Package 'FisherEM'

January 20, 2025

Type Package

Title The FisherEM Algorithm to Simultaneously Cluster and Visualize High-Dimensional Data

Version 1.6

Date 2020-09-28

Encoding UTF-8

Author Charles Bouveyron, Camille Brunet & Nicolas Jouvin.

Maintainer Charles Bouveyron <charles.bouveyron@gmail.com>

Depends MASS, parallel, elasticnet, ggplot2

Imports ellipse, plyr

Description The FisherEM algorithm, proposed by Bouveyron & Brunet (2012) <doi:10.1007/s11222-011-9249-9>, is an efficient method for the clustering of high-dimensional data. FisherEM models and clusters the data in a discriminative and low-dimensional latent subspace. It also provides a low-dimensional representation of the clustered data. A sparse version of Fisher-EM algorithm is also provided.

License GPL-2

LazyLoad yes

NeedsCompilation no

RoxygenNote 7.1.1

Suggests testthat, aricode

Repository CRAN

Date/Publication 2020-09-28 14:10:02 UTC

Contents

`FisherEM-package` *The FisherEM Algorithm to Simultaneously Cluster and Visualize High-Dimensional Data*

Description

The FisherEM algorithm, proposed by Bouveyron & Brunet (201) <doi:10.1007/s11222-011-9249-9>, is an efficient method for the clustering of high-dimensional data. FisherEM models and clusters the data in a discriminative and low-dimensional latent subspace. It also provides a low-dimensional representation of the clustered data. A sparse version of Fisher-EM algorithm is also provided.

Details

Package:	FisherEM
Type:	Package
Version:	1.2
Date:	2012-07-09
License:	GPL-2
LazyLoad:	yes

Author(s)

Charles Bouveyron, Camille Brunet & Nicolas Jouvin.

Maintainer: Charles Bouveyron <charles.bouveyron@gmail.com>

References

Charles Bouveyron, Camille Brunet (2012), "Simultaneous model-based clustering and visualization in the Fisher discriminative subspace.", Statistics and Computing, 22(1), 301-324 <doi:10.1007/s11222-011-9249-9>.

Charles Bouveyron and Camille Brunet (2014), "Discriminative variable selection for clustering with the sparse Fisher-EM algorithm", Computational Statistics, vol. 29(3-4), pp. 489-513 <10.1007/s00180-013-0433-6>.

bfem *The Bayesian Fisher-EM algorithm.*

Description

The Bayesian Fisher-EM algorithm is built on a Bayesian formulation of the model used in the fem. It is a subspace clustering method for high-dimensional data. It is based on a Gaussian Mixture Model and on the idea that the data lives in a common and low dimensional subspace. A VEM-like algorithm estimates both the discriminative subspace and the parameters of the mixture model.

Usage

```
bfem(
  Y,
  K = 2:6,
  model = "AkjBk",
  method = "gs",
  crit = "icl",
  maxit.em = 100,
  eps.em = 1e-06,
  maxit.ve = 3,
  eps.ve = 1e-04,
  lambda = 1000,
  emp.bayes = T,
  init = "kmeans",
  nstart = 10,
  Tinit = c(),
  kernel = "",
  disp = FALSE,
  mc.cores = (detectCores() - 1),
  subset = NULL
)
```

Arguments

Y	The data matrix. Categorical variables and missing values are not allowed.
K	An integer vector specifying the numbers of mixture components (clusters) among which the model selection criterion will choose the most appropriate number of groups. Default is 2:6.
model	A vector of Bayesian discriminative latent mixture (BDLM) models to fit. There are 12 different models: "DkBk", "DkB", "DBk", "DB", "AkjBk", "AkjB", "AkBk", "AkBk", "AjBk", "AjB", "ABk", "AB". The option "all" executes the Fisher-EM algorithm on the 12 DLM models and select the best model according to the maximum value obtained by model selection criterion. Similar to fem

