Package ‘BiclustBHMM’

June 23, 2023

Type Package

Title BiclustBHMM: A Bayesian model-based biclustering via Hidden Markov Models

Version 1.0

Date 2023-06-20

Author Thierry Chekouo

Maintainer Thierry Chekouo <tchekouo@umn.edu>

Description The BiclustBHMM package implements an MCMC algorithm for a Bayesian model-based biclustering that accounts for prior dependence between features while classifying them into under-expressed, over-expressed, and irrelevant features.

License GPL (>= 2.0)

Imports mc2d, truncnorm, invgamma

RoxygenNote 7.2.3

Encoding UTF-8

NeedsCompilation no

Depends R (>= 3.5.0)

R topics documented:

`BiclustBHMM`	*An MCMC algorithm to perform a Bayesian model-based biclustering using Hidden Markov Models*

Description

The method accounts for prior dependence between features, and performs a clustering of samples. Within each cluster, features are classified in 3 groups: over-expressed, under-expressed and irrelevant features. Hence, if K is the number of clusters, we have in total 3*K biclusters. The algorithm is implemented on the four methods: i) HMMBi-C: this method determines biclusters by assuming an order structure between features through a hidden Markov structure on the data, and imposing a constraint on the mean parameters μ_{kl} (positive, negative and zero depending on the feature state) in order to have a better biological interpretation of obtained clusters; ii) HMMBi-NoC: this method assumes an order structure between features through a hidden Markov structure on the data, but does not impose a constraint on the mean parameters μ_{kl}; iii) NoHMMBi-C: this method does not assume an order structure between features, but does impose a constraint on the mean parameters μ_{kl}; and iv) NoHMMBi-NoC: this method does not assume an order structure between features, and does not impose a constraint on the mean parameters μ_{kl}. The algorithm computes mainly (i) the posterior probabilities of inclusions for each subject to be in each cluster and ii) For each cluster, the posterior probabilities of inclusions for each feature to be in a group (or state).

Usage

```
BiclustBHMM(method="HMMBi-C",Data=Data,K=K,TruncValue=.2,Mu0=0,sigma20Mu=1000,
hyperErrVar=c(1,1),alphaP=rep(1,K),delt=rep(0.5,3),Nsample=3000, burnin=1000,seed=1)
```

Arguments

`method`	It's one of the four methods: "HMMBi-C","HMMBi-NoC","NoHMMBi-C" or "NoHMMBi-NoC".
`Data`	An expression data matrix of dimension $n \times p$ where n is the number of subjects, and p is the number of features. Features are ordered with respect to some criteria. For instance, we sorted features using similarities computed Gene Ontology.
`K`	It's the number of subject clusters.
`TruncValue`	It's the minimum value imposed on the mean parameters to identify either under-expressed or over-expressed features.
`Mu0`	Prior mean of μ_{kl}, parameter means of biclusters.
`sigma20Mu`	Prior variance of μ_{kl}, parameter means of biclusters.
`hyperErrVar`	Shape and rate parameters of the inverse prior distribution of both the error variance and the variance of μ_{kl}. They are denoted α_0 and β_0 in the manuscript.
`alphaP`	Scale parameters of the dirichlet prior distributions of proba. of inclusion of subjects. They are denoted α_1, ..., α_K in the manuscript.
`delt`	Scale parameters of the dirichlet prior distributions of transition prob. of features for each cluster. They are denoted δ_1, δ_3 and δ_3 in the manuscript.
`Nsample`	Total number of MCMC draws. It must be larger than burnin.
`burnin`	Number of draws to discard for burn-in.
`seed`	Set a seed number to generate distributions in the MCMC algorithm.

Details

The function will return several R objects, which can be assigned to a variable. To see the results, use the “$" operator.

