Package ‘clustEff’

December 15, 2017

Type Package

Title Clusters of Effects Curves in Quantile Regression Models

Version 0.1.2


Depends qrcm, cluster, fda

License GPL-2

Encoding UTF-8

LazyData true

RoxygenNote 6.0.1

NeedsCompilation no

Author Gianluca Sottile [aut, cre], Giada Adelfio [aut]

Maintainer Gianluca Sottile <gianluca.sottile@unipa.it>

Repository CRAN

Date/Publication 2017-12-15 08:16:17 UTC

R topics documented:

clustEff-package .................................................. 2
clustEff .............................................................. 5
distshape .............................................................. 8
extract.object ....................................................... 9
plot.clustEff ......................................................... 10
summary.clustEff .................................................. 11

Index 13
Description

This package implements a general algorithm to cluster coefficient functions (i.e. clusters of effects) obtained from a quantile regression (qrcm; Frumento and Bottai, 2016). This algorithm is also used for clustering curves observed in time, as in functional data analysis. The objectives of this algorithm vary with the scenario in which it is used, i.e. in the case of a cluster of effects, in a univariate case the objective may be to reduce its dimensionality or in the multivariate case to group similar effects on a covariate. In the case of a functional data analysis the main objective is to cluster waves or any other function of time or space. Sottile G. and Adelfio G. (2017) <https://iwsm2017.webhosting.rug.nl/IWSM_2017_V2.pdf>.

Details

<table>
<thead>
<tr>
<th>Package:</th>
<th>clustEff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type:</td>
<td>Package</td>
</tr>
<tr>
<td>Version:</td>
<td>0.1.2</td>
</tr>
<tr>
<td>Date:</td>
<td>2017-12-15</td>
</tr>
<tr>
<td>License:</td>
<td>GPL-2</td>
</tr>
</tbody>
</table>

The function clustEff allows to specify the type of the curves to apply the proposed clustering algorithm. The function extract.object extracts the matrices, in case of multivariate response, through the quantile regression coefficient modeling, useful to run the main algorithm. The auxiliary functions summary.clustEff and plot.clustEff can be used to extract information from the main algorithm.

Author(s)

Gianluca Sottile
Maintainer: Gianluca Sottile <gianluca.sottile@unipa.it>

References


Examples

# use simulated data
# CURVES EFFECTS CLUSTERING
set.seed(1234)
n <- 300
q <- 2
k <- 5
x1 <- runif(n, 0, 5)
x2 <- runif(n, 0, 5)
X <- cbind(x1, x2)
rownames(X) <- 1:n
colnames(X) <- paste0("X", 1:q)

theta1 <- matrix(c(1, 1, 0, 0, .5, 0, .5, 1, 2, .5, 0, 2, 1, .5),
ncol=k, byrow=TRUE)
theta2 <- matrix(c(1, 1, 0, 0, -.3, 0, .5, 1, .5, -1.5, 0, -1, -.5, 1),
ncol=k, byrow=TRUE)
theta3 <- matrix(c(1, 1, 0, 0, .3, 0, -.5, -1, 2, -.5, 0, 1, -.5, -1),
ncol=k, byrow=TRUE)
rownames(theta3) <- rownames(theta2) <- rownames(theta1) <-
c("(intercept)", paste("X", 1:q, sep=""))
colnames(theta3) <- colnames(theta2) <- colnames(theta1) <-
c("(intercept)", "qnorm(p)", "p", "p^2", "p^3")

Theta <- list(theta1, theta2, theta3)

B <- function(p, k)(matrix(cbind(1, qnorm(p), p, p^2, p^3), nrow=k, byrow=TRUE))
Q <- function(p, theta, B, k, X)(rowSums(X * t(theta %*% B(p, k))))

Y <- matrix(NA, nrow(X), 15)
for(i in 1:15){
  if(i <= 5) Y[, i] <- Q(runif(n), Theta[[1]], B, k, cbind(1, X))
  if(i <= 10 & i > 5) Y[, i] <- Q(runif(n), Theta[[2]], B, k, cbind(1, X))
  if(i <= 15 & i > 10) Y[, i] <- Q(runif(n), Theta[[3]], B, k, cbind(1, X))
}

