Package 'EMMIXSSL'

February 8, 2022

Type Package

Title Semi-Supervised Learning via Gaussian Mixture Model

Version 1.1.0

Author Ziyang Lyu, Daniel Ahfock, Geoffrey J. McLachlan

Maintainer Ziyang Lyu <ziyang.lyu@unsw.edu.au>

Description The algorithm of semi-supervised learning for a partially classified sample via Gaussian mixture model with the missing-label mechanism is designed for a fitting g-component Gaussian mixture model via maximum likelihood (ML). The classifier is proposed to treat the labels of the unclassified features as missing-data and to introduce a framework for their missing as in the pioneering work of Rubin (1976) for missing in incomplete data analysis. It suggests that the missingness of the labels of the features can be modelled by representing the probability of a missing-label for a feature via the logistic model depending on the entropy of the feature or an appropriate proxy for it.

Depends R (>= 3.1.0), mvtnorm,stats

License GPL-3

Encoding UTF-8

RoxygenNote 7.1.2

NeedsCompilation no

Repository CRAN

Date/Publication 2022-02-08 11:00:02 UTC

R topics documented:

cov2vec 2
discriminant_beta 3
EMMIXSSL 3
get_clusterprobs 5
get_entropy 6
initialvalue 7
list2par 8
loglk_full 9
loglk_ig 10

loglk_miss 11
logsumexp 12
makelabelmatrix 12
neg_objective_function 13
normalise_logprob 13
par2list 14
pro2vec 14
rlabel 15
rmix 16
vec2cov 17
vec2pro 17

Index **18**

`cov2vec` *Transform a variance matrix into a vector*

Description

Transform a variance matrix into a vector i.e., Sigma=R^T*R

Usage

```
cov2vec(sigma)
```

Arguments

`sigma`	A variance matrix

Details

The variance matrix is decomposed by computing the Choleski factorization of a real symmetric positive-definite square matrix. Then, storing the upper triangular factor of the Choleski decomposition into a vector.

Value

par A vector representing a variance matrix discriminant_beta *Discriminant function*

Description

Discriminant function in the particular case of g=2 groups under the equal covariance consideration

Usage

```
discriminant_beta(pi, mu, sigma)
```

Arguments

pi	A 2-dimensional initial vector of the mixing proportions.
mu	A initial $p \times 2$ matrix of the location parameters.
sigma	A $p \times p$ common covariance matrix

Details

Discriminant function in the particular case of g=2 groups under the equal covariance consideration can be expressed

$$d(y_i, \beta) = \beta_0 + \beta_1 y_i,$$

where $\beta_0 = \log \frac{\pi_1}{\pi_2} - \frac{1}{2} \frac{\mu_1^2 - \mu_2^2}{\sigma^2}$ and $\beta_1 = \frac{\mu_1 - \mu_2}{\sigma^2}$.

Value

beta0	An intercept of discriminant function
beta	A coefficient of discriminant function

EMMIXSSL *Fitting Gaussian mixture model to the incompleted dataset with missing-data mechanism*

Description

Fit normal distribution to the classified data and fit a Gaussian mixture model to the unclassified data based on the missing-data mechanism

Usage

```
EMMIXSSL(
  dat,
  zm,
  pi,
  mu,
  sigma,
  ncov,
  xi = NULL,
  type,
  iter.max = 500,
  eval.max = 500,
  rel.tol = 1e-15,
  sing.tol = 1e-20
)
```

Arguments

`dat`	An $n \times p$ matrix where each row represents an individual observation
`zm`	An n-dimensional vector of group partition including the missing-label, denoted as NA.
`pi`	A g-dimensional initial vector of the mixing proportions.
`mu`	A initial $p \times g$ matrix of the location parameters.
`sigma`	A $p \times p$ covariance matrix if `ncov=1`, or a list of g covariance matrices with dimension $p \times p \times g$ if `ncov=2`.
`ncov`	Options of structure of sigma matrix; the default value is 2; `ncov` = 1 for a common covariance matrix; `ncov` = 2 for the unequal covariance/scale matrices.#'
`xi`	A 2-dimensional initial coefficient vector for a logistic function of the Shannon entropy.
`type`	Two types to fit to the model, 'ign' indicates fitting the model on the basis of the missing-label mechanism ignored, and 'full' indicates fitting the model on the basis of the missing-label mechanism
`iter.max`	Maximum number of iterations allowed. Defaults to 500
`eval.max`	Maximum number of evaluations of the objective function allowed. Defaults to 500
`rel.tol`	Relative tolerance. Defaults to 1e-15
`sing.tol`	Singular convergence tolerance; defaults to 1e-20.

