Package ‘mixedClust’

March 29, 2021

Type Package
Title Co-Clustering of Mixed Type Data
Version 1.0.2
Date 2021-03-21
Author Margot Selosse, Julien Jacques, Christophe Biernacki
Maintainer Margot Selosse <margot.selosse@gmail.com>
Description Implementation of the co-clustering method for mixed type data proposed in M. Selosse, J. Jacques, C. Biernacki (2018) <https://hal.archives-ouvertes.fr/hal-01893457>. It consists in clustering simultaneously the rows (observations) and the columns (features) of a heterogeneous data set.
License GPL (>= 2)
Imports Rcpp (>= 0.12.11), fda, methods
LinkingTo Rcpp, RcppProgress, RcppArmadillo
Suggests rmarkdown, ordinalClust, knitr
VignetteBuilder knitr
LazyData true
Depends R (>= 3.5.0)
SystemRequirements C++11
NeedsCompilation yes
Repository CRAN
Date/Publication 2021-03-29 14:00:02 UTC

R topics documented:

M1 ................................................................. 2
mixedCoclust ................................................... 2

Index 5
mixedCoclust

Matrix of simulated ordinal data

Description
This is a toy dataset for running simple examples.

Usage
M1

Format
A mixed type data matrix with 50 lines and 120 columns. There are 40 categorical variables, 40 continuous variables, and 40 ordinal variables.

mixedCoclust
Function to perform a co-clustering

Description
This function performs a co-clustering on heterogeneous data sets by using the Multiple Latent Block model (cf references for further details).

Usage
mixedCoclust(x=matrix(0,nrow=1,ncol=1), idx_list=c(1), distrib_names, kr, kc, init, nbSEM, nbSEMBurn, nbRepeat=1, nbIndmini, m=0, functionalData=array(0, c(1,1,1)), zrinit= 0 , zcinit=0, percentRandomB=0, percentRandomP=0)

Arguments

x
Data matrix, of dimension N*Jtot. The features with same type should be aside. The missing values should be coded as NA.

idx_list
Vector of length D. This argument is useful when variables are of different types. Element d should indicate where the variables of type d begins in matrix x.

distrib_names
Vector of length D. indicates the type of distribution to use. Must be among "Gaussian", "Multinomial", "BOS", "Poisson" or "Functional". Functional data must always be at the end.

kr
Number of row classes.

kc
Vector of length D. d^th element indicates the number of column clusters.

m
Vector of length D. d^th element defines the ordinal and categorical data's number of levels.
mixedCoclust

**functionalData**
Data tensor of dimension N*J*T.

**nbSEM**
Number of SEM-Gibbs iterations realized to estimate parameters.

**nbSEMburn**
Number of SEM-Gibbs burning iterations for estimating parameters. This parameter must be inferior to nbSEM.

**nbRepeat**
Number of times sampling on rows and on columns will be done at each SEM-Gibbs iteration.

**nbindmini**
Minimum number of cells belonging to a block.

**init**
String that indicates the kind of initialisation. Must be one of the following words: "kmeans", "random", "provided", "randomParams" or "randomBurnin".

**zrinit**
Vector of length N. When init="provided", indicates the labels of each row.

**zcinit**
Vector of length Jtot. When init="provided", indicates the labels of each column.

**percentRandomB**
Vector of length 2. Indicates the percentage of resampling when init is equal to "randomBurnin".

**percentRandomP**
Vector of length 2. Indicates the percentage of resampling when init is equal to "randomParams".

**Value**

@V
Matrix of dimension N*kr such that V[i,g]=1 if i belongs to cluster g.

@icl
ICL value for co-clustering.

@name

@paramschain
List of length nbSEMburn. For each iteration of the SEM-Gibbs algorithm, the parameters of the blocks are stored.

@pichain
List of length nbSEM. Item i is a vector of length kr which contains the row mixing proportions at iteration i.

@rhochain
List of length nbSEM. Item i is a list of length D whose d^th contains the column mixing proportions of groups of variables d, at iteration i.

@zc
List of length D. d^th item is a vector of length J[d] representing the columns partitions for the group of variables d.

@zr
Vector of length N with resulting row partitions.

@W
List of length D. Item d is a matrix of dimension J*kc[d] such that W[j,h]=1 if j belongs to cluster h.

@m
Vector of length D. d^th element represents the number of levels of d^th group of variables.

@params
List of length D. d^th item represents the blocks parameters for group of variables d.

@pi
Vector of length kr. Row mixing proportions.

@rho
List of length D. d^th item represents the column mixing proportion for d^th group of variables.

@xhat
List of length D. d^th item represents the d^th group of variables dataset, with missing values completed.

@zrchain
Matrix of dimension nbSEM*N. Row i represents the row cluster partitions at iteration i.

@zrchain
List of length D. Item d is a matrix of dimension nbSEM*J[d]. Row i represents the column cluster partitions at iteration i.
Author(s)
Margot Selosse, Julien Jacques, Christophe Biernacki.

Examples

data(M1)
nbSEM=30
nbSEMburn=20
nbbindmini=1
init = "random"

kr=2
kc=c(2,2,2)
m=c(6,3)
d.list <- c(1,41,81)
distributions <- c("Multinomial","Gaussian","Bos")
res <- mixedCoclust(x = M1, idx_list = d.list, distrib_names = distributions,
kr = kr, kc = kc, m = m, init = init, nbSEM = nbSEM,
nbSEMburn = nbSEMburn, nbbindmini = nbbindmini)
Index

* datasets
  M1, 2

M1, 2
mixedCoclust, 2