Package ‘MBCbook’

July 2, 2019

Encoding UTF-8
Type Package
Title Companion Package for the Book "Model-Based Clustering and Classification for Data Science" by Bouveyron et al. (2019, ISBN:9781108644181).
Version 0.1.1
Date 2019-07-1
Author Charles Bouveyron and Gilles Celeux and T. Brendan Murphy and Adrian E. Raftery
Maintainer Charles Bouveyron <charles.bouveyron@gmail.com>
Depends R (>= 3.1.0), mclust, Rmixmod, MASS, mvtnorm
Suggests network, jpeg
Description The companion package provides all original data sets and functions that are used in the book "Model-Based Clustering and Classification for Data Science" by Charles Bouveyron, Gilles Celeux, T. Brendan Murphy and Adrian E. Raftery (2019, ISBN:9781108644181).
License GPL (>= 2)
NeedsCompilation no
Repository CRAN
Date/Publication 2019-07-02 07:00:03 UTC

R topics documented:

MBCbook-package .................................................. 2
Advice .............................................................. 3
AIDSBlogs ..........................................................., 4
amazonFineFoods .................................................. 5
constrEM ........................................................... 5
Coworker ........................................................... 7
credit ................................................................. 7
denoisePatches .................................................... 8
Friend ............................................................... 9
Description

The companion package provides all original data sets and functions that are used in the book "Model-Based Clustering and Classification for Data Science" by Charles Bouveyron, Gilles Celeux, T. Brendan Murphy and Adrian E. Raftery (2019, ISBN:9781108644181).

Details

The DESCRIPTION file:

Encoding: UTF-8
Package: MBCbook
Type: Package
Title: Companion Package for the Book "Model-Based Clustering and Classification for Data Science" by Bouveyron et al. (2019, ISBN:9781108644181).
Version: 0.1.1
Date: 2019-07-1
Author: Charles Bouveyron and Gilles Celeux and T. Brendan Murphy and Adrian E. Raftery
Maintainer: Charles Bouveyron <charles.bouveyron@gmail.com>
Depends: R (>= 3.1.0), mclust, Rmixmod, MASS, mvtnorm
Suggests: network, jpeg
Description: The companion package provides all original data sets and functions that are used in the book "Model-Based Clustering and Classification for Data Science" by Bouveyron et al. (2019, ISBN:9781108644181).
License: GPL (>= 2)

Index of help topics:

AIDSblogs The AIDSBlogs data set
### Description

Lazega (2001) <doi:10.2307/3556688> collected a network data set detailing interactions between a set of 71 lawyers in a corporate law firm in the USA. The data include measurements of the advice network, friendship network and co-worker network between the lawyers within the firm. Further covariates associated with each lawyer in the firm are also available including age, seniority, college education and office location.
Usage
data("Advice")

Format
A large network object, which can be managed with the network library, with 71 nodes.

References

Examples
data(Advice)

AIDS Blogs
The AIDS Blogs data set

Description
The AIDS blog data set records the pattern of citation among 146 unique blogs related to AIDS patients and their support networks. The data were originally collected by Gopal (2007) <doi:10.1007/1-4020-5427-0_18> over a randomly selected three-day period in August 2005. The nodes in the network correspond to blogs and a directed edge from one blog to another indicates that the former had a link to the latter in their web page.

Usage
data("AIDS Blogs")

Format
A large network object, which can be managed with the network library, with 146 nodes.

References

Examples
data(AIDS Blogs)
amazonFineFoods

The Amazon Fine Foods data set

Description

The Amazon Fine Foods data set has 1646 rows and 1735 columns, describing whether an user (row) has noted and reviewed a product (column) or not.

Usage

data("amazonFineFoods")

Format

A data frame with binary values indicating whether an user (row) has noted and reviewed a product (column) or not.

Source


Examples

data(amazonFineFoods)

constrEM

Semi-supervised clustering with must-link constraints

Description

Semi-supervised clustering with must-link constraints allows to cluster data for which must-link constraints are available. This function implements the method described in Shental et al. (2003, ISBN:9781615679119).

