Package ‘RegClust’

February 19, 2015

Type Package

Title Cluster analysis via regression coefficients.

Version 1.0

Date 2014-02-17

Author Weichao Bao, Xin Tong, Meredith Ray, Hongmei Zhang

Maintainer Meredith Ray `<mere2110@yahoo.com>`

Depends R (>= 3.0.2)

Description This package clusters regression coefficients using the methods of clustering through linear regression models (CLM) (Qin and Self 2006). Maximum likelihood approach is used to infer the parameters for each cluster. Bayesian information criterion (BIC) combined with Bootstrapped maximum volume (BMV) criterion are used to determine the number of clusters.

License Artistic-2.0

NeedsCompilation no

Repository CRAN

Date/Publication 2014-02-24 18:22:43

R topics documented:

`RegClust-package` *A clustering process via regression.*

Description

A clustering process built upon linear regression analysis (Qin and Self 2006), maximum likelihood, Bayesian inference criteria, and Bootstrapping maximum volume.

Details

Package:	RegClust
Type:	Package
Version:	1.0
Date:	2014-02-17
License:	Artistic-2.0

Author(s)

Weichao Bao, Xin Tong, Meredith Ray, Hongmei Zhang

References

Qin, Li-Xuan, and Steven G. Self. "The clustering of regression models method with applications in gene expression data." Biometrics 62.2 (2006): 526-533.

`cluster.reg` *Clustering analysis of regression coefficients.*

Description

This package performs clustering on regression coefficients using the methods of clustering through linear regression models (CLM) (Qin and Self 2006). Maximum likelihood approach is used to infer the parameters for each cluster. Bayesian information criterion (BIC) combined with Bootstrapped maximum volume (BMV) criterion are used to determine the number of clusters. The Bootstrap method is used to estimate the uncertainty on the number of clusters.

Usage

```
cluster.reg(Y, X, loop = 1000)
```

Arguments

`Y`	An n x k data matrix for n number of observations and k dependent variables.
`X`	An n x m data matrix for n number of observations and m covariates of interest. Restriction of m < n.
`loop`	A numeric scalar identfying the number of iterations for the bootstrappng process. Default is 1000.

Value

`cluster`	A numeric vector of length k identifying which cluster the respective dependent variables belong to.
`parm`	Dataframe containing estimates of regression coefficients, σ^2, and π for each cluster, where σ^2 is the variance in the random error, and π is the probability that a variable is in a cluster.
`likelihood`	Likelihood at final iteration or at convergence.
`BIC`	Bayesian information criterion at final interation or at convergence.

Note

Although the number of covariates is unlimited, it is recommended to only allow up to 3 covariates to prevent potential difficulty in clustering the variables.

Author(s)

Weichao Bao, Xin Tong, Meredith Ray, Hongmei Zhang

References

Qin, Li-Xuan, and Steven G. Self. The clustering of regression models method with applications in gene expression data. Biometrics 62.2 (2006): 526-533.

Examples

```
## Not run:
beta0=1
beta1=0.02
beta2=-0.04
beta3=0.1
set.seed(1234)
sim=200
age=runif(sim, min=18, max=70)
Rerror=rnorm(sim,-3,3)

y1=matrix(NA,sim,4,dimnames = list(NULL,c(″y1″, ″y2″, ″y3″, ″y4″)))
for (g in 1:4){
set.seed(1234)
y1[,g]=beta0+beta1*runif(sim, min=18, max=70)+beta1*rnorm(sim,-3,3)
set.seed(1134+g)
y1[,g]=y1[,g]+rnorm(sim,0,1)
}
y2=matrix(NA,sim,5,dimnames = list(NULL,c(″y5″, ″y6″, ″y7″,″y8″,″y9″)))
for (g in 1:5){
set.seed(1234)
y2[,g]=beta0+beta2*runif(sim, min=18, max=70)+beta2*rnorm(sim,-3,3)
set.seed(2234+g)
y2[,g]=y2[,g]+rnorm(sim,0,0.5^0.5)
}
y3=matrix(NA,sim,6,dimnames = list(NULL,c(″y10″, ″y11″, ″y12″,″y13″,″y14″,″y15″)))
```

```
for (g in 1:6){
set.seed(1234)
y3[,g]=beta0+beta3*runif(sim, min=18, max=70)+beta3*rnorm(sim,-3,3)
set.seed(3334+g)
y3[,g]=y3[,g]+rnorm(sim,0,1)
}
X=data.frame(round(cbind(Rerror=Rerror,age=age),2))
Y=data.frame(round(cbind(y1,y2,y3),2))

run<-cluster.reg(Y,X)
run

## End(Not run)
```

Index