Value

`objective`	Value of objective likelihood
`convergence`	Value of convergence
`iteration`	Number of iteration
`pi`	Estimated vector of the mixing proportions.

mu	Estimated matrix of the location parameters.
sigma	Estimated covariance matrix
xi	Estimated coefficient vector for a logistic function of the Shannon entropy

Examples

```
n<-150
pi<-c(0.25,0.25,0.25,0.25)
sigma<-array(0,dim=c(3,3,4))
sigma[,,1]<-diag(1,3)
sigma[,,2]<-diag(2,3)
sigma[,,3]<-diag(3,3)
sigma[,,4]<-diag(4,3)
mu<-matrix(c(0.2,0.3,0.4,0.2,0.7,0.6,0.1,0.7,1.6,0.2,1.7,0.6),3,4)
dat<-rmix(n=n,pi=pi,mu=mu,sigma=sigma,ncov=2)
xi<-c(-0.5,1)
m<-rlabel(dat=dat$Y,pi=pi,mu=mu,sigma=sigma,xi=xi,ncov=2)
zm<-dat$clust
zm[m==1]<-NA
inits<-initialvalue(g=4,zm=zm,dat=dat$Y,ncov=2)
## Not run:
fit_pc<-EMMIXSSL(dat=dat$Y,zm=zm,pi=inits$pi,mu=inits$mu,sigma=inits$sigma,xi=xi,type='full',ncov=2)

## End(Not run)
```

---

get_clusterprobs *Posterior probability*

Description

Get the posterior probability for each cluster

Usage

```
get_clusterprobs(dat, n, p, g, pi, mu, sigma, ncov = 2)
```

Arguments

dat	An $n \times p$ matrix where each row represents an individual observation
n	Number of observations.
p	Dimension of observation vecor.
g	Number of multivariate Gaussian groups.
pi	A g-dimensional initial vector of the mixing proportions.
mu	A initial $p \times g$ matrix of the location parameters.
sigma	A $p \times p$ covariance matrix if ncov=1, or a list of g covariance matrices with dimension $p \times p \times g$ if ncov=2.

ncov — Options of structure of sigma matrix; the default value is 2; `ncov = 1` for a common covariance matrix that `sigma` is a $p \times p$ matrix. `ncov = 2` for the unequal covariance/scale matrices that `sigma` represents a list of g matrices with dimension $p \times p \times g$.

Details

The posterior probability can be expressed as

$$\tau_i(y_j;\theta) = Prob\{z_{ij} = 1|y_j\} = \frac{\pi_i \phi(y_j;\mu_i,\Sigma_i)}{\sum_{h=1}^{g} \pi_h \phi(y_j;\mu_h,\Sigma_h)},$$

where ϕ is a normal probability density function, and z_{ij} is a zero-one indicator variable defining the known group of origin of each.

Value

clusprobs — The posterior probabilities of the i-th entity that belongs to the j-th group.

Examples

```
n<-150
pi<-c(0.25,0.25,0.25,0.25)
sigma<-array(0,dim=c(3,3,4))
sigma[,,1]<-diag(1,3)
sigma[,,2]<-diag(2,3)
sigma[,,3]<-diag(3,3)
sigma[,,4]<-diag(4,3)
mu<-matrix(c(0.2,0.3,0.4,0.2,0.7,0.6,0.1,0.7,1.6,0.2,1.7,0.6),3,4)
dat<-rmix(n=n,pi=pi,mu=mu,sigma=sigma,ncov=2)
tau<-get_clusterprobs(dat=dat$Y,n=150,p=3,g=4,mu=mu,sigma=sigma,pi=pi,ncov=2)
```

---

get_entropy — *Shannon entropy*

Description

Shannon entropy

Usage

```
get_entropy(dat, n, p, g, pi, mu, sigma, ncov = 2)
```

Arguments

dat — An $n \times p$ matrix where each row represents an individual observation

n — Number of observations.

p — Dimension of observation vecor.