Usage

constrEM(X, K, C, maxit = 30)
Arguments

- **X**: a data frame of observations, assuming the rows are the observations and the columns the variables. Note that NAs are not allowed.
- **K**: the number of desired groups.
- **C**: a vector encoding the must-link constraints through chunklets. This vector has to be of the length of the number of observations. Two observations that have to be in the same group must be in the same chunklet. For instance, the chunklet vector \((1, 2, 3, 4, 3, 5)\) indicate that 3rd and the 5th observations have a must-link constraint. If there is no must-link constraints, this vector should be simply \(1:nrow(X)\).
- **maxit**: the maximum number of iterations.

Value

A list is returned with the following fields:

- **cls**: a vector containing the group memberships of the observations.
- **T**: the posterior probabilities that the observations belong to the K groups.
- **prop**: the estimated mixture proportions.
- **mu**: the estimated mixture means.
- **S**: the estimated mixture covariance matrices.
- **ll**: the log-likelihood value at convergence.

Author(s)

C. Bouveyron

References


Examples

```r
# Simulation of some data
set.seed(123)
n = 200
m1 = c(0, 0); m2 = 4 * c(1, 1); m3 = 4 * c(1, 1)
S1 = diag(2); S2 = rbind(c(1, 0), c(0, 0.05))
S3 = rbind(c(0.05, 0), c(0, 1))
X = rbind(mvrnorm(n, m1, S1), mvrnorm(n, m2, S2), mvrnorm(n, m3, S3))
cls = rep(1:3, c(n, n, n))

# Encoding the constraints through chunklets
# Observations 397 and 408 are in the same chunklet
a = 398
```
Coworker

\[ b = 430 \]
\[ C = c(1:(b-1),a,b:(nrow(X)-1)) \]

# Clustering with constrEM
\[ \text{res} = \text{constrEM}(X,K=3,C,\text{maxit}=20) \]

**Coworker**

*The Coworker data set from Lazega (2001)*

**Description**

Lazega (2001) <doi:10.2307/3556688> collected a network data set detailing interactions between a set of 71 lawyers in a corporate law firm in the USA. The data include measurements of the advice network, friendship network and co-worker network between the lawyers within the firm. Further covariates associated with each lawyer in the firm are also available including age, seniority, college education and office location.

**Usage**

`data("Coworker")`

**Format**

A large network object, which can be managed with the network library, with 71 nodes.

**References**


**Examples**

`data(Coworker)`

---

**Credit**

*The Credit data set*

**Description**

The Credit data set has 66 rows and 11 columns, describing customers who took out loans from a credit company described with 11 categorical or ordinal variables.

**Usage**

`data("credit")`
denoisePatches

Format
A data frame with 66 observations and 11 categorical or ordinal variables.

Source
https://husson.github.io/data.html

Examples
data(credit)

denoisePatches  Denoising of image patches

Description
Denoising of image patches based on the clustering of patches.

Usage
denoisePatches(Y, out, P, sigma=10)

Arguments
Y a data frame containing as rows the image patches to denoise
out the mixmodCluster object that contains mixture parameters
P the posterior probabilities that patches belong to the clusters
sigma the noise standard deviation

Value
A data frame of the denoised patches is returned.

Note
C. Bouveyron & J. Delon

Examples
Im = diag(16)
ImNoise = Im + rnorm(256, 0, 0.1)
X = imageToPatch(ImNoise, 4)
out = mixmodCluster(X, 10, model=mixmodGaussianModel(family=c("spherical")))
res = mixmodPredict(X, out@bestResult)
Xdenoised = denoisePatches(X, out, P = res@proba, sigma = 0.1)
ImRec = reconstructImage(Xdenoised, 16, 16)
par(mfrow=c(1,3)); imshow(Im); imshow(ImNoise); imshow(ImRec)
Description

Lazega (2001) <doi:10.2307/3556688> collected a network data set detailing interactions between a set of 71 lawyers in a corporate law firm in the USA. The data include measurements of the advice network, friendship network and co-worker network between the lawyers within the firm. Further covariates associated with each lawyer in the firm are also available including age, seniority, college education and office location.