Value

probZMean	Posterior probability of cluster membership for each sample/subject
probkappajk	Posterior probability of group membership for each feature within each sample cluster. It's an array of dimension $p \times K \times 3$.
TransMatMean	Overall (posterior) transition matrix between the three states (or groups) of featurues.
MuMean	Estimated posterior mean of cluster means. The last column is zero, and the first column should be means of over-expressed biclusters, the second cloumn is the mean of the under-expressed bicluster. The dimension is $K \times 3$..
logpostsample	Provide the log-posterior for each mcmc sample.
DIC	Provide the Deviance information criteria (DIC)

References

Thierry Chekouo and Himadri Mukherjee (2023), *A Bayesian Hierarchical Hidden Markov Model for clustering and gene selection: Application to Kidney cancer gene expression data, submitted.*

See Also

GenDataHMM

Examples

```
### We run Setting 1 in the manuscript
library(BiclustBHMM);
s2=1 # Variance of the error
K=2 # number of clusters
n=100 # number of subjects
p=1000 # number of features
Dat=GenDataHMM(OrderFeatured=TRUE,K=K,p=p,seed=1,n=n,SigmaOE=rep(s2,K),
SigmaUE=rep(s2,K),MeanOverExpre=c(2:(K+1))/2,MeanUnderExpre=-c(2:(K+1))/2)

dat=Dat$Y

Result=BiclustBHMM(method="HMMBi-C",Data=dat,K=K,TruncValue=.2,Mu0=0,
sigma20Mu=1000,hyperErrVar=c(1,1), alphaP=rep(1,K),delt=rep(0.5,3),Nsample=300,
burnin=100,seed=1)
Result=BiclustBHMM(method="NoHMMBi-NoC",Data=dat,K=K,TruncValue=.2,Mu0=0,
sigma20Mu=1000,hyperErrVar=c(1,1), alphaP=rep(1,K),delt=rep(0.5,3),Nsample=300,
burnin=100,seed=1)
str(Result)
F11=F1(TrueZ=Dat$Zk,TrueKappajk=Dat$kappajk,
probZmean=Result$probZMean,probkappajk=Result$probkappajk,EstK=K)

F11

## Not run:
# this is a long running example on the real data application
data(SimilarityBP)
HC=hclust(as.dist(1-SimilarityBP), method = "average", members = NULL)
OrderEntrez=HC$order
data(mRNAExpression)
dat=mRNAExpression[OrderEntrez,]
K=2
```

```
Result=BiclustBHMM(method="HMMBi-C",Data=dat,K=K,TruncValue=.2,Mu0=0,
sigma20Mu=1000,hyperErrVar=c(1,1), alphaP=rep(1,K),delt=rep(0.5,3),Nsample=300,
burnin=100,seed=1)

## End(Not run)
```

F1 *Compute criteria to evaluate biclustering performamce based on the F1-measure*

Description

This function computes F1-measure (F1) as defined in the manuscript.

Usage

```
F1(TrueZ, TrueKappajk,probZmean,probkappajk,EstK)
```

Arguments

TrueZ	True membership values of subjects/samples of dimension n, sample size. Possible values are $1, ..., K$.
TrueKappajk	Binary vector with 1 as positive and 0 as negative.
probZmean	(Posterior or Estimated) probability of cluster membership for each sample/subject.
probkappajk	(Posterior or Estimated) probability of group membership for each feature within each sample cluster. It's an array of dimension $p \times K \times 3$.
EstK	(Estimated) number of cluster

Details

The function returns F1 as compute in the manuscript

References

Thierry Chekouo and Himadri Mukherjee (2023), *A Bayesian Hierarchical Hidden Markov Model for clustering and gene selection: Application to Kidney cancer gene expression data*, *submitted*.

