

Package ‘segclust’

April 19, 2009

Version 0.74

Date 2008-11-22

Title SegClust : a package for Segmentation and Segmentation/Clustering.

Author Franck Picard <picard@biomserv.univ-lyon1.fr>

Maintainer Mark Hoebeke <Mark.Hoebeke@jouy.inra.fr>

Depends R (>= 2.4)

SystemRequirements libgsl (>= 1.4)

Description SegClust corresponds to the implementation of the statistical model described in : Picard et al., A segmentation/clustering model for the analysis of array CGH data. *Biometrics*, 63(3) 2007. Segmentation functions are also available (from Picard et al. A statistical approach for array CGH data analysis. *BMC Bioinformatics*. 2005 Feb 11;6:27).

License GPL (>= 2)

Repository CRAN

Date/Publication 2008-12-01 19:20:14

R topics documented:

clustersegments	2
DPEM	3
hybrid	4
segclustout	5
segclustselect	7
segmean	8
segmixt	9
segout	10
segselect	12

Index	13
--------------	-----------

clustersegments *clustersegments*

Description

Cluster segments using the EM algorithm when the number of clusters and the segmentation are given. The segmentation is not revised after clustering (contrary to hybrid)

Usage

```
out <- clustersegments(x, P, bp, vh=TRUE)
```

Arguments

x	data vector (without missing values)
P	number of clusters
bp	vector (size length(x)), such that bp[t]=1 if t is a breakpoint and 0 otherwise. (t corresponds to the end of a segment). bp[length(x)] is always equal to 1.
vh	TRUE for homogeneous variances (default), FALSE otherwise

Value

out	dataframe
output\$signal	input signal x
output\$mean	estimated mean using a mixture model with P cluster, AFTER a segmentation
output\$sd	estimated standard deviation using a mixture model with P cluster, AFTER a segmentation
output\$cluster	cluster for each point AFTER segmentation
output\$bp	breakpoint coordinates, equals 1 for a breakpoint (corresponding to the end of the segments)

Author(s)

F. Picard, M. Hoebecke

References

Picard, F., Robin, S., Lebarbier, E., & Daudin, J. -J. (2007). A segmentation/clustering model for the analysis of array CGH data. *Biometrics*, 63(3) 758-766

Examples

```

x1          <- rnorm(20,0,1)
x2          <- rnorm(30,2,1)
x           <- c(x1,x2)
bp         <- rep(0,length(x))
bp[c(20,50)] <- 1
P          <- 2
out        <- clustersegments(x,P,bp)

```

DPEM

*Dynamic Programming / EM segmentation/clustering***Description**

Estimation of segmentation/clustering parameters in the Gaussian case, using dynamic programming and the EM algorithm. This function performs a global analysis with estimation and model selection. It uses functions `hybrid()`, `segclustselect()`, and `seclustout()`. `Pmin` and `Pmax` must be different.

Usage

```
output <- DPEM(x,Pmin,Pmax,Kmax,method,draw,lmin=1,lmax=length(x),vh=TRUE,S
```

Arguments

<code>x</code>	data vector (without missing values)
<code>Pmin</code>	minimum number of clusters
<code>Pmax</code>	maximum number of clusters
<code>Kmax</code>	max number of segments. <code>Kmax</code> must be greater than <code>Pmax</code>
<code>method</code>	model selection method. Equals "sequential" or "BIC"
<code>draw</code>	equals TRUE for a graphical display
<code>lmin</code>	minimum segment length, default value <code>lmin = 1</code>
<code>lmax</code>	maximum segment length, default value <code>lmax = length(x)</code>
<code>vh</code>	TRUE for homogeneous variances (default), FALSE otherwise
<code>S</code>	Threshold for model selection, set at 0.5

Value

<code>output</code>	dataframe containing results of the estimation procedure
<code>output\$signal</code>	input signal <code>x</code>
<code>output\$mean</code>	estimated mean according to the model, for each position
<code>output\$sd</code>	estimated standard deviation according to the model, for each position

```

output$cluster
      cluster for each point
output$bp
      breakpoint coordinates, equals 1 for a breakpoint (corresponding to the end of
      the segments)

```

Author(s)

F. Picard, M. Hoebecke

References

Picard, F., Robin, S., Lebarbier, E., & Daudin, J. -J. (2007). A segmentation/clustering model for the analysis of array CGH data. *Biometrics*, 63(3) 758-766