Usage

data("Friend")

Format

A large network object, which can be managed with the network library, with 71 nodes.

References


Examples

data(Friend)

imageToPatch

Transform an image into a collection of patches

Description

Transform an image into a collection of small images (patches) that cover the original image.

Usage

imageToPatch(Im, f)

Arguments

Im the image for which one wants to extract local patches.
f the size of the desired patches (fxf).
Value

A data frame of all extracted patches is returned.

Author(s)

C. Bouveyron & J. Delon

Examples

```r
Im = diag(16)
ImNoise = Im + rnorm(256, 0, 0.1)
X = imageToPatch(ImNoise, 4)
out = mixmodCluster(X, 10, model=mixmodGaussianModel(family=c("spherical")))
res = mixmodPredict(X, out@bestResult)
Xdenoised = denoisePatches(X, out, P = res@proba, sigma = 0.1)
ImRec = reconstructImage(Xdenoised, 16, 16)
par(mfrow=c(1,3)); imshow(Im); imshow(ImNoise); imshow(ImRec)
```

Description

A simple way of displaying an image, using the `image` function.

Usage

```r
imshow(x, col=palette(gray(0:255/255)), useRaster = TRUE,...)
```

Arguments

- `x`: the image to display as a matrix.
- `col`: the color palette to use when displaying the image.
- `useRaster`: logical; if TRUE a bitmap raster is used to plot the image instead of polygons. The grid must be regular in that case, otherwise an error is raised. For the behaviour when this is not specified, see the ‘Details’ section of the `image` function.
- `...`: additional arguments to provide to subfunctions.

See Also

`image`

Examples

```r
Im = diag(16)
imshow(Im)
```
NIR

*The chemometrics near-infrared (NIR) data set*

**Description**

The chemometrics near-infrared (NIR) data set has 202 observations and 2801 variables: 2800 near-infrared wavelength measures and 1 class variable. The data were obtained from the analysis of three types of textiles. The data set was first introduced in Devos et al. (2009) [10.1016/j.chemolab.2008.11.005].

**Usage**

```R
data("velibCount")
```

**Format**

A data frame with 202 observations and 2801 variables. The first variable indicates the class-memberships of the observations.

**References**


**Examples**

```R
data(NIR)
matplot(t(NIR[,][-1]),type='l',col=NIR[,1])
```

PoliticalBlogs

*The political blog data set*

**Description**

The political blog data set shows the linking structure in online blogs which commentate on French political issues; the data were collected by Observatoire Presidentielle in October 2006. The data were first used by Latouche et al. (2011) [10.1214/10-AOAS382].

**Usage**

```R
data("PoliticalBlogs")
```

**Format**

A large network object, which can be managed with the network library, with 196 nodes.
References


Examples

data(PoliticalBlogs)

# Visualization with the network library
library(network)
plot(PoliticalBlogs)

puffin

The puffin data set

Description

The puffin data set contains 69 individuals (birds) described by 5 categorical variables, in addition to class labels.

Usage

data("puffin")

Format

A data frame with 69 observations and 6 variables.

class  the class of the observations
gender  gender of the bird
eyebrow gender of the bird
collar  gender of the bird
sub.caudal gender of the bird
border  gender of the bird

Source

The data were provided by Bretagnolle, V., Museum d’Histoire Naturelle, Paris.

Examples

data(puffin)
reconstructImage

Reconstructing an image from a patch decomposition

Description
A simple way of reconstructing an image from a patch decomposition.

Usage
reconstructImage(X,nl,nc)

Arguments
X  the matrix of patches to be used for reconstructing the image.
nl  the number of rows of the image.
nc  the number of columns of the image.