method	The method used for the fitting of the projection matrix associated to the discriminative subspace. Three methods are available: 'gs' (Gram-Schmidt, the original proposition), 'svd' (based on SVD, faster) and 'reg' (the Fisher criterion is rewritten as a regression problem). The 'gs' method is the default method.
crit	The model selection criterion to use for selecting the most appropriate model for the data. There are 3 possibilities: "bic", "aic" or "icl". Default is "icl".
maxit.em	The maximum number of iterations before the stop of the main EM loop in the BFEM algorithm.
eps.em	The threshold value for the likelihood differences (Aitken's criterion) to stop the BFEM algorithm.
maxit.ve	The maximum number of iterations before the stop of the VE-step loop (fixed point algorithm)
eps.ve	The threshold value for the likelihood differences (Aitken's criterion) to stop the BFEM algorithm.
lambda	The initial value for the variance of the Gaussian prior on the means in the latent space.
emp.bayes	Should the hyper-parameters (mean and variance) of the prior be updated ? Default to TRUE.
init	The initialization method for the Fisher-EM algorithm. There are 4 options: "random" for a randomized initialization, "kmeans" for an initialization by the kmeans algorithm, "hclust" for hierarchical clustering initialization or "user" for a specific initialization through the parameter "Tinit". Default is "kmeans". Notice that for "kmeans" and "random", several initializations are asked and the initialization associated with the highest likelihood is kept (see "nstart").
nstart	The number of restart if the initialization is "kmeans" or "random". In such a case, the initialization associated with the highest likelihood is kept.
Tinit	A n x K matrix which contains posterior probabilities for initializing the algorithm (each line corresponds to an individual).
kernel	It enables to deal with the n < p problem. By default, no kernel (" ") is used. But the user has the choice between 3 options for the kernel: "linear", "sigmoid" or "rbf".
disp	If true, some messages are printed during the clustering. Default is false.
mc.cores	The number of CPUs to use to fit in parallel the different models (only for non-Windows platforms). Default is the number of available cores minus 1.
subset	A positive integer defining the size of the subsample, default is NULL. In case of large data sets, it might be useful to fit a FisherEM model on a subsample of the data, and then use this model to predict cluster assignments for the whole data set. Notice that in, such a case, likelihood values and model selection criteria are computed for the subsample and not the whole data set.

Value

A list is returned:

- K - The number of groups.

- cls - the group membership of each individual estimated by the BFEM algorithm
- Tinit - The initial posterior probalities used to start the algorithm
- d - the dimension of the discriminative subspace
- elbos - A vector containing the evolution of the variational lower bound at each iteration
- loglik - The final value of the variational lower bound
- n_ite - The number of iteration until convergence of the BFEM algorithm
- P - the posterior probabilities of each individual for each group
- U - The loading matrix which determines the orientation of the discriminative subspace
- param - A list containing the estimated parameters of the model
  - PI - The mixture proportions
  - Sigmak - An array containing estimated cluster covariances in the latent space
  - Beta - The noise variance in each cluster
- var_param - A list containing the variational distribution parameters
  - logtau - A n x K matrix containing the logarithm of the multinomial parameters of q(Z)
  - Varmeank - A K x d matrix containing the variational mean
  - Varcovk - A d x d x K array containing the variational covariance matrices.
- proj - The projected data on the discriminative subspace.
- aic - The value of the Akaike information criterion
- bic - The value of the Bayesian information criterion
- icl - The value of the integrated completed likelihood criterion
- method - The method used in the F-step
- call - The call of the function
- crit - The model selection criterion used

See Also

fem

Examples

```
# Chang's 1983 setting
simu = simu_bfem(300, which = "Chang1983")
Y = simu$Y
res.bfem = bfem(Y, K = 2:6, model=c('AB'), init = 'kmeans', nstart = 1,
               maxit.em = 10, eps.em = 1e-3, maxit.ve = 3, mc.cores = 2)
```

`fem` *The Fisher-EM algorithm*

Description

The Fisher-EM algorithm is a subspace clustering method for high-dimensional data. It is based on the Gaussian Mixture Model and on the idea that the data lives in a common and low dimensional subspace. An EM-like algorithm estimates both the discriminative subspace and the parameters of the mixture model.