Examples

```

x1      <- rnorm(20, 0, 1)
x2      <- rnorm(30, 2, 1)
x3      <- rnorm(10, 0, 1)
x4      <- rnorm(40, 2, 1)
x       <- c(x1, x2, x3, x4)
Pmin    <- 1
Pmax    <- 4
Kmax    <- 20
output  <- DPEM(x, Pmin, Pmax, Kmax, method="BIC", draw=TRUE)

```

hybrid *hybrid algorithm for segmentation/clustering*

Description

Estimation of segmentation/clustering parameters in the Gaussian case, using dynamic programming and the EM algorithm. This function estimates the parameters for a model with P clusters and for a number of segments from P to Kmax (Kmin=P).

Usage

```
out.hybrid <- hybrid(x, P, Kmax, lmin=1, lmax=length(x), vh=TRUE)
```

Arguments

x	data vector (without missing values)
P	number of clusters
Kmax	max number of segments. The minimum number of segments is P
lmin	minimum segment length, default value lmin = 1
lmax	maximum segment length, default value lmax = length(x)
vh	TRUE for homogeneous variances (default), FALSE otherwise

Value

`Linc` incomplete-data log-likelihood for a model with P clusters and P to K_{\max} segments

`param[[K]]` list of estimated parameters for a segmentation/clustering model with P clusters and K segments

`param[[K]]$phi` mixture parameters of size $3 \times P$: P means, P standard deviations, P mixture proportions

`param[[K]]$rupt` breakpoints position : matrix with 2 columns (begin and end of segments) and K rows

Author(s)

F. Picard, M. Hoebecke

References

Picard, F., Robin, S., Lebarbier, E., & Daudin, J. -J. (2007). A segmentation/clustering model for the analysis of array CGH data. *Biometrics*, 63(3) 758-766

Examples

```
x1      <- rnorm(20,0,1)
x2      <- rnorm(30,2,1)
x3      <- rnorm(10,0,1)
x4      <- rnorm(40,2,1)
x       <- c(x1,x2,x3,x4)
Pmin    <- 1
Pmax    <- 4
Kmax    <- 20
Linc    <- matrix(-Inf, ncol=Pmax,nrow= Kmax)
param.list <- list()

for (P in (Pmin:Pmax)){
  out.hybrid      <- hybrid(x,P,Kmax)
  param.list[[P]] <- out.hybrid$param
  Linc[,P]       <- out.hybrid$Linc
}
out.select <- segclustselect(x,param=param.list,Pmin,Pmax,Kmax,Linc,method = "BIC")
output    <- segclustout(x,param.list[[out.select$Pselect]],out.select$Pselect,out.
```

segclustout

segclustout

Description

Extraction of parameters for a segmentation/clustering model

Usage

```
out <- segclustout(x, param, P, K, draw)
```

Arguments

x	data vector (without missing values)
param	list of parameters estimated by hybrid for a given P
P	number of clusters
K	number of segments (must be smaller than P)
draw	TRUE for plotting

Value

output	dataframe containing results of the estimation procedure
output\$signal	input signal x
output\$mean	estimated mean according to the model, for each position
output\$sd	estimated standard deviation according to the model, for each position
output\$cluster	cluster for each point
output\$bp	breakpoint coordinates, equals 1 for a breakpoint (corresponding to the end of the segments)

Author(s)

F. Picard, M. Hoebecke

References

Picard, F., Robin, S., Lebarbier, E., & Daudin, J. -J. (2007). A segmentation/clustering model for the analysis of array CGH data. *Biometrics*, 63(3) 758-766

Examples

```
x1 <- rnorm(20, 0, 1)
x2 <- rnorm(30, 2, 1)
x3 <- rnorm(10, 0, 1)
x4 <- rnorm(40, 2, 1)
x <- c(x1, x2, x3, x4)

Pmin <- 1
Pmax <- 4
Kmax <- 20
Linc <- matrix(-Inf, ncol=Pmax, nrow= Kmax)
param.list <- list()

for (P in (Pmin:Pmax)){
```

```

    out.hybrid      <- hybrid(x,P,Kmax)
    param.list[[P]] <- out.hybrid$param
    Linc[,P]       <- out.hybrid$Linc
  }
  out.select <- segclustselect(x,param,Pmin,Pmax,Kmax,Linc, method = "sequential")
  output    <- segclustout(x,param.list[[out.select$Pselect]],out.select$Pselect,out.

