Package 'MixtureMissing'

January 17, 2024

Type Package

Title Robust and Flexible Model-Based Clustering for Data Sets with Missing Values at Random

Version 3.0.1

Description Implementations of various robust and flexible model-based clustering methods for data sets with missing values at random. Two main models are: Multivariate Contaminated Normal Mixture (MCNM, Tong and Tortora, 2022, <doi:10.1007/s11634-021-00476-1>) and Multivariate Generalized Hyperbolic Mixture (MGHM, Wei et al., 2019, <doi:10.1016/j.csda.2018.08.016>). Mixtures via some special or limiting cases of the multivariate generalized hyperbolic distribution are also included: Normal-Inverse Gaussian, Symmetric Normal-Inverse Gaussian, Skew-Cauchy, Cauchy, Skew-t, Student's t, Normal, Symmetric Generalized Hyperbolic, Hyperbolic Univariate Marginals, Hyperbolic, and Symmetric Hyperbolic.

Imports mvtnorm (>= 1.1-2), mnormt (>= 2.0.2), cluster (>= 2.1.2), MASS (>= 7.3), numDeriv (>= 8.1.1), Bessel (>= 0.6.0)

Suggests mice (>= 3.10.0)

License GPL (>= 2)

Encoding UTF-8

LazyData true

Repository CRAN

RoxygenNote 7.2.3

Depends R (>= 3.5.0)

NeedsCompilation no

Author Hung Tong [aut, cre], Cristina Tortora [aut, ths, dgs]

Maintainer Hung Tong <hungtongmx@gmail.com>

Date/Publication 2024-01-17 17:52:06 UTC

R topics documented:

`auto` *Automobile Data Set*

Description

This data set consists of three types of entities: (a) the specification of an auto in terms of various characteristics, (b) its assigned insurance risk rating, (c) its normalized losses in use as compared to other cars. The second rating corresponds to the degree to which the auto is more risky than its price indicates. Cars are initially assigned a risk factor symbol associated with its price. Then, if it is more risky (or less), this symbol is adjusted by moving it up (or down) the scale. Actuarians call this process "symboling". A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.

Usage

```
auto
```

Format

A data frame with 205 rows and 26 variables. The first 15 variables are continuous, while the last 11 variables are categorical. There are 45 rows with missing values.

normalized_losses continuous from 65 to 256.

wheel_base continuous from 86.6 120.9.

length continuous from 141.1 to 208.1.

width continuous from 60.3 to 72.3.

height continuous from 47.8 to 59.8.

curb_weight continuous from 1488 to 4066.

engine_size continuous from 61 to 326.

bore continuous from 2.54 to 3.94.

stroke continuous from 2.07 to 4.17.

compression_ratio continuous from 7 to 23.

horsepower continuous from 48 to 288.

peak_rpm continuous from 4150 to 6600.

city_mpg continuous from 13 to 49.

highway_mpg continuous from 16 to 54.

price continuous from 5118 to 45400.

symboling -3, -2, -1, 0, 1, 2, 3.

make alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo

fuel_type diesel, gas.

aspiration std, turbo.

num_doors four, two.

body_style hardtop, wagon, sedan, hatchback, convertible.

drive_wheels 4wd, fwd, rwd.

engine_location front, rear.

engine_type dohc, dohcv, l, ohc, ohcf, ohcv, rotor.

num_cylinders eight, five, four, six, three, twelve, two.

fuel_system 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.

Source

Kibler, D., Aha, D.W., & Albert,M. (1989). Instance-based prediction of real-valued attributes. Computational Intelligence, Vol 5, 51–57. https://archive.ics.uci.edu/ml/datasets/automobile

---

`bankruptcy` *Bankruptcy Data Set*

Description

The data set contains the ratio of retained earnings (RE) to total assets, and the ratio of earnings before interests and taxes (EBIT) to total assets of 66 American firms recorded in the form of ratios. Half of the selected firms had filed for bankruptcy.

Usage

```
bankruptcy
```

Format

A data frame with 66 rows and 3 variables:

Y Status of the firm: 0 for bankruptcy and 1 for financially sound.

RE Ratio of retained earnings.

EBIT Ratio of earnings before interests and taxes.