`g`	Number of multivariate Gaussian groups.
`pi`	A g-dimensional initial vector of the mixing proportions.
`mu`	A initial $p \times g$ matrix of the location parameters.
`sigma`	A $p \times p$ covariance matrix if `ncov=1`, or a list of g covariance matrices with dimension $p \times p \times g$ if `ncov=2`.
`ncov`	Options of structure of sigma matrix; the default value is 2; `ncov = 1` for a common covariance matrix that `sigma` is a $p \times p$ matrix. `ncov = 2` for the unequal covariance/scale matrices that `sigma` represents a list of g matrices with dimension $p \times p \times g$.

Details

The concept of information entropy was introduced by *shannon1948mathematical*. The entropy of y_j is formally defined as

$$e_j(y_j; \theta) = -\sum_{i=1}^{g} \tau_i(y_j; \theta) \log \tau_i(y_j; \theta).$$

Value

`clusprobs`	The posterior probabilities of the i-th entity that belongs to the j-th group.

Examples

```
n<-150
pi<-c(0.25,0.25,0.25,0.25)
sigma<-array(0,dim=c(3,3,4))
sigma[,,1]<-diag(1,3)
sigma[,,2]<-diag(2,3)
sigma[,,3]<-diag(3,3)
sigma[,,4]<-diag(4,3)
mu<-matrix(c(0.2,0.3,0.4,0.2,0.7,0.6,0.1,0.7,1.6,0.2,1.7,0.6),3,4)
dat<-rmix(n=n,pi=pi,mu=mu,sigma=sigma,ncov=2)
en<-get_entropy(dat=dat$Y,n=150,p=3,g=4,mu=mu,sigma=sigma,pi=pi,ncov=2)
```

`initialvalue` *Initial values for ECM*

Description

Inittial values for claculating the estimates based on solely on the classified features.

Usage

```
initialvalue(dat, zm, g, ncov = 2)
```

Arguments

- `dat` An $n \times p$ matrix where each row represents an individual observation
- `zm` An n-dimensional vector of group partition including the missing-label, denoted as NA.
- `g` Number of multivariate Gaussian groups.
- `ncov` Options of structure of sigma matrix; the default value is 2; `ncov` = 1 for a common covariance matrix; `ncov` = 2 for the unequal covariance/scale matrices.

Value

- `pi` A g-dimensional initial vector of the mixing proportions.
- `mu` A initial $p \times g$ matrix of the location parameters.
- `sigma` A $p \times p$ covariance matrix if `ncov=1`, or a list of g covariance matrices with dimension $p \times p \times g$ if `ncov=2`.

Examples

```
n<-150
pi<-c(0.25,0.25,0.25,0.25)
sigma<-array(0,dim=c(3,3,4))
sigma[,,1]<-diag(1,3)
sigma[,,2]<-diag(2,3)
sigma[,,3]<-diag(3,3)
sigma[,,4]<-diag(4,3)
mu<-matrix(c(0.2,0.3,0.4,0.2,0.7,0.6,0.1,0.7,1.6,0.2,1.7,0.6),3,4)
dat<-rmix(n=n,pi=pi,mu=mu,sigma=sigma,ncov=2)
xi<-c(-0.5,1)
m<-rlabel(dat=dat$Y,pi=pi,mu=mu,sigma=sigma,xi=xi,ncov=2)
zm<-dat$clust
zm[m==1]<-NA
inits<-initialvalue(g=4,zm=zm,dat=dat$Y,ncov=2)
```

`list2par` *Transfer a list into a vector*

Description

Transfer a list into a vector

Usage

```
list2par(p, g, pi, mu, sigma, ncov = 2, xi = NULL, type = c("ign", "full"))
```

Arguments

- `p` Dimension of observation vecor.
- `g` Number of multivariate Gaussian groups.
- `pi` A g-dimensional initial vector of the mixing proportions.
- `mu` A initial $p \times g$ matrix of the location parameters.
- `sigma` A $p \times p$ covariance matrix if `ncov=1`, or a list of g covariance matrices with dimension $p \times p \times g$ if `ncov=2`.
- `ncov` Options of structure of sigma matrix; the default value is 2; `ncov` = 1 for a common covariance matrix that `sigma` is a $p \times p$ matrix. `ncov` = 2 for the unequal covariance/scale matrices that `sigma` represents a list of g matrices with dimension $p \times p \times g$.
- `xi` A 2-dimensional coefficient vector for a logistic function of the Shannon entropy.
- `type` Two types to fit to the model, 'ign' indicates fitting the model on the basis of the missing-label mechanism ignored, and 'full' indicates fitting the model on the basis of the missing-label mechanism