Usage

```
fem(Y,K=2:6,model='AkjBk',method='gs',crit='icl',maxit=50,eps=1e-4,init='kmeans',
            nstart=5,Tinit=c(),kernel='',disp=FALSE,mc.cores=(detectCores()-1),
                subset=NULL)
```

Arguments

`Y`	The data matrix. Categorical variables and missing values are not allowed.
`K`	An integer vector specifying the numbers of mixture components (clusters) among which the model selection criterion will choose the most appropriate number of groups. Default is 2:6.
`model`	A vector of discriminative latent mixture (DLM) models to fit. There are 12 different models: "DkBk", "DkB", "DBk", "DB", "AkjBk", "AkjB", "AkBk", "AkBk", "AjBk", "AjB", "ABk", "AB". The option "all" executes the Fisher-EM algorithm on the 12 DLM models and select the best model according to the maximum value obtained by model selection criterion.
`method`	The method used for the fitting of the projection matrix associated to the discriminative subspace. Three methods are available: 'gs' (Gram-Schmidt, the original proposition), 'svd' (based on SVD, fastest approach, it should be preferred on large data sets) and 'reg' (the Fisher criterion is rewritten as a regression problem). The 'gs' method is the default method since it is the most efficient one on most data sets.
`crit`	The model selection criterion to use for selecting the most appropriate model for the data. There are 3 possibilities: "bic", "aic" or "icl". Default is "icl".
`maxit`	The maximum number of iterations before the stop of the Fisher-EM algorithm.
`eps`	The threshold value for the likelihood differences to stop the Fisher-EM algorithm.
`init`	The initialization method for the Fisher-EM algorithm. There are 4 options: "random" for a randomized initialization, "kmeans" for an initialization by the kmeans algorithm, "hclust" for hierarchical clustering initialization or "user" for a specific initialization through the parameter "Tinit". Default is "kmeans". Notice that for "kmeans" and "random", several initializations are asked and the initialization associated with the highest likelihood is kept (see "nstart").

nstart	The number of restart if the initialization is "kmeans" or "random". In such a case, the initialization associated with the highest likelihood is kept.
Tinit	A n x K matrix which contains posterior probabilities for initializing the algorithm (each line corresponds to an individual).
kernel	It enables to deal with the n < p problem. By default, no kernel (" ") is used. But the user has the choice between 3 options for the kernel: "linear", "sigmoid" or "rbf".
disp	If true, some messages are printed during the clustering. Default is false.
mc.cores	The number of CPUs to use to fit in parallel the different models (only for non-Windows platforms). Default is the number of available cores minus 1.
subset	A positive integer defining the size of the subsample, default is NULL. In case of large data sets, it might be useful to fit a FisherEM model on a subsample of the data, and then use this model to predict cluster assignments for the whole data set. Notice that in, such a case, likelihood values and model selection criteria are computed for the subsample and not the whole data set.

Value

A list is returned:

K	The number of groups.
cls	the group membership of each individual estimated by the Fisher-EM algorithm.
P	the posterior probabilities of each individual for each group.
U	The loading matrix which determines the orientation of the discriminative subspace.
mean	The estimated mean in the subspace.
my	The estimated mean in the observation space.
prop	The estimated mixture proportion.
D	The covariance matrices in the subspace.
aic	The value of the Akaike information criterion.
bic	The value of the Bayesian information criterion.
icl	The value of the integrated completed likelihood criterion.
loglik	The log-likelihood values computed at each iteration of the FEM algorithm.
ll	the log-likelihood value obtained at the last iteration of the FEM algorithm.
method	The method used.
call	The call of the function.
plot	Some information to pass to the plot.fem function.
crit	The model selction criterion used.