Source

Altman E.I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J Finance* 23(4): 589-609 https://www.jstor.org/stable/2978933

---

evaluation_metrics *Binary Classification Evaluation*

Description

Evaluate the performance of a classification model by comparing its predicted labels to the true labels. Various metrics are returned to give an insight on how well the model classifies the observations. This function is added to aid outlier detection evaluation of MCNM and MtM in case that true outliers are known in advance.

Usage

```
evaluation_metrics(true_labels, pred_labels)
```

Arguments

`true_labels`	An 0-1 or logical vector denoting the true labels. The meaning of 0 and 1 (or TRUE and FALSE) is up to the user.
`pred_labels`	An 0-1 or logical vector denoting the true labels. The meaning of 0 and 1 (or TRUE and FALSE) is up to the user.

Value

A list with the following slots:

`matr`	The confusion matrix built upon true labels and predicted labels.
`TN`	True negative.
`FP`	False positive (type I error).
`FN`	False negative (type II error).
`TP`	True positive.
`TPR`	True positive rate (sensitivy).
`FPR`	False positive rate.
`TNR`	True negative rate (specificity).

FNR	False negative rate.
precision	Precision or positive predictive value (PPV).
accuracy	Accuracy.
error_rate	Error rate.
FDR	False discovery rate.

Examples

```
#++++ Inputs are 0-1 vectors ++++#

evaluation_metrics(
  true_labels = c(1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1),
  pred_labels = c(1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1)
)

#++++ Inputs are logical vectors ++++#

evaluation_metrics(
  true_labels = c(TRUE, FALSE, FALSE, FALSE, TRUE, TRUE, TRUE, TRUE, FALSE, FALSE),
  pred_labels = c(FALSE, FALSE, TRUE, FALSE, TRUE, FALSE, FALSE, TRUE, FALSE, FALSE)
)
```

---

generate_patterns *Missing-Data Pattern Generation*

Description

Generate all possible missing patterns in a multivariate data set. The function can be used to complement the function `ampute()` from package `mice` in which a matrix of patterns is needed to allow for general missing-data patterns with missing-data mechanism missing at random (MAR). Using this function, each observation can have more than one missing value.

Usage

```
generate_patterns(d)
```

Arguments

d	The number of variables or columns of the data set. d must be an integer greater than 1.

Details

An observation cannot have all values missing values. A complete observation is not qualified for missing-data pattern. Note that a large value of d may result in memory allocation error.

Value

A matrix where 0 indicates that a variable should have missing values and 1 indicates that a variable should remain complete. This matrix has d columns and $2^d - 2$ rows.

Examples

```
generate_patterns(4)

#++++ To use with the function ampute() from package mice ++++#
library(mice)

patterns_matr <- generate_patterns(4)
data_missing <- ampute(iris[1:4], prop = 0.5, patterns = patterns_matr)$amp
```

---

hide_values *Missing Values Generation*

Description

A convenient function that randomly introduces missing values to an at-least-bivariate data set. The user can specify either the proportion of observations that contain some missing values or the exact number of observations that contain some missing values. Note that the function does not guarantee that underlying missing-data mechanism to be missing at random (MAR).

Usage

```
hide_values(X, prop_cases = 0.1, n_cases = NULL)
```

Arguments

X	An n by d matrix or data frame where n is the number of observations and d is the number of columns or variables. X must have at least 2 rows and 2 columns.
prop_cases	(optional) Proportion of observations that contain some missing values. prop_cases must be a number in $(0, 1)$. prop_cases = 0.1 by default, but will be ignored if n_cases is specified.
n_cases	(optional) Number of observations that contain some missing values. n_cases must be an integer ranging from 1 to nrow(X) - 1.

Details

If subject to missingness, an observation can have at least 1 and at most ncol(X) - 1 missing values. Depending on the data set, it is not guaranteed that the resulting matrix will have the number of rows with missing values matches the specified proportion.

Value

The orginal n by d matrix or data frame with missing values.

Examples

```
set.seed(1234)

hide_values(iris[1:4])
hide_values(iris[1:4], prop_cases = 0.5)
hide_values(iris[1:4], n_cases = 80)
```

`initialize_clusters` *Cluster Initialization using a Heuristic Method*

Description

Initialize cluster memberships and component parameters to start the EM algorithm using a heuristic clustering method or user-defined labels.