XX <- extract.object(Y, X, intercept=TRUE, formula.p= - I(p) + I(p^2) + I(p^3))
seqP <- XX$p

obj <- clustEff(XX$X$X1, Beta.lower=XX$X$l$X1, Beta.upper=XX$X$r$X1)
summary(obj)
plot(obj, xvar="clusters", add=TRUE)
par(mfrow=c(1,3));plot(obj, xvar="clusters", add=FALSE);par(mfrow=c(1,1))
plot(obj, xvar="dendrogram")
plot(obj, xvar="boxplot")

## Not run:
obj2 <- clustEff(XX$X$X2, Beta.lower=XX$X$l$X2, Beta.upper=XX$X$r$X2)
summary(obj2)
plot(obj2, xvar="clusters", add=TRUE)
par(mfrow=c(1,3));plot(obj2, xvar="clusters", add=FALSE);par(mfrow=c(1,1))
plot(obj2, xvar="dendrogram")
plot(obj2, xvar="boxplot")

set.seed(1234)
n <- 300
q <- 15
k <- 5
X <- matrix(rnorm(n*q), n, q); X <- scale(X)
rownames(X) <- 1:n
colnames(X) <- paste0("X", 1:q)

Theta <- matrix(c(1, 1, 0, 0, 0,
                   .5, 0, .5, 1, 1,
                   .5, 0, 1, 2, .5,
                   .5, 0, 1, 1, .5,
                   .5, 0, .5, 1, 1,
                   .5, 0, .5, 1, .5,
                   -1.5, 0, -.5, 1, 1,
                   -1, 0, .5, -1, -1,
                   -.5, 0, -.5, -1, .5,
                   -1, 0, .5, -1, -.5,
                   -1.5, 0, -.5, -1, -.5,
                   2, 0, 1, 1.5, 2,
                   2, 0, .5, 1.5, 2,
                   2.5, 0, 1, 1, 2,
                   1.5, 0, 1.5, 1, 2,
                   3, 0, 2, 1, .5),
                   ncol=k, byrow=TRUE)
rownames(Theta) <- c("(intercept)", paste("X", 1:q, sep=""))
colnames(Theta) <- c("(intercept)", "qnorm(p)", "p", "p^2", "p^3")

B <- function(p, k) {matrix(cbind(1, qnorm(p), p, p^2, p^3), nrow=k, byrow=TRUE))
Q <- function(p, theta, B, k, X) {rowSums(X * t(theta %*% B(p, k)))}

s <- matrix(1, q+1, k)
s[2:(q+1), 2] <- 0
s[1, 3:k] <- 0

Y <- Q(runif(n), Theta, B, k, cbind(1, X))
XX <- iqr(Y ~ X, formula.p= - I(p) + I(p^2) + I(p^3))
seqP <- seq(0.01, .99, l=100)
predObj <- predict(XX, type="beta", p=seqP)
X2 <- X1 <- Xr <- matrix(NA, nrow=length(seqP), ncol=(dim(coef(XX))[1]-1))
for(i in 2:dim(coef(XX))[1]){X2[, (i-1)] <- predObj[i][, 2];
X1[, (i-1)] <- predObj[i][, 4];
Xr[, (i-1)] <- predObj[i][, 5];}

obj <- clustEff(X2, Beta.lower=X1, Beta.upper=Xr)
summary(obj)
plot(obj, xvar="clusters", add=TRUE)
clustEff

Cluster Effects Algorithm

description

This function implements the algorithm to cluster curves of effects obtained from a quantile regression (qrcm; Frumento and Bottai, 2015) in which the coefficients are described by flexible parametric functions of the order of the quantile. This algorithm can be also used for clustering of curves observed in time, as in functional data analysis.

usage

clustEff(Beta, k, alpha, cluster.effects=TRUE, step=c("both", "shape", "distance"),
Arguments

Beta
A matrix n x q. q represents the number of curves to cluster and n is either the length of percentiles used in the quantile regression or the length of the time vector.

k
If fixed, it represents the number of clusters.

alpha
It is the alpha-percentile used for computing the dissimilarity matrix. If not fixed, the algorithm choose alpha=.25 (cluster.effects=TRUE) or alpha=.5 (cluster.effects=FALSE).

cluster.effects
If TRUE, it selects the framework (quantile regression or curves clustering) in which to apply the clustering algorithm.

step
The steps used in computing the dissimilarity matrix. Default is "both"="shape" and "distance"

k.max
The maximum number of clusters to let the algorithm to choose the best.

