text.pvclust](#), [pvrect](#) and [vpick](#).

Examples

```
## using Boston data in package MASS
library(MASS)
data(Boston)

## multiscale bootstrap resampling
boston.pv <- pvclust(Boston, 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 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

## Not run:
## parallel computation via snow package
library(snow)
cl <- makeCluster(10, type="MPI")

## parallel version of pvclust
boston.pv <- parPvclust(cl, Boston, nboot=1000)
## 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

<code>x</code>	object of class <code>pvclust</code> .
<code>alpha</code>	threshold value for p -values.
<code>pv</code>	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 <code>plot.pvclust</code> for details.
<code>type</code>	one of "geq", "leq", "gt" or "lt". If "geq" is specified, clusters with p -value <i>greater than or equals</i> the threshold given by "alpha" are returned or displayed. Likewise "leq" stands for <i>lower than or equals</i> , "gt" for <i>greater than</i> and "lt" for <i>lower than</i> the threshold value. The default is "geq".
<code>max.only</code>	logical. If some of clusters with high/low p -values have inclusion relation, only the largest cluster is returned (or displayed) when <code>max.only=TRUE</code> .

<code>border</code>	numeric value which specifies the color of borders of rectangles.
<code>col</code>	numeric value which specifies the color of lines.
<code>lwd</code>	numeric value which specifies the width of lines.
<code>...</code>	other graphic parameters to be used.

Value

`pvpick` returns a list which contains the following values.

<code>clusters</code>	a list of character string vectors. Each vector corresponds to the names of objects in each cluster.
<code>edges</code>	numeric vector of edge numbers. The i 'th element (number) corresponds to the i 'th name vector in <code>clusters</code> .

Author(s)

Ryota Suzuki <ryota.suzuki@is.titech.ac.jp>

`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

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

Author(s)

Ryota Suzuki <ryota.suzuki@is.titech.ac.jp>

Index

*Topic **aplot**
 pvpick, 10

*Topic **cluster**
 pvclust, 7

*Topic **datasets**
 lung, 1

*Topic **hplot**
 msplot, 4
 plot.pvclust, 5
 seplot, 11

*Topic **htest**
 msfit, 3

*Topic **print**
 print.pvclust, 6

cor, 7

dist, 7

hclust, 7, 8

identify, 11

lines.msfit (*msfit*), 3
lines.pvclust, 9
lines.pvclust (*pvpick*), 10
lung, 1

msfit, 3, 8, 9
msplot, 4

par, 5, 11
parPvclust (*pvclust*), 7
plot.msfit, 4
plot.msfit (*msfit*), 3
plot.pvclust, 5, 9
print.pvclust, 6, 9
pvclust, 5, 7
pvpick, 9, 10
pvrect, 9
pvrect (*pvpick*), 10

seplot, 11
snow-startstop, 7
summary.msfit (*msfit*), 3
text.pvclust, 6, 9
text.pvclust (*plot.pvclust*), 5