Value
an image is returned as a matrix object, that can be display with the imshow function.

Author(s)
C. Bouveyron & J. Delon

Examples
Im = diag(16)
ImNoise = Im + rnorm(256,0,0.1)
X = imageToPatch(ImNoise,4)
out = mixmodCluster(X,10,model=mixmodGaussianModel(family=c("spherical")))
res = mixmodPredict(X,out@bestResult)
Xdenoised = denoisePatches(X,out,P = res@proba,sigma = 0.1)
ImRec = reconstructImage(Xdenoised,16,16)
par(mfrow=c(1,3)); imshow(Im); imshow(ImNoise); imshow(ImRec)

rqda
Robust (quadratic) discriminant analysis

Description
Robust (quadratic) discriminant analysis implements a discriminant analysis method which is robust
to label noise. This function implements the method described in Lawrence and Scholkopf (2003,
Usage

rqda(X, lb1, Y, maxit=50, disp=FALSE,...)

Arguments

X a data frame containing the learning observations.

lb1 the class labels of the learning observations.

Y a data frame containing the new observations to classify.

maxit the maximum number of iterations.

disp logical, if TRUE, several plots are displayed.

... additional arguments to provide to subfunctions.

Value

A list is returned with the following elements:

nu the estimated class proportions.

mu the estimated class means.

S the estimated covariance matrices.

gamma the estimated purity level of the labels.

Ti the posterior probabilities of the labels knowing the observed labels for the learning observations.

Pi the class posterior probabilities of the observations to classify.

cls the class assignments of the observations to classify.

ll the log-likelihood value.

Author(s)

C. Bouveyron

References


Examples

n = 50

m1 = c(0,0); m2 = 1.5*c(1,-1)

S1 = 0.1*diag(2); S2 = 0.25*diag(2)

X = rbind(mvrnorm(n,m1,S1),mvrnorm(2*n,m2,S2))

cls = rep(1:2,c(n,2*n))

# Label perturbation

ind = rbinom(3*n,1,0.4); lb = cls
UScongress

The US congress vote data set

Description

The US congress vote data set contains the votes (yes, no, abstained or absent) of 434 members of the 98th US Congress on 16 different key issues. This data set involves three-level categorical data.

Usage

data("UScongress")

Format

A data frame with 434 observations on 16 different key issues. The first variables indicates the political party of the congressmen.

Source


Examples

data(UScongress)

usps358

The handwritten digits usps358 data set

Description

The handwritten digits usps358 data set is a subset of the famous USPS data from UCI, which contains only the 1 756 images of the digits 3, 5 and 8.

Usage

data("usps358")
Format

A data frame with 1756 observations on the following 257 variables: cls is a numeric vector encoding the class of the digits, V1 to V256 are numeric vectors corresponding to the pixels of the 8x8 images.

Source

The data set is a subset of the famous USPS data from UCI (https://archive.ics.uci.edu/ml/index.php). The usps358 data set contains only the 1756 images of the digits 3, 5 and 8 which are the most difficult digits to discriminate.

Examples

data(usps358)

Description


Usage

varSelEM(X,G,maxit=100,eps=1e-6)

Arguments

X a data frame containing the observations to cluster.
G the expected number of groups (integer).
maxit the maximum number of iterations (integer). The default value is 100.
eps the convergence threshold. The default value is 1e-6.

Value

A list is returned with the following elements:

mu the group means for relevant variables.
sigma the group variances for relevant variables.
lambda the group means for irrelevant variables
alpha the group variances for irrelevant variables.
rho the feature saliency.
P the group posterior probabilities.
c1s the group memberships.
ll the log-likelihood value.
The bivariate Vélib data set

Description

The bivariate Vélib data set contains data from the bike sharing system of Paris, called Vélib. The data are loading profiles and percentage of broken docks of the bike stations over one week. The data were collected every hour during the period Sunday 1st Sept. to Sunday 7th Sept., 2014. The data were first used in Bouveyron et al. (2015) <doi:10.1214/15-AOAS861>.