See Also

GenDataHMM BiclustBHMM

Examples

```
### We run Setting 1 in the manuscript
library(BiclustBHMM);
s2=1 # Variance of the error
K=2 # number of clusters
n=100 # number of subjects
p=1000 # number of features
Dat=GenDataHMM(OrderFeatured=TRUE,K=K,p=p,seed=1,n=n,SigmaOE=rep(s2,K),
SigmaUE=rep(s2,K),MeanOverExpre=c(2:(K+1))/2,MeanUnderExpre=-c(2:(K+1))/2)
```

```
dat=Dat$Y

Result=BiclustBHMM(method="HMMBi-C",Data=dat,K=K,TruncValue=.2,Mu0=0,
sigma20Mu=1000,hyperErrVar=c(1,1), alphaP=rep(1,K),delt=rep(0.5,3),Nsample=300,
burnin=100,seed=1)
str(Result)
F11=F1(TrueZ=Dat$Zk,TrueKappajk=Dat$kappajk,
             probZmean=Result$probZMean,probkappajk=Result$probkappajk,EstK=K)
F11
```

GenDataHMM — *Generation of simulated data as explained in the reference manuscript.*

Description

This function generates data described in the manuscript.

Usage

```
GenDataHMM(OrderFeatured=FALSE,K=K,MeanOverExpre=1:K,MeanUnderExpre=-c(1:K),p=p,seed=1,n=n,
                    SigmaOE=rep(1,K),SigmaUE=rep(1,K))
```

Arguments

`OrderFeatured`	If TRUE, then consecutive features belong to the same group. Otherwise, features in any group are chosen randomly without a predefined order.
`K`	Number of sample clusters
`MeanOverExpre`	(positive) Means of over-expressed biclusters.
`MeanUnderExpre`	(negative) Means of under-expressed biclusters..
`p`	Number of features.
`seed`	Seed to generate random numbers.
`n`	Number of subjects.
`SigmaOE`	Variance to generate over-expressed features.
`SigmaUE`	Variance to generate under-expressed feature.

Details

The function will generate data as explained in the manuscript. To see the results, use the “$" operator.

Value

`Y`	A expression matrix of dimension $n \times p$
`kappajk`	A matrix of feature membership of dimension $p \times K$. Values of the matrix are 1, 2 or 3 that correspond respectively to over-expressed, under-expressed and irrelevant features.
`TransMatMean`	Overall (posterior) transition matrix between the three states (or groups) of featurues.
`Zk`	A vector of length n of subject membership. Values are $1, ..., K$.

References

Thierry Chekouo and Himadri Mukherjee (2023), *A Bayesian Hierarchical Hidden Markov Model for clustering and gene selection: Application to Kidney cancer gene expression data, submitted.*

See Also

BiclustBHMM

Examples

```
library(BiclustBHMM);
s2=1 # Variance of the error
K=2 # number of clusters
n=100 # number of subjects
p=1000 # number of features
Dat=GenDataHMM(OrderFeatured=TRUE,K=K,p=p,seed=1,n=n,SigmaOE=rep(s2,K),
SigmaUE=rep(s2,K),MeanOverExpre=c(2:(K+1))/2,MeanUnderExpre=-c(2:(K+1))/2)

str(Dat)
```

---

mRNAExpression *mRNA expression data from TCGA - KIRC*

Description

mRNA expression of 1009 genes expressed under 534 samples

Usage

```
mRNAExpression
```

Format

A data frame

References

Thierry Chekouo and Himadri Mukherjee (2023), *A Bayesian Hierarchical Hidden Markov Model for clustering and gene selection: Application to Kidney cancer gene expression data, submitted.*

Examples

```
data("mRNAExpression")
```

`SimilarityBP` *Similarities between genes*

Description

Similarities between genes computed using the Biological Process in GO

Usage

```
SimilarityBP
```

Format

A data frame

References

Thierry Chekouo and Himadri Mukherjee (2023), *A Bayesian Hierarchical Hidden Markov Model for clustering and gene selection: Application to Kidney cancer gene expression data*, *submitted*.

Examples

```
data("SimilarityBP")
```

Index