Author(s)

Charles Bouveyron, Camille Brunet & Nicolas Jouvin.

References

Charles Bouveyron and Camille Brunet (2012), Simultaneous model-based clustering and visualization in the Fisher discriminative subspace, Statistics and Computing, 22(1), 301-324 <doi:10.1007/s11222-011-9249-9>.

Charles Bouveyron and Camille Brunet (2014), "Discriminative variable selection for clustering with the sparse Fisher-EM algorithm", Computational Statistics, vol. 29(3-4), pp. 489-513 <10.1007/s00180-013-0433-6>.

See Also

sfem, plot.fem, fem.ari, summary.fem

Examples

```
data(iris)
res = fem(iris[,-5],K=3,model='AkBk',method='gs')
res
plot(res)
fem.ari(res,as.numeric(iris$Species))
table(iris$Species,res$cls)


# Fit several models and numbers of groups (use by default on non-Windows
# platforms the parallel computing).
res = fem(iris[,-5],K=2:6,model='all',method='gs', mc.cores=2)
res
plot(res)
fem.ari(res,as.numeric(iris$Species))
table(iris$Species,res$cls)
```

`fem.ari` *Adjusted Rand index*

Description

The function computes the adjusted Rand index (ARI) which allows to compare two clustering partitions.

Usage

```
fem.ari(x,y)
```

Arguments

x	A 'fem' object containing the first partition to compare.
y	The second partition to compare (as vector).

Value

ari The value of the ARI.

See Also

fem, sfem, plot.fem, summary.fem

Examples

```
data(iris)
res = fem(iris[,-5],K=3,model='DkBk',method='reg')
res
plot(res)
fem.ari(res,as.numeric(iris[,5]))
```

---

plot.bfem *Plotting function*

Description

Utility function to plot the results of the BFEM algorithm. The S3 plot function is a wrapper function over the 3 other functions

Usage

```
## S3 method for class 'bfem'
plot(x, type = "subspace", ...)

plot_subspace(
  x,
  alpha_levels = c(0.95),
  plot.dims = c(1, 2),
  show.ellipses = T,
  show.uncertainty = T,
  size = 2,
  cex.uncertainty = 1,
  ...
)

plot_bound(x, ...)

plot_crit(x, crit = NULL, ...)
```

Arguments

x	The results of bfem.
type	The plot type: • "subspace" (default) - Uses plot_subspace() to plot the projected data • "criterion" - Uses plot_crit() to plot the criterion value. • "elbo" - Uses plot_bound() to plot the variational lower bound evolution.
...	Additional parameter to pass to corxponding functions:
alpha_levels	A vector giving the desired Gaussian ellipses level set. Default to 0.95.
plot.dims	The dimension to be plotted. Default to the first two dimensions.
show.ellipses	Should Gaussian ellipses be plotted. Default to TRUE
show.uncertainty	Should uncertainty be plotted. A point is considered uncertain if its posterior probability of membership is peaked toward 2 or more clusters. Graphically, it can be displayed with a bigger point size depending on the uncertainty level, bigger points being more uncertain.
size	The point size.
cex.uncertainty	The multiplicative factor for the basic point size controlling the size of uncertain points.
crit	Used to specify which criterion should be plotted. Possible values are "aic", "bic" and 'icl. The default is the criterion used in the algorithm.

Value

a ggplot2 plot object

Functions

- plot_subspace: Plot Y projected on the ‘plot.dims‘ dimensions of the latent space
- plot_bound: plot the variational bound evolution
- plot_crit: Plot the criterion xult

Examples

```
data(iris)
Y = iris[,-5]
res = bfem(Y, 3, model = 'DB')
gg = plot(x=res, type = "subspace")
print(gg)
```

plot.fem *The plot function for 'fem' objects.*

Description

This function plots different information about 'fem' objects such as model selection, log-likelihood evolution and visualization of the clustered data into the discriminative subspace fitted by the Fisher-EM algorithm.