```

segclustselect *segclustselect*

Description

Model selection for segmentation/clustering

Usage

```
out <- segclustselect(x,param,Pmin,Pmax,Kmax,Linc,method,S=0.5,lmin=1,lmax=
```

Arguments

x	data vector
param	parameters estimated by hybrid()
Pmin	minimum number of clusters
Pmax	maximum number of clusters
Kmax	maximum number of segments
Linc	incomplete-data log-likelihood calculated by hybrid()
method	Method used of the selection. Equals "sequential" or "BIC"
S	threshold for the adaptive method, default value S = 0.5
lmin	minimal segment length, default value lmin = 1
lmax	maximal segment length, default value lmax = length(x)
vh	TRUE for homogeneous variances (default), FALSE otherwise

Details

This function is used to select simultaneously Pselect and Kselect, the number of clusters and the number of segments in a segmentation/clustering model. It is based on the penalization of the incomplete-data log-likelihood Linc. Two methods are implemented. The first one is based on a sequential choice of Pselect and Kselect as described in Picard et al. (2007). The second one is based on a modified BIC criterion as described in Zhang et al. (2007). The function uses the Stirling approximation of the Gamma function, such that :

$$\log \Gamma(x) \sim (x - 1/2) \times \log(x) - x + 1/2 \times \log(2\pi)$$

Value

Pselect Selected number of clusters
 Kselect Selected number of segments

References

Picard, F., Robin, S., Lebarbier, E., & Daudin, J. -J. (2007). A segmentation/clustering model for the analysis of array CGH data. *Biometrics*, 63(3) 758-766 \ Zhang NR, Siegmund DO. A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*. 2007 63(1):22-32.

Examples

```
x1      <- rnorm(20,0,1)
x2      <- rnorm(30,2,1)
x3      <- rnorm(10,0,1)
x4      <- rnorm(40,2,1)
x       <- c(x1,x2,x3,x4)
Pmin    <- 1
Pmax    <- 4
Kmax    <- 20
Linc    <- matrix(-Inf, ncol=Pmax,nrow= Kmax)
param.list <- list()
for (P in (Pmin:Pmax)){
  out.hybrid <- hybrid(x,P,Kmax)
  param.list[[P]] <- out.hybrid$param
  Linc[,P] <- out.hybrid$Linc
}
out.select <- segclustselect(x,param=param.list,Pmin,Pmax,Kmax,Linc,method = "BIC")
```

 segmean

segmean

Description

segmentation of a signal when considering changes in the mean

Usage

```
out <- segmean(x,Kmax,lmin=1,lmax=length(x),vh=TRUE)
```

Arguments

x data vector
 Kmax maximum number of segments
 lmin minimal segment length, default value lmin = 1
 lmax maximal segment length, default value lmax = length(x)
 vh TRUE for an homogeneous variance (default), FALSE otherwise

Details

This function can be used for a segmentation model such that:

$$\forall t \in I_k \quad Y_t = \mu_k + \varepsilon_t$$

and the variance of ε can be either homoscedastic (vh=TRUE) or heteroscedastic (vh=FALSE). It uses dynamic programming to find the best breakpoints, and is based on the calculus of the Residual Sum of Squares:

$$J.est_K = \sum_{k=1}^K \sum_{t \in I_k} (y_t - \hat{\mu}_k)^2.$$

Value

J.est	Residual Sum of Squares for segmentation models up to Kmax segments
t.est	estimated positions for breakpoints for segmentation models up to Kmax segments

References

Picard, F., Robin, S., Lavielle, M., Vaisse, C., & Daudin, J. -J. (2005). A statistical approach for array CGH data analysis. *BMC Bioinformatics*, 6(1), 1-14.