Beta.lower
A matrix n x q. q represents the number of lower interval of the curves to cluster and n the length of percentiles used in quantile regression. Used only if cluster.effects=TRUE.

Beta.upper
A matrix n x q. q represents the number of upper interval of the curves to cluster and n the length of percentiles used in quantile regression. Used only if cluster.effects=TRUE.

ask
If TRUE, after plotting the dendrogram, the user make is own choice about how many cluster to use.

approx.spline
If TRUE, Beta is approximated by a smooth spline.

nbasis
An integer variable specifying the number of basis functions. Only when approx.spline=TRUE

method
The agglomeration method to be used.

plot
If TRUE, dendrogram, boxplot and clusters are plotted.

trace
If TRUE, some informations are printed.

Details

Quantile regression models conditional quantiles of a response variable, given a set of covariates. Assume that each coefficient can be expressed as a parametric function of \( p \) in the form:

\[
\beta(p|\theta) = \theta_0 + \theta_1 b_1(p) + \theta_2 b_2(p) + \ldots
\]

where \( b_1(p), b_2(p), \ldots \) are known functions of \( p \).
Value

An object of class “clustEff”, a list containing the following items:

call the matched call.
X The curves matrix.
X.mean The mean curves matrix of dimension n x k.
X.mean.dist The within cluster distance from the mean curve.
X.lower The lower interval matrix.
X.mean.lower The mean lower interval of dimension n x k.
X.upper The upper interval matrix.
X.mean.upper The mean upper interval of dimension n x k.
k The number of selected clusters.
p The percentiles used in quantile regression coefficient modeling or the time otherwise.
diss.matrix The dissimilarity matrix.
X.mean.diss The within cluster dissimilarity.
oggSilhouette An object of class “silhouette”.
oggHclust An object of class “hclust”.
clusters The vector of clusters.
distance A vector of goodness measures used to select the best number of clusters.
step The selected step.
method The used agglomeration method.
cut.method The used method to select the best number of clusters.
alpha The selected alpha-percentile.

Author(s)

Gianluca Sottile <gianluca.sottile@unipa.it>

References


See Also

summary.clustEff, plot.clustEff, for summary and plotting. extract.object to extract useful objects for the clustering algorithm through a quantile regression coefficient modeling in a multivariate case.
Examples

#### Using simulated data in all examples

# see the documentation for 'clustEff-package'

---

**distshape**

**Dissimilarity matrix**

---

**Description**

This function implements the dissimilarity matrix based on shape and distance of curves.

**Usage**

```r
distshape(Beta, alpha=.5, step=c("both", "shape", "distance"), trace=TRUE)
```

**Arguments**

- **Beta**: A matrix of dimension \( n \times q \). \( q \) represents the number of curves to cluster and \( n \) is either the length of percentiles used in the quantile regression or the length of the time vector.
- **alpha**: It is the alpha-percentile used for computing the dissimilarity matrix. If not fixed, the algorithm chooses \( \alpha = 0.25 \) (cluster.effects=TRUE) or \( \alpha = 0.5 \) (cluster.effects=FALSE).
- **step**: The steps used in computing the dissimilarity matrix. Default is "both"=("shape" and "distance")
- **trace**: If TRUE, some informations are printed.

**Value**

The dissimilarity matrix of class "dist".

**Author(s)**

Gianluca Sottile <gianluca.sottile@unipa.it>

**References**


extract.object

See Also
clustEff, summary.clustEff, plot.clustEff, for summary and plotting. extract.object to extract useful objects for the clustering algorithm through a quantile regression coefficient modeling in a multivariate case.

Examples

### Using simulated data in all examples

# see the documentation for 'clustEff-package'

---

extract.object  extract.object fits a multivariate quantile regression and extracts objects for the cluster effects algorithm.

Description

extract.object fits a multivariate quantile regression and extracts objects for the cluster effects algorithm.

Usage

extract.object(Y, X, intercept=TRUE, formula.p=~slp(p, 3), s, object, p, which)

Arguments

Y  A multivariate response matrix of dimension n x q1, or a vector of length n.
X  The covariates matrix of dimension n x q2.
intercept  If TRUE, the intercept is included in the model.
formula.p  a one-sided formula of the form ~ b1(p, ...) + b2(p, ...) + ...
s  An optional 0/1 matrix that allows to exclude some model coefficients (see ‘Examples’).
object  An object of class “iqr”. If missing, Y and X have to be supplied.
p  The percentiles used in quantile regression coefficient modeling. If missing a default sequence is choosen.
which  If fixed, only the selected covariates are extracted from the model. If missing all the covariates are extracted.