Value

- `par` a vector including all list information

`loglk_full` *Full log-likelihood function*

Description

Full log-likelihood function with both terms of ignoring and missing

Usage

```
loglk_full(dat, zm, pi, mu, sigma, ncov = 2, xi)
```

Arguments

- `dat` An $n \times p$ matrix where each row represents an individual observation
- `zm` An n-dimensional vector of group partition including the missing-label, denoted as NA.
- `pi` A g-dimensional initial vector of the mixing proportions.
- `mu` A initial $p \times g$ matrix of the location parameters.
- `sigma` A $p \times p$ covariance matrix if `ncov=1`, or a list of g covariance matrices with dimension $p \times p \times g$ if `ncov=2`.
- `ncov` Options of structure of sigma matrix; the default value is 2; `ncov` = 1 for a common covariance matrix; `ncov` = 2 for the unequal covariance/scale matrices.#'
- `xi` A 2-dimensional coefficient vector for a logistic function of the Shannon entropy.

Details

The full log-likelihood function can be expressed as

$$\log L_{PC}^{(full)}(\mathbf{\Psi}) = \log L_{PC}^{(ig)}(\theta) + \log L_{PC}^{(miss)}(\theta, \boldsymbol{\xi}),$$

where$\log L_{PC}^{(ig)}(\theta)$is the log likelihood function formed ignoring the missing in the label of the unclassified features, and $\log L_{PC}^{(miss)}(\theta, \boldsymbol{\xi})$ is the log likelihood function formed on the basis of the missing-label indicator.

Value

`lk`	Log-likelihood value

`loglk_ig` *Log likelihood for partially classified data with ingoring the missing mechanism*

Description

Log likelihood for partially classified data with ingoring the missing mechanism

Usage

```
loglk_ig(dat, zm, pi, mu, sigma, ncov = 2)
```

Arguments

`dat`	An $n \times p$ matrix where each row represents an individual observation
`zm`	An n-dimensional vector of group partition including the missing-label, denoted as NA.
`pi`	A g-dimensional initial vector of the mixing proportions.
`mu`	A initial $p \times g$ matrix of the location parameters.
`sigma`	A $p \times p$ covariance matrix if `ncov=1`, or a list of g covariance matrices with dimension $p \times p \times g$ if `ncov=2`.
`ncov`	Options of structure of sigma matrix; the default value is 2; `ncov = 1` for a common covariance matrix that `sigma` is a $p \times p$ matrix. `ncov = 2` for the unequal covariance/scale matrices that `sigma` represents a list of g matrices with dimension $p \times p \times g$.

Details

The log-likelihood function for partially classified data with ingoring the missing mechanism can be expressed as

$$\log L_{PC}^{(ig)}(\theta) = \sum_{j=1}^{n} \left[(1-m_j) \sum_{i=1}^{g} z_{ij} \left\{ \log \pi_i + \log f_i(y_j; \omega_i) \right\} + m_j \log \left\{ \sum_{i=1}^{g} \pi_i f_i(y_j; \omega_i) \right\} \right],$$

where m_j is a missing label indicator, z_{ij} is a zero-one indicator variable defining the known group of origin of each, and $f_i(y_j; \omega_i)$ is a probability density function with parameters ω_i.

Value

`lk`	Log-likelihood value.