Usage

```
initialize_clusters(
  X,
  G,
  init_method = c("kmedoids", "kmeans", "hierarchical", "manual"),
  clusters = NULL
)
```

Arguments

`X`	An n x d matrix or data frame where n is the number of observations and d is the number of columns or variables. Alternately, `X` can be a vector of n observations.
`G`	The number of clusters, which must be at least 1. If `G = 1`, then user-defined `clusters` is ignored.
`init_method`	(optional) A string specifying the method to initialize the EM algorithm. "kmedoids" clustering is used by default. Alternative methods include "kmeans", "hierarchical", "manual". When "manual" is chosen, a vector `clusters` of length n must be specified. When `G = 1` and "kmedoids" clustering is used, the medoid will be returned, not the sample mean.
`clusters`	A numeric vector of length n that specifies the initial cluster memberships of the user when `init_method` is set to "manual". This argument is NULL by default, so that it is ignored whenever other given initialization methods are chosen.

Details

Available heuristic methods include k-medoids clustering, k-means clustering, and hierarchical clustering. Alternately, the user can also enter pre-specified cluster memberships, making other initialization methods possible. If the given data set contains missing values, only observations with complete records will be used to initialize clusters. However, in this case, except when `G = 1`, the resulting cluster memberships will be set to `NULL` since they represent those complete records rather than the original data set as a whole.

Value

A list with the following slots:

`pi`	Component mixing proportions.
`mu`	A G by d matrix where each row is the component mean vector.
`Sigma`	A G-dimensional array where each d by d matrix is the component covariance matrix.
`clusters`	An numeric vector with values from 1 to G indicating initial cluster memberships if `X` is a complete data set; NULL otherwise.

References

Everitt, B., Landau, S., Leese, M., and Stahl, D. (2011). *Cluster Analysis*. John Wiley & Sons.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Applied Statistics*, **28**, 100-108. doi: 10.2307/2346830.

Examples

```
#++++ Initialization using a heuristic method ++++#


#++++ Initialization using user-defined labels ++++#

init <- initialize_clusters(iris[1:4], G = 3, init_method = 'manual',
                            clusters = as.numeric(iris$Species))

#++++ Initial parameters and pairwise scatterplot showing the mapping ++++#

init$pi
init$mu
init$Sigma
init$clusters

pairs(iris[1:4], col = init$clusters, pch = 16)
```

Description

Carries out model-based clustering using a multivariate contaminated normal mixture (MCNM). The function will determine itself if the data set is complete or incomplete and fit the appropriate model accordingly. In the incomplete case, the data set must be at least bivariate, and missing values are assumed to be missing at random (MAR).

Usage

```
MCNM(
  X,
  G,
  max_iter = 20,
  epsilon = 0.01,
  init_method = c("kmedoids", "kmeans", "hierarchical", "manual"),
  clusters = NULL,
  eta_min = 1.001,
  progress = TRUE
)
```

Arguments

`X`	An n x d matrix or data frame where n is the number of observations and d is the number of variables.
`G`	The number of clusters, which must be at least 1. If `G = 1`, then both `init_method` and `clusters` are ignored.
`max_iter`	(optional) A numeric value giving the maximum number of iterations each EM algorithm is allowed to use; 20 by default.
`epsilon`	(optional) A number specifying the epsilon value for the Aitken-based stopping criterion used in the EM algorithm: 0.01 by default.
`init_method`	(optional) A string specifying the method to initialize the EM algorithm. "kmedoids" clustering is used by default. Alternative methods include "kmeans", "hierarchical", and "manual". When "manual" is chosen, a vector `clusters` of length n must be specified. If the data set is incomplete, missing values will be first filled based on the mean imputation method.
`clusters`	(optional) A numeric vector of length n that specifies the initial cluster memberships of the user when `init_method` is set to "manual". This argument is NULL by default, so that it is ignored whenever other given initialization methods are chosen.
`eta_min`	(optional) A numeric value close to 1 to the right specifying the minimum value of eta; 1.001 by default.
`progress`	(optional) A logical value indicating whether the fitting progress should be displayed; TRUE by default.

