Package 'MBCbook'

April 9, 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

Date 2019-04-11

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 2019-04-09 15:15:17 UTC

R topics documented:

`MBCbook-package` *Companion Package for the Book "Model-Based Clustering and Classification for Data Science" by Bouveyron et al. (2019, ISBN:9781108644181).*

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
Version:	0.1
Date:	2019-04-11
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 C
License:	GPL (>= 2)

Index of help topics:

`AIDSBlogs` `The AIDSBlogs data set`

Advice	The Advice data set from Lazega (2001)
Coworker	The Coworker data set from Lazega (2001)
Friend	The Friend data set from Lazega (2001)
MBCbook-package	Companion Package for the Book "Model-Based Clustering and Classification for Data Science" by Bouveyron et al. (2019, ISBN:9781108644181).
NIR	The chemometrics near-infrared (NIR) data set
PoliticalBlogs	The political blog data set
UScongress	The US congress vote data set
amazonFineFoods	The Amazon Fine Foods data set
constrEM	Semi-supervised clustering with must-link constraints
credit	The Credit data set
denoisePatches	Denoising of image patches
imageToPatch	Transform an image into a collection of patches
imshow	Display an image
puffin	The puffin data set
reconstructImage	Reconstructing an image from a patch decomposition
rqda	Robust (quadratic) discriminant analysis
usps358	The handwritten digits usps358 data set
varSelEM	A variable selection algorithm for clustering
velib2D	The bivariate Vélib data set
velibCount	The discrete version (count data) of the Vélib data set
wine27	The (27-dimensional) Italian Wine data set

Author(s)

Charles Bouveyron and Gilles Celeux and T. Brendan Murphy and Adrian E. Raftery

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

References

Charles Bouveyron and Gilles Celeux and T. Brendan Murphy and Adrian E. Raftery, Model-Based Clustering and Classification for Data Science: with Applications in R, Cambridge University Press, 2019.

`Advice` *The Advice 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("Advice")
```

Format

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

References

Lazega, E., The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership, Oxford University Press, 2001 <doi:10.2307/3556688>.

Examples

```
data(Advice)
```

AIDSBlogs *The AIDSBlogs 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("AIDSBlogs")
```

Format

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

References

Gopal, S., The evolving social geography of blogs, in Miller, H. J. (ed.), Societies and Cities in the Age of Instant Access, The GeoJournal Library, vol. 88., pp. 275–293, 2007 <doi:10.1007/1-4020-5427-0_18>.

Examples

```
data(AIDSBlogs)
```

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

https://snap.stanford.edu/data/web-FineFoods.html.

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 chuncklets. 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 chuncklet. For instance, the chuncklet 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 containg 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

This function implements the method described in Shental, N., Bar-Hillel, A., Hertz, T., and Weinshall, D., Computing Gaussian mixture models with EM using equivalence constraints, Proceedings of the 16th International Conference on Neural Information Processing Systems, pages 465–472, 2003 (ISBN:9781615679119).

Examples

```
# 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
```

```
b = 430
C = c(1:(b-1),a,b:(nrow(X)-1))

# Clustering with constrEM
res = constrEM(X,K=3,C)
```

`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

Lazega, E., The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership, Oxford University Press, 2001 <doi:10.2307/3556688>.

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")
```

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 fame 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)
```

Friend *The Friend 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("Friend")
```

Format

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

References

Lazega, E., The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership, Oxford University Press, 2001 <doi:10.2307/3556688>.

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

```
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)
```

---

imshow *Display an image*

Description

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

Usage

```
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.
...	additionial arguments to provide to subfunctions.

See Also

image

Examples

```
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 introduce in Devos et al. (2009) <doi:10.1016/j.chemolab.2008.11.005>.

Usage

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

Format

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

References

Devos, O., Ruckebusch, C., Durand, A., Duponchel, L., and Huvenne, J.-P., Support vector machines (SVM) in near infrared (NIR) spectroscopy: Focus on parameters optimization and model interpretation, Chemometrics and Intelligent Laboratory Systems, 96, 27–33, 2009 <doi:10.1016/j.chemolab.2008.11.005>.

Examples

```
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) <doi:10.1214/10-AOAS382>.

Usage

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

Format

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

References

P. Latouche, E. Birmelé, and C. Ambroise. "Overlapping stochastic block models with application to the French political blogosphere". In : Annals of Applied Statistics 5.1, p. 309-336, 2011 <doi:10.1214/10-AOAS382>.

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, ISBN:1-55860-778-1).

Usage

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

Arguments

X	a data frame containing the learning observations.
lbl	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 probabilties 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

Lawrence, N., and Scholkopf, B., Estimating a kernel Fisher discriminant in the presence of label noise, Pages 306–313 of: Proceedings of the Eighteenth International Conference on Machine Learning. ICML'01. San Francisco, CA, USA, 2001 (ISBN:1-55860-778-1).

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
```

```
lb[ind==1 & cls==1] = 2
lb[ind==1 & cls==2] = 1

# Classification with RQDA
res = rqda(X,lb,X)
table(cls,res$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

http://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records

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 ot 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 1 756 images of the digits 3, 5 and 8 which are the most difficult digits to discriminate.

Examples

```
data(usps358)
```

---

varSelEM *A variable selection algorithm for clustering*

Description

A variable selection algorithm for clustering which implements the method described in Law et al. (2004) <doi:10.1109/TPAMI.2004.71>.

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.
`cls`	the group memberships.
`ll`	the log-likelihood value.

Author(s)

C. Bouveyron

References

Law, M. H., Figueiredo, M. A. T., and Jain, A. K., Simultaneous feature selection and clustering using mixture models, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, pp. 1154–1166, 2004 <doi:10.1109/TPAMI.2004.71>.

Examples

```
data(wine27)
X = scale(wine27[,1:27])
cls = wine27$Type

# Clustering and variable selection with VarSelEM
res = varSelEM(X,G=3)

# Clustering table
table(cls,res$cls)
```

---

velib2D *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. - 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.

- brokenDockss: 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

The data were first used in C. Bouveyron, E. Côme and J. Jacques, The discriminative functional mixture model for the analysis of bike sharing systems, The Annals of Applied Statistics, vol. 9 (4), pp. 1726-1760, 2015 <doi:10.1214/15-AOAS861>.

Examples

```
data(velib2D)
```

`velibCount` *The discrete version (count data) of the Vélib data set*

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

```
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

The data were first used in C. Bouveyron, E. Côme and J. Jacques, The discriminative functional mixture model for the analysis of bike sharing systems, The Annals of Applied Statistics, vol. 9 (4), pp. 1726-1760, 2015 <doi:10.1214/15-AOAS861>.

Examples

```
data(velib2D)
```

`wine27` *The (27-dimensional) Italian Wine data set*

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

```
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

`OD280.OD315_of_diluted_wines` a numeric vector

`OD280.OD315_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