Usage

data("velib2D")

Format

The format is:
- availableBikes: the loading profiles (nb of available bikes / nb of bike docks) of the 1189 stations at 181 time points.
- brokenDocks: the percentage of broken docks of the 1189 stations at 181 time points.
- position: the longitude and latitude of the 1189 bike stations.
- dates: the download dates.
- bonus: indicates if the station is on a hill (bonus = 1).
- names: the names of the stations.
### Source

The real time data are available at https://developer.jcdecaux.com/ (with an api key).

### References


### Examples

```r
data(velib2D)
```

---

### Description

The discrete version (count data) of Vélib data set contains data from the bike sharing system of Paris, called Vélib. The data consist in the number of bikes at stations over one week. The data were collected every hour during the period Sunday 1st Sept. - Sunday 7th Sept., 2014. The data were first used in Bouveyron et al. (2015) <doi:10.1214/15-AOAS861>.

### Usage

```r
data("velibCount")
```

### Format

The format is:
- data: the nb of available bikes of the 1189 stations at 181 time points.
- position: the longitude and latitude of the 1189 bike stations.
- dates: the download dates.
- bonus: indicates if the station is on a hill (bonus = 1).
- names: the names of the stations.

### Source

The real time data are available at https://developer.jcdecaux.com/ (with an api key).

### References

**Examples**

```r
data(wine27)
```

**Description**

The (27-dimensional) Italian Wine data set is the result of a chemical analysis of 178 wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 27 constituents found in each of the three types of wines.

**Usage**

```r
data("wine27")
```

**Format**

A data frame with 178 observations on the following 29 variables.

- Alcohol: a numeric vector
- Sugar.free_extract: a numeric vector
- Fixed.acidity: a numeric vector
- Tartaric.acid: a numeric vector
- Malic.acid: a numeric vector
- Uronic.acids: a numeric vector
- pH: a numeric vector
- Ash: a numeric vector
- Alcalinity.of.ash: a numeric vector
- Potassium: a numeric vector
- Calcium: a numeric vector
- Magnesium: a numeric vector
- Phosphate: a numeric vector
- Chloride: a numeric vector
- Total.phenols: a numeric vector
- Flavanoids: a numeric vector
- Nonflavanoid.phenols: a numeric vector
- Proanthocyanins: a numeric vector
- Color.Intensity: a numeric vector
- Hue: a numeric vector
- O.D.280._O.D.315.of.diluted.wines: a numeric vector
OD280.0D315_of_flavanoids a numeric vector
Glycerol a numeric vector
X2_3.butanediol a numeric vector
Total_nitrogen a numeric vector
Proline a numeric vector
Methanol a numeric vector
Type a factor with levels Barbera, Barolo, Grignolino
Year a numeric vector

Details

This data set is an expended version of the popular one from the UCI machine learning repository (http://archive.ics.uci.edu/ml/datasets/Wine).

Examples

data(wine27)
Index

*Topic datasets
  Advice, 3
  AIDSBlogs, 4
  amazonFineFoods, 5
  Coworker, 7
  credit, 7
  Friend, 9
  NIR, 11
  PoliticalBlogs, 11
  puffin, 12
  UScongress, 15
  usps358, 15
  velib2D, 17
  velibCount, 18
  wine27, 19

*Topic package
  MBCbook-package, 2
  reconstructImage, 13
  rqda, 13
  UScongress, 15
  usps358, 15
  varSelEM, 16
  velib2D, 17
  velibCount, 18
  wine27, 19

Advice, 3
AIDSBlogs, 4
amazonFineFoods, 5

constrEM, 5
Coworker, 7
credit, 7
denoisePatches, 8
Friend, 9

image, 10
imageToPatch, 9
imshow, 10, 13

MBCbook (MBCbook-package), 2
MBCbook-package, 2

NIR, 11
PoliticalBlogs, 11
puffin, 12