Value

An object of class `MixtureMissing` with:

model	The model used to fit the data set.
pi	Mixing proportions.
mu	Component mean vectors (location).
Sigma	Component covariance matrices (dispersion).
alpha	Component proportions of good observations.
eta	Component degrees of contamination.
z_tilde	An n by G matrix where each row indicates the expected probabilities that the corresponding observation belongs to each cluster.
v_tilde	An n by G matrix where each row indicates the expected probabilities that the corresponding observation is good with respect to each cluster.
clusters	A numeric vector of length n indicating cluster memberships determined by the model.
outliers	A logical vector of length n indicating observations that are outliers.
data	The original data set if it is complete; otherwise, this is the data set with missing values imputed by appropriate expectations.
complete	A logical vector of length n indicating which observation(s) have no missing values.
npar	The breakdown of the number of parameters to estimate.
max_iter	Maximum number of iterations allowed in the EM algorithm.
iter_stop	The actual number of iterations needed when fitting the data set.
final_loglik	The final value of log-likelihood.
loglik	All the values of log-likelihood.
AIC	Akaike information criterion.
BIC	Bayesian information criterion.
KIC	Kullback information criterion.
KICc	Corrected Kullback information criterion.
AIC3	Modified AIC.
CAIC	Bozdogan's consistent AIC.
AICc	Small-sample version of AIC.
ent	Entropy.
ICL	Integrated Completed Likelihood criterion.
AWE	Approximate weight of evidence.
CLC	Classification likelihood criterion.
init_method	The initialization method used in model fitting.

References

Punzo, A. and McNicholas, P.D., 2016. Parsimonious mixtures of multivariate contaminated normal distributions. *Biometrical Journal, 58*(6), pp.1506-1537.

Tong, H. and, Tortora, C., 2022. Model-based clustering and outlier detection with missing data. *Advances in Data Analysis and Classification*.

Examples

```
data('auto')

#++++ With no missing values ++++#

X <- auto[, c('engine_size', 'city_mpg', 'highway_mpg')]
mod <- MCNM(X, G = 2, init_method = 'kmedoids', max_iter = 10)

summary(mod)
plot(mod)

#++++ With missing values +++++#

X <- auto[, c('normalized_losses', 'horsepower', 'highway_mpg', 'price')]
mod <- MCNM(X, G = 2, init_method = 'kmedoids', max_iter = 10)

summary(mod)
plot(mod)
```

---

mean_impute *Mean Imputation*

Description

Replace missing values of data set by the mean of other observed values.

Usage

```
mean_impute(X)
```

Arguments

X	An n x d matrix or data frame where n is the number of observations and d is the number of columns or variables. Alternately, X can be a vector of n observations.

Value

A complete data matrix with missing values imputed accordingly.

References

Schafer, J. L. and Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7(2):147–177.

Little, R. J. A. and Rubin, D. B. (2020). *Statistical analysis with missing data*. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ, 3rd edition

Examples

```
X <- matrix(nrow = 6, ncol = 3, byrow = TRUE, c(
  NA,  2,  2,
   3, NA,  5,
   4,  3,  2,
  NA, NA,  3,
   7,  2, NA,
  NA,  4,  2
))


mean_impute(X)
```

MGHM *Multivariate Generalized Hyperbolic Mixture (MGHM)*

Description

Carries out model-based clustering using a multivariate generalized hyperbolic mixture (MGHM). The function will determine itself if the data set is complete or incomplete and fit the appropriate model accordingly. In the incomplete case, the data set must be at least bivariate, and missing values are assumed to be missing at random (MAR).

Usage

```
MGHM(
  X,
  G,
 model = c("GH", "NIG", "SNIG", "SC", "C", "St", "t", "N", "SGH", "HUM", "H", "SH"),
  max_iter = 20,
  epsilon = 0.01,
  init_method = c("kmedoids", "kmeans", "hierarchical", "manual"),
  clusters = NULL,
  outlier_cutoff = 0.95,
 deriv_ctrl = list(eps = 1e-08, d = 1e-04, zero.tol = sqrt(.Machine$double.eps/7e-07), r
    = 6, v = 2, show.details = FALSE),
  progress = TRUE
)
```

Arguments

X	An n x d matrix or data frame where n is the number of observations and d is the number of variables.
G	The number of clusters, which must be at least 1. If G = 1, then both `init_method` and `clusters` are ignored.

`model` A string indicating the mixture model to be fitted; "GH" for generalized hyperbolic by default. See the details section for a list of available distributions.