`loglk_miss` — *Log likelihood function formed on the basis of the missing-label indicator*

Description

Log likelihood for partially classified data based on the missing mechanism with the Shanon entropy

Usage

```
loglk_miss(dat, zm, pi, mu, sigma, ncov = 2, xi)
```

Arguments

`dat`	An $n \times p$ matrix where each row represents an individual observation
`zm`	An n-dimensional vector of group partition including the missing-label, denoted as NA.
`pi`	A g-dimensional initial vector of the mixing proportions.
`mu`	A initial $p \times g$ matrix of the location parameters.
`sigma`	A $p \times p$ covariance matrix if `ncov=1`, or a list of g covariance matrices with dimension $p \times p \times g$ if `ncov=2`.
`ncov`	Options of structure of sigma matrix; the default value is 2; `ncov = 1` for a common covariance matrix that `sigma` is a $p \times p$ matrix. `ncov = 2` for the unequal covariance/scale matrices that `sigma` represents a list of g matrices with dimension $p \times p \times g$.
`xi`	A 2-dimensional coefficient vector for a logistic function of the Shannon entropy.

Details

The log-likelihood function formed on the basis of the missing-label indicator can be expressed by

$$\log L_{PC}^{(miss)}(\theta, \boldsymbol{\xi}) = \sum_{j=1}^{n} \left[(1 - m_j) \log \{1 - q(y_j; \theta, \boldsymbol{\xi})\} + m_j \log q(y_j; \theta, \boldsymbol{\xi})\right],$$

where $q(y_j; \theta, \boldsymbol{\xi})$ is a logistic function of the Shannon entropy $e_j(y_j; \theta)$, and m_j is a missing label indicator.

Value

`lk`	loglikelihood value

`logsumexp` *log summation of exponential function*

Description

log summation of exponential variable vector.

Usage

```
logsumexp(x)
```

Arguments

`x`	A variable vector.

Value

`val`	log summation of exponential variable vector.

`makelabelmatrix` *Label matrix*

Description

Convert group indicator into a label maxtrix.

Usage

```
makelabelmatrix(clust)
```

Arguments

`clust`	An n-dimensional vector of group partition.

Value

`Z`	A matrix of group indicator.

Examples

```
cluster<-c(1,1,2,2,3,3)
label_maxtrix<-makelabelmatrix(cluster)
```

`neg_objective_function` *Negative objective function for EMMIXSSL*

Description

Negative objective function for EMMIXSSL

Usage

```
neg_objective_function(dat, zm, g, par, ncov = 2, type = c("ign", "full"))
```

Arguments

`dat`	An $n \times p$ matrix where each row represents an individual observation
`zm`	An n-dimensional vector of group partition including the missing-label, denoted as NA.
`g`	Number of multivariate Gaussian groups.
`par`	An informative vector including `mu`, `pi`,`sigma` and `xi`.
`ncov`	Options of structure of sigma matrix; the default value is 2; `ncov` = 1 for a common covariance matrix; `ncov` = 2 for the unequal covariance/scale matrices.
`type`	Two types to fit to the model, 'ign' indicates fitting the model on the basis of the missing-label mechanism ignored,and 'full' indicates fitting the model on the basis of the missing-label mechanism.

Value

`val`	Value of negatvie objective function.

`normalise_logprob` *Normalize log-probability*

Description

Normalize log-probability.

Usage

```
normalise_logprob(x)
```

Arguments

`x`	A variable vector.

Value

`val`	A normalize log probability of variable vector.

par2list *Transfer a vector into a list*

Description

Transfer a vector into a list

Usage

```
par2list(par, g, p, ncov = 2, type = c("ign", "full"))
```

Arguments

par	A vector with list information.
g	Number of multivariate Gaussian groups.
p	Dimension of observation vecor.
ncov	Options of structure of sigma matrix; the default value is 2; ncov = 1 for a common covariance matrix that sigma is a $p \times p$ matrix. ncov = 2 for the unequal covariance/scale matrices that sigma represents a list of g matrices with dimension $p \times p \times g$.
type	Two types to fit to the model, 'ign' indicates fitting the model on the basis of the missing-label mechanism ignored,and 'full' indicates fitting the model on the basis of the missing-label mechanism.