Usage

```
 ## S3 method for class 'fem'
plot(x, frame=0, crit=c(),...)
```

Arguments

x	The fem object.
frame	0: all plots; 1: selection of the number of groups; 2: log-likelihood; projection of the data into the discriminative subspace.
crit	The model selection criterion to display. Default is the criterion used in the 'fem' function ('icl' by default).
...	Additional options to pass to the plot function.

See Also

fem, sfem, fem.ari, summary.fem

Examples

```
data(iris)
res = fem(iris[,-5],K=3,model='DkBk',method='reg')
res
plot(res)
fem.ari(res,as.numeric(iris[,5]))
```

print.fem *The print function for 'fem' objects.*

Description

This function summarizes 'fem' objects. It in particular indicates which DLM model has been chosen and displays the loading matrix 'U' if the original dimension is smaller than 10.

Usage

```
 ## S3 method for class 'fem'
print(x,...)
```

Arguments

x	The fem object.
...	Additional options to pass to the summary function.

See Also

fem, sfem, fem.ari, plot.fem

Examples

```
data(iris)
res = fem(iris[,-5],K=3,model='DkBk',method='reg')
res
plot(res)
fem.ari(res,as.numeric(iris[,5]))
```

---

sfem *The sparse Fisher-EM algorithm*

Description

The sparse Fisher-EM algorithm is a sparse version of the Fisher-EM algorithm. The sparsity is introduced within the F step which estimates the discriminative subspace. The sparsity on U is obtained by adding a l1 penalty to the optimization problem of the F step.

Usage

```
sfem(Y,K=2:6,obj=NULL,model='AkjBk',method='reg',crit='icl',maxit=50,eps=1e-6,
    init='kmeans',nstart=5,Tinit=c(),kernel='',disp=FALSE,l1=0.1,l2=0,nbit=2)
```

Arguments

Y	The data matrix. Categorical variables and missing values are not allowed.
K	An integer vector specifying the numbers of mixture components (clusters) among which the model selection criterion will choose the most appropriate number of groups. Default is 2:6.
obj	An object of class 'fem' previously learned with the 'fem' function which will be used as initialization of the sparse FisherEM algorithm.
model	A vector of discriminative latent mixture (DLM) models to fit. There are 12 different models: "DkBk", "DkB", "DBk", "DB", "AkjBk", "AkjB", "AkBk", "AkBk", "AjBk", "AjB", "ABk", "AB". The option "all" executes the Fisher-EM algorithm on the 12 DLM models and select the best model according to the maximum value obtained by model selection criterion.

method	The method use for the fitting of the projection matrix associated to the discriminative subspace. Three methods are available: 'svd', 'reg' and 'gs'. The 'reg' method is the default.
crit	The model selection criterion to use for selecting the most appropriate model for the data. There are 3 possibilities: "bic", "aic" or "icl". Default is "icl".
maxit	The maximum number of iterations before the stop of the Fisher-EM algorithm.
eps	The threshold value for the likelihood differences to stop the Fisher-EM algorithm.
init	The initialization method for the Fisher-EM algorithm. There are 4 options: "random" for a randomized initialization, "kmeans" for an initialization by the kmeans algorithm, "hclust" for hierarchical clustering initialization or "user" for a specific initialization through the parameter "Tinit". Default is "kmeans". Notice that for "kmeans" and "random", several initializations are asked and the initialization associated with the highest likelihood is kept (see "nstart").
nstart	The number of restart if the initialization is "kmeans" or "random". In such a case, the initialization associated with the highest likelihood is kept.
Tinit	A n x K matrix which contains posterior probabilities for initializing the algorithm (each line corresponds to an individual).
kernel	It enables to deal with the n < p problem. By default, no kernel (" ") is used. But the user has the choice between 3 options for the kernel: "linear", "sigmoid" or "rbf".
disp	If true, some messages are printed during the clustering. Default is false.
l1	The l1 penalty value (lasso) which has to be in [0,1]. A small value (close to 0) leads to a very sparse loading matrix whereas a value equals to 1 corresponds to no sparsity. Default is 0.1.
l2	The l2 penalty value (elasticnet). Defaults is 0 (no regularization).
nbit	The number of iterations for the lasso procedure. Defaults is 2.