`max_iter` (optional) A numeric value giving the maximum number of iterations each EM algorithm is allowed to use; 20 by default.

`epsilon` (optional) A number specifying the epsilon value for the Aitken-based stopping criterion used in the EM algorithm: 0.01 by default.

`init_method` (optional) A string specifying the method to initialize the EM algorithm. "kmedoids" clustering is used by default. Alternative methods include "kmeans", "hierarchical", and "manual". When "manual" is chosen, a vector `clusters` of length n must be specified. If the data set is incomplete, missing values will be first filled based on the mean imputation method.

`clusters` (optional) A vector of length n that specifies the initial cluster memberships of the user when `init_method` is set to "manual". Both numeric and character vectors are acceptable. This argument is NULL by default, so that it is ignored whenever other given initialization methods are chosen.

`outlier_cutoff` (optional) A number between 0 and 1 indicating the percentile cutoff used for outlier detection. This is only relevant for t mixture.

`deriv_ctrl` (optional) A list containing arguments to control the numerical procedures for calculating the first and second derivatives. Some values are suggested by default. Refer to functions `grad` and `hessian` under the package `numDeriv` for more information.

`progress` (optional) A logical value indicating whether the fitting progress should be displayed; TRUE by default.

Details

Beside the generalized hyperbolic distribution, the function can fit mixture via its special and limiting cases. Available distributions include

- GH - Generalized Hyperbolic
- NIG - Normal-Inverse Gaussian
- SNIG - Symmetric Normal-Inverse Gaussian
- SC - Skew-Cauchy
- C - Cauchy
- St - Skew-*t*
- t - Student's *t*
- N - Normal or Gaussian
- SGH - Symmetric Generalized Hyperbolic
- HUM- Hyperbolic Univariate Marginals
- H - Hyperbolic
- SH - Symmetric Hyperbolic

Value

An object of class `MixtureMissing` with:

`model`	The model used to fit the data set.
`pi`	Mixing proportions.
`mu`	Component mean vectors (location).
`Sigma`	Component covariance matrices (dispersion).
`beta`	Component skewness vectors. Only available if `model` is GH, NIG, SNIG, SC, SGH, HUM, H, or SH; NULL otherwise.
`lambda`	Component index parameters. Only available if `model` is GH, NIG, SNIG, SGH, HUM, H, or SH; NULL otherwise.
`omega`	Component concentration parameters. Only available if `model` is GH, NIG, SNIG, SGH, HUM, H, or SH; NULL otherwise.
`df`	Component degrees of freedom. Only available if `model` is St or t; NULL otherwise.
`z_tilde`	An n by G matrix where each row indicates the expected probabilities that the corresponding observation belongs to each cluster.
`clusters`	A numeric vector of length n indicating cluster memberships determined by the model.
`outliers`	A logical vector of length n indicating observations that are outliers. Only available if `model` is t
`data`	The original data set if it is complete; otherwise, this is the data set with missing values imputed by appropriate expectations.
`complete`	A logical vector of length n indicating which observation(s) have no missing values.
`npar`	The breakdown of the number of parameters to estimate.
`max_iter`	Maximum number of iterations allowed in the EM algorithm.
`iter_stop`	The actual number of iterations needed when fitting the data set.
`final_loglik`	The final value of log-likelihood.
`loglik`	All the values of log-likelihood.
`AIC`	Akaike information criterion.
`BIC`	Bayesian information criterion.
`KIC`	Kullback information criterion.
`KICc`	Corrected Kullback information criterion.
`AIC3`	Modified AIC.
`CAIC`	Bozdogan's consistent AIC.
`AICc`	Small-sample version of AIC.
`ent`	Entropy.
`ICL`	Integrated Completed Likelihood criterion.
`AWE`	Approximate weight of evidence.
`CLC`	Classification likelihood criterion.
`init_method`	The initialization method used in model fitting.

References

Browne, R. P. and McNicholas, P. D. (2015). A mixture of generalized hyperbolic distributions. *Canadian Journal of Statistics*, 43(2):176–198.

Wei, Y., Tang, Y., and McNicholas, P. D. (2019). Mixtures of generalized hyperbolic distributions and mixtures of skew-*t* distributions for model-based clustering with incomplete data. *Computational Statistics & Data Analysis*, 130:18–41.