Value

parlist	Return a list including mu, pi, sigma and xi.

pro2vec *Transfer a probability vector into a vector*

Description

Transfer a probability vector into an informative vector

Usage

```
pro2vec(pro)
```

Arguments

pro	An propability vector

Value

y An informative vector

`rlabel` *Generation of a missing-data indicator*

Description

Generate the missing label indicator

Usage

```
rlabel(dat, pi, mu, sigma, ncov = 2, xi)
```

Arguments

`dat`	An $n \times p$ matrix where each row represents an individual observation.
`pi`	A g-dimensional initial vector of the mixing proportions.
`mu`	A initial $p \times g$ matrix of the location parameters.
`sigma`	A $p \times p$ covariance matrix if `ncov=1`, or a list of g covariance matrices with dimension $p \times p \times g$ if `ncov=2`.
`ncov`	Options of structure of sigma matrix; the default value is 2; `ncov = 1` for a common covariance matrix that `sigma` is a $p \times p$ matrix. `ncov = 2` for the unequal covariance/scale matrices that `sigma` represents a list of g matrices with dimension $p \times p \times g$.
`xi`	A 2-dimensional coefficient vector for a logistic function of the Shannon entropy.

Value

`m`	A n-dimensional vector of missing label indicator. The element of outputs `m` represents its label indicator is missing if m equals 1, otherwise its label indicator is available if m equals to 0.

Examples

```
n<-150
pi<-c(0.25,0.25,0.25,0.25)
sigma<-array(0,dim=c(3,3,4))
sigma[,,1]<-diag(1,3)
sigma[,,2]<-diag(2,3)
sigma[,,3]<-diag(3,3)
sigma[,,4]<-diag(4,3)
mu<-matrix(c(0.2,0.3,0.4,0.2,0.7,0.6,0.1,0.7,1.6,0.2,1.7,0.6),3,4)
dat<-rmix(n=n,pi=pi,mu=mu,sigma=sigma,ncov=2)
xi<-c(-0.5,1)
m<-rlabel(dat=dat$Y,pi=pi,mu=mu,sigma=sigma,xi=xi,ncov=2)
```

rmix *Gaussian mixture model generator.*

Description

Generate random observations from the Gaussian mixture distributions.

Usage

```
rmix(n, pi, mu, sigma, ncov = 2)
```

Arguments

n	Number of observations.
pi	A g-dimensional initial vector of the mixing proportions.
mu	A initial $p \times g$ matrix of the location parameters.
sigma	A $p \times p$ covariance matrix if ncov=1, or a list of g covariance matrices with dimension $p \times p \times g$ if ncov=2.
ncov	Options of structure of sigma matrix; the default value is 2; ncov = 1 for a common covariance matrix that sigma is a $p \times p$ matrix. ncov = 2 for the unequal covariance/scale matrices that sigma represents a list of g matrices with dimension $p \times p \times g$.

Value

Y	An $n \times p$ numeric matrix with samples drawn in rows.
Z	An $n \times g$ numeric matrix; each row represents zero-one indicator variables defining the known group of origin of each.
clust	An n-dimensional vector of group partition.

Examples

```
n<-150
pi<-c(0.25,0.25,0.25,0.25)
sigma<-array(0,dim=c(3,3,4))
sigma[,,1]<-diag(1,3)
sigma[,,2]<-diag(2,3)
sigma[,,3]<-diag(3,3)
sigma[,,4]<-diag(4,3)
mu<-matrix(c(0.2,0.3,0.4,0.2,0.7,0.6,0.1,0.7,1.6,0.2,1.7,0.6),3,4)
dat<-rmix(n=n,pi=pi,mu=mu,sigma=sigma,ncov=2)
```

vec2cov *Transform a vector into a matrix*

Description

Transform a vector into a matrix i.e., Sigma=R^T*R

Usage

```
vec2cov(par)
```

Arguments

`par`	A vector representing a variance matrix

Details

The variance matrix is decomposed by computing the Choleski factorization of a real symmetric positive-definite square matrix. Then, storing the upper triangular factor of the Choleski decomposition into a vector.

Value

sigma A variance matrix

vec2pro *Transfer an informative vector to a probability vector*

Description

Transfer an informative vector to a probability vector

Usage

```
vec2pro(vec)
```

Arguments

`vec`	An informative vector

Value

pro A probability vector

Index

cov2vec, 2

discriminant_beta, 3

EMMIXSSL, 3

get_clusterprobs, 5
get_entropy, 6

initialvalue, 7

list2par, 8
loglk_full, 9
loglk_ig, 10
loglk_miss, 11
logsumexp, 12

makelabelmatrix, 12

neg_objective_function, 13
normalise_logprob, 13

par2list, 14
pro2vec, 14

rlabel, 15
rmix, 16

vec2cov, 17
vec2pro, 17