Examples

```
x1      <- rnorm(20, 0, 1)
x2      <- rnorm(30, 2, 1)
x       <- c(x1, x2)
Kmax    <- 20
out     <- segmean(x, Kmax)
Kselect <- segselect(out$J.est, Kmax)
output  <- segout(x, K=Kselect, th = out$t.est, draw=TRUE)
```

 segmixt

segmixt

Description

segmentation of a signal when considering changes in a mixture of Gaussian vectors

Usage

```
out <- segmixt(x, P, Kmax, phi, lmin=1, lmax=length(x))
```

Arguments

<code>x</code>	data vector
<code>P</code>	number of clusters
<code>Kmax</code>	maximum number of segments
<code>phi</code>	parameters of the mixture
<code>lmin</code>	minimal segment length, default value <code>lmin = 1</code>
<code>lmax</code>	maximal segment length, default value <code>lmax = length(x)</code>

Details

This function can be used for a segmentation/clustering model with P clusters and up to K_{\max} segments. `phi` gives the parameters of the mixture (P means, P standard deviations, P mixture proportions). It uses dynamic programming to find the best breakpoints, and is based on the calculus of the incomplete-data log-likelihood:

$$J.est_{K,P} = \sum_{k=1}^K \log \left\{ \sum_{p=1}^P \pi_p f(y^k; \theta_p) \right\}$$

where $f(y^k; \theta_p)$ is the density of a Gaussian vector of length n_k .

Value

<code>J.est</code>	incomplete-data log-likelihood for a segmentation/clustering model with P clusters and up to K_{\max} segments
<code>t.est</code>	estimated positions of breakpoints for a segmentation/clustering model up to K_{\max} segments

Author(s)

F. Picard, M. Hoebecke

References

Picard, F., Robin, S., Lebarbier, E., & Daudin, J. -J. (2007). A segmentation/clustering model for the analysis of array CGH data. *Biometrics*, 63(3) 758-766

segout

segout

Description

Extraction of parameters for segmentation model

Usage

```
out <- segclustout(x, K, th, draw, vh=TRUE)
```

Arguments

<code>x</code>	data vector (without missing values)
<code>K</code>	number of segments
<code>th</code>	estimated positions for breakpoints for segmentation models up to <code>Kmax</code> segments (<code>t.est</code> from <code>segmean</code>)
<code>draw</code>	TRUE for plotting
<code>vh</code>	Variance homogeneity, default = TRUE

Value

<code>output</code>	dataframe containing results of the estimation procedure
<code>output\$signal</code>	input signal <code>x</code>
<code>output\$mean</code>	estimated mean according to the model, for each position
<code>output\$sd</code>	estimated standard deviation according to the model, for each position
<code>output\$bp</code>	breakpoint coordinates, equals 1 for a breakpoint (corresponding to the end of the segments)

Author(s)

F. Picard, M. Hoebecke

References

Picard, F., Robin, S., Lebarbier, E., & Daudin, J. -J. (2007). A segmentation/clustering model for the analysis of array CGH data. *Biometrics*, 63(3) 758-766

Examples

```
x1      <- rnorm(20, 0, 1)
x2      <- rnorm(30, 2, 1)
x3      <- rnorm(10, 0, 1)
x4      <- rnorm(40, 2, 1)
x       <- c(x1, x2, x3, x4)
Kmax    <- 20
out     <- segmean(x, Kmax)
Kselect <- segselect(out$J.est, Kmax)
output  <- segout(x, K=Kselect, th = out$t.est, draw=TRUE)
```

segselect

segselect

Description

model selection for segmentation

Usage

```
out <- segselect(J, Kmax, S=0.5)
```

Arguments

J	Residual Sum of Squares for segmentation models up to Kmax segments
Kmax	maximum number of segments
S	threshold for the adaptive method, set to 0.5 (default)

Value

Kselect	Selected number of segments
---------	-----------------------------

References

Picard, F., Robin, S., Lavielle, M., Vaisse, C., & Daudin, J. -J. (2005). A statistical approach for array CGH data analysis. *BMC Bioinformatics*, 6(1), 1-14.

Examples

```
x1 <- rnorm(20, 0, 1)
x2 <- rnorm(30, 2, 1)
x <- c(x1, x2)
Kmax <- 20
vh <- TRUE
out <- segmean(x, Kmax)
Kselect <- segselect(out$J.est, Kmax)
output <- segout(x, K=Kselect, th = out$t.est, draw=TRUE)
```

Index

*Topic **cluster**

clustersegments, 1

DPEM, 3

hybrid, 4

segclustout, 5

segclustselect, 7

segout, 10

segselect, 12

*Topic **ts**

segmean, 8

segmixt, 9

clustersegments, 1

DPEM, 3

hybrid, 4

segclustout, 5

segclustselect, 7

segmean, 8

segmixt, 9

segout, 10

segselect, 12