Examples

```
data('bankruptcy')

#++++ With no missing values ++++#

X <- bankruptcy[, 2:3]
mod <- MGHM(X, G = 2, init_method = 'kmedoids', max_iter = 10)

summary(mod)
plot(mod)

#++++ With missing values ++++#

set.seed(1234)

X <- hide_values(bankruptcy[, 2:3], prop_cases = 0.1)
mod <- MGHM(X, G = 2, init_method = 'kmedoids', max_iter = 10)

summary(mod)
plot(mod)
```

---

plot.MixtureMissing *Mixture Missing Plotting*

Description

Provide a parallel plot of up to the first 10 variables of a multivariate data sets, and a line plot showing log-likelihood values at every iteration during the EM algorithm. When applicable, pairwise scatter plots highlighting outliers denoted by triangles and/or observations whose values are missing but are replaced by expectations obtained in the EM algorithm will be included.

Usage

```
## S3 method for class 'MixtureMissing'
plot(x, ...)
```

Arguments

x	A MixtureMissing object.
...	Arguments to be passed to methods, such as graphical parameters.

Value

No return value, called to visualize the fitted model's results

Examples

```
data('auto')

#++++ With no missing values ++++#

X <- auto[, c('engine_size', 'city_mpg', 'highway_mpg')]
mod <- MCNM(X, G = 2, init_method = 'kmedoids', max_iter = 10)
plot(mod)

#++++ With missing values ++++#

X <- auto[, c('normalized_losses', 'horsepower', 'highway_mpg', 'price')]
mod <- MCNM(X, G = 2, init_method = 'kmedoids', max_iter = 10)
plot(mod)
```

---

select_mixture *Mixture Model Selection*

Description

Fit mixtures via various distributions and decide the best model based on a given information criterion. The distributions include multivariate contaminated normal, multivariate generalized hyperbolic, special and limiting cases of multivariate generalized hyperbolic.

Usage

```
select_mixture(
  X,
  G,
 model = c("CN", "GH", "NIG", "SNIG", "SC", "C", "St", "t", "N", "SGH", "HUM", "H",
    "SH"),
 criterion = c("BIC", "AIC", "KIC", "KICc", "AIC3", "CAIC", "AICc", "ICL", "AWE", "CLC"),
  max_iter = 20,
  epsilon = 0.01,
  init_method = c("kmedoids", "kmeans", "hierarchical", "manual"),
  clusters = NULL,
  eta_min = 1.001,
```

```
  outlier_cutoff = 0.95,
 deriv_ctrl = list(eps = 1e-08, d = 1e-04, zero.tol = sqrt(.Machine$double.eps/7e-07), r
    = 6, v = 2, show.details = FALSE),
  progress = TRUE
)
```

Arguments

X	An n x d matrix or data frame where n is the number of observations and d is the number of variables.
G	The number of clusters, which must be at least 1. If G = 1, then both init_method and clusters are ignored.
model	A vector of character strings indicating the mixture model(s) to be fitted. See the details section for a list of available distributions. However, all distributions will be considered by default.
criterion	A character string indicating the information criterion for model selection. See the details section for a list of available information criteria.
max_iter	(optional) A numeric value giving the maximum number of iterations each EM algorithm is allowed to use; 20 by default.
epsilon	(optional) A number specifying the epsilon value for the Aitken-based stopping criterion used in the EM algorithm: 0.01 by default.
init_method	(optional) A string specifying the method to initialize the EM algorithm. "kmedoids" clustering is used by default. Alternative methods include "kmeans", "hierarchical", and "manual". When "manual" is chosen, a vector clusters of length n must be specified. If the data set is incomplete, missing values will be first filled based on the mean imputation method.
clusters	(optional) A vector of length n that specifies the initial cluster memberships of the user when init_method is set to "manual". Both numeric and character vectors are acceptable. This argument is NULL by default, so that it is ignored whenever other given initialization methods are chosen.
eta_min	(optional) A numeric value close to 1 to the right specifying the minimum value of eta; 1.001 by default. This is only relevant for CN mixture
outlier_cutoff	(optional) A number between 0 and 1 indicating the percentile cutoff used for outlier detection. This is only relevant for t mixture.
deriv_ctrl	(optional) A list containing arguments to control the numerical procedures for calculating the first and second derivatives. Some values are suggested by default. Refer to functions grad and hessian under the package numDeriv for more information.
progress	(optional) A logical value indicating whether the fitting progress should be displayed; TRUE by default.