Details

A list of objects useful to run the cluster effect algorithm is created.
plot.clustEff

Value

\( p \)

The percentiles used in the quantile regression.

\( X \)

A list containing as many matrices as covariates, where for each matrix the number of columns corresponds to the number of the responses. Each column of a matrix corresponds to one curve effect.

\( X_l \)

A list as \( X \). Each column of a matrix corresponds to the lower interval of the curve effect.

\( X_r \)

A list as \( X \). Each column of a matrix corresponds to the upper interval of the curve effect.

Author(s)

Gianluca Sottile <gianluca.sottile@unipa.it>

See Also

clustEff, for clustering algorithm; summary.clustEff and plot.clustEff, for summarizing and plotting clustEff objects.

Examples

```r
# using simulated data
# see the documentation for 'clustEff-package'
```

---

### plot.clustEff

**Plot Clustering Effects**

Description

Produces a dendrogram, a cluster plot and a boxplot of average distance cluster class "piqr".

Usage

```r
## S3 method for class 'clustEff'
plot(x, xvar=c("clusters", "dendrogram", "boxplot", "numclust"), which,
     add=FALSE, all=TRUE, polygon=TRUE, dissimilarity=TRUE, ...)
```

Arguments

- \( x \)
  An object of class "clustEdd", typically the result of a call to clustEff.

- \( xvar \)
  Clusters: plot of the k clusters; Dendrogram: plot of the tree after computing the dissimilarity measure and applying a hierarchical clustering algorithm; Boxplot: plot the average distance within clusters; Numclust: plot the curve to minimize to select the best number of clusters;
which If missing all curves effect are plotted.
add If TRUE and xvar="clusters", a unique plot of clusters is done.
all If TRUE and add=FALSE and xvar="clusters", a unique window of clusters is done.
polygon If TRUE confidence intervals are represented by shaded areas via polygon. Otherwise, dashed lines are used.
dissimilarity If TRUE dissimilarity measure within each cluster is used to do boxplot representation.
... additional graphical parameters, that can include xlim, ylim, xlab, ylab, col, lwd, lty. See par.

Details

Different plot for the clustering algorithm.

Author(s)
Gianluca Sottile <gianluca.sottile@unipa.it>

See Also
clustEff for cluster algorithm; extract.object for extracting information through a quantile regression coefficient modeling in a multivariate case; summary.clustEff for clustering summary.

Examples

```r
# using simulated data

# see the documentation for 'clustEff-package'
```

summary.clustEff Summary after Clustering Algorithm

Description

Summary of an object of class "clustEff".

Usage

```r
## S3 method for class 'clustEff'
summary(object, ...)
```

Arguments

- **object** An object of class "clustEff", the result of a call to clustEff.
- **...** for future methods.
Details

A summary of the clustering algorithm is printed.

Value

The following items is returned:

k  The number of selected clusters.
n  The number of observations.
p  The number of curves.
step  The selected step for computing the dissimilarity matrix.
alpha  The alpha-percentile used for computing the dissimilarity matrix.
method  The selected method to compute the hierarchical cluster analysis.
cut.method  The selected method to choose the best number of clusters.
tabClust  The table of clusters.
avClust  The average distance within clusters.
avSilhouette  Silhouette widths for clusters.
avDiss  The average dissimilarity measure within clusters.

Author(s)

Gianluca Sottile <gianluca.sottile@unipa.it>

See Also

clustEff, for cluster algorithm

Examples

# using simulated data

# see the documentation for 'clustEff-package'
Index

*Topic **clustering algorithm**
  clustEff, 5
  distshape, 8
*Topic **methods**
  plot.clustEff, 10
*Topic **models**
  clustEff, 5
  distshape, 8
*Topic **package**
  clustEff-package, 2

clustEff, 2, 5, 9–12
clustEff-package, 2

distshape, 8

extract.object, 2, 7, 9, 9, 11, 12

par, 11

plot.clustEff, 2, 7, 9, 10, 10

summary.clustEff, 2, 7, 9–11, 11