Value

A list is returned:

K	The number of groups.
cls	the group membership of each individual estimated by the Fisher-EM algorithm.
P	the posterior probabilities of each individual for each group.
U	The loading matrix which determines the orientation of the discriminative subspace.
mean	The estimated mean in the subspace.
my	The estimated mean in the observation space.
prop	The estimated mixture proportion.
D	The covariance matrices in the subspace.
aic	The value of the Akaike information criterion.
bic	The value of the Bayesian information criterion.

icl	The value of the integrated completed likelihood criterion.
loglik	The log-likelihood values computed at each iteration of the FEM algorithm.
ll	the log-likelihood value obtained at the last iteration of the FEM algorithm.
method	The method used.
call	The call of the function.
plot	Some information to pass to the plot.fem function.
crit	The model selction criterion used.
l1	The l1 value.
l2	The l2 value.

Author(s)

Charles Bouveyron and Camille Brunet

References

Charles Bouveyron and Camille Brunet (2012), Simultaneous model-based clustering and visualization in the Fisher discriminative subspace, Statistics and Computing, 22(1), 301-324 <doi:10.1007/s11222-011-9249-9>.

Charles Bouveyron and Camille Brunet (2014), "Discriminative variable selection for clustering with the sparse Fisher-EM algorithm", Computational Statistics, vol. 29(3-4), pp. 489-513 <10.1007/s00180-013-0433-6>.

See Also

fem, plot.fem, fem.ari, summary.fem

Examples

```
data(iris)
res = sfem(iris[,-5],K=3,model='DkBk',l1=seq(.01,.3,.05))
res
plot(res)
fem.ari(res,as.numeric(iris[,5]))
```

---

simu_bfem *Experimental setting of the chapter BFEM*

Description

Experimental setting of the chapter BFEM

Usage

```
simu_bfem(n, which = "Chang1983", ...)
```

Arguments

n	Number of observations
which	Type of simulation, either: • "Chang1983" - Simulate the dataset of Chang's (1983) paper : a mixture of 2 Gaussian with in dimension p=15. • "section4.2" - Experimental setting of Section 4.2: DLM model in dimension p with d=2 and K=3, with noisy dimensions. • "section4.3" - Experimental setting of Section 4.3: Same as ‘"section4.2"‘ except the noise is expressed in term of signal-to-noise ration (decibels).
...	Additional param controlling the simulation • p - The desired observed space dimension, the latent dimension is kept fixed to d=2 and noisy Gaussian dimensions are added (useless for ‘"Chang1983"‘) • noise (for ‘"section4.2"‘ only) - Variance of the noise • snr (for ‘"section4.3"‘ only) - Signal-to-noise ratio (in decibels) representing the ratio of signal and noise variances in logarithmic scale. The greater snr, the smaller noise variance.

Value

A list with slots

- Y - The simulated data.
- cls - The true clustering.

Examples

```
n = 300

# Chang's 1983 setting
simu = simu_bfem(n = n, which = "Chang1983")

# Section 4.2 setting
p = 25
noise = 1
simu = simu_bfem(n, which = "section4.2", p = p, noise = noise)

# Section4.3 setting
snr = 3 # noise variance is 2 times smaller than that of the signal.
simu = simu_bfem(n, which = "section4.3", snr = 10)
```

Index