Details

The function can fit mixtures via the contaminated normal distribution, generalized hyperbolic distribution, and special and limiting cases of the generalized hyperbolic distribution. Available distributions include

- CN - Contaminated Normal
- GH - Generalized Hyperbolic
- NIG - Normal-Inverse Gaussian
- SNIG - Symmetric Normal-Inverse Gaussian
- SC - Skew-Cauchy
- C - Cauchy
- St - Skew-*t*
- t - Student's *t*
- N - Normal or Gaussian
- SGH - Symmetric Generalized Hyperbolic
- HUM- Hyperbolic Univariate Marginals
- H - Hyperbolic
- SH - Symmetric Hyperbolic

Available information criteria include

- AIC - Akaike information criterion
- BIC - Bayesian information criterion
- KIC - Kullback information criterion
- KICc - Corrected Kullback information criterion
- AIC3 - Modified AIC
- CAIC - Bozdogan's consistent AIC
- AICc - Small-sample version of AIC
- ICL - Integrated Completed Likelihood criterion
- AWE - Approximate weight of evidence
- CLC - Classification likelihood criterion

Value

If the best model is CN, the function returns an object of class `MixtureMissing` with

`model`	The model used to fit the data set.
`pi`	Mixing proportions.
`mu`	Component mean vectors (location).
`Sigma`	Component covariance matrices (dispersion).
`alpha`	Component proportions of good observations.
`eta`	Component degrees of contamination.
`z_tilde`	An n by G matrix where each row indicates the expected probabilities that the corresponding observation belongs to each cluster.
`v_tilde`	An n by G matrix where each row indicates the expected probabilities that the corresponding observation is good with respect to each cluster.

clusters	A numeric vector of length n indicating cluster memberships determined by the model.
outliers	A logical vector of length n indicating observations that are outliers.
data	The original data set if it is complete; otherwise, this is the data set with missing values imputed by appropriate expectations.
complete	A logical vector of length n indicating which observation(s) have no missing values.
npar	The breakdown of the number of parameters to estimate.
max_iter	Maximum number of iterations allowed in the EM algorithm.
iter_stop	The actual number of iterations needed when fitting the data set.
final_loglik	The final value of log-likelihood.
loglik	All the values of log-likelihood.
AIC	Akaike information criterion.
BIC	Bayesian information criterion.
KIC	Kullback information criterion.
KICc	Corrected Kullback information criterion.
AIC3	Modified AIC.
CAIC	Bozdogan's consistent AIC.
AICc	Small-sample version of AIC.
ent	Entropy.
ICL	Integrated Completed Likelihood criterion.
AWE	Approximate weight of evidence.
CLC	Classification likelihood criterion.
init_method	The initialization method used in model fitting.

If the best model is GH, NIG, SNIG, SC, C, St, t, N, SGH, HUM, H, or SH, the function returns an object of class `MixtureMissing` with

model	The model used to fit the data set.
pi	Mixing proportions.
mu	Component mean vectors (location).
Sigma	Component covariance matrices (dispersion).
beta	Component skewness vectors. Only available if `model` is GH, NIG, SNIG, SC, SGH, HUM, H, or SH; NULL otherwise.
lambda	Component index parameters. Only available if `model` is GH, NIG, SNIG, SGH, HUM, H, or SH; NULL otherwise.
omega	Component concentration parameters. Only available if `model` is GH, NIG, SNIG, SGH, HUM, H, or SH; NULL otherwise.
df	Component degrees of freedom. Only available if `model` is St or t; NULL otherwise.

z_tilde	An n by G matrix where each row indicates the expected probabilities that the corresponding observation belongs to each cluster.
clusters	A numeric vector of length n indicating cluster memberships determined by the model.
outliers	A logical vector of length n indicating observations that are outliers. Only available if model is t
data	The original data set if it is complete; otherwise, this is the data set with missing values imputed by appropriate expectations.
complete	A logical vector of length n indicating which observation(s) have no missing values.
npar	The breakdown of the number of parameters to estimate.
max_iter	Maximum number of iterations allowed in the EM algorithm.
iter_stop	The actual number of iterations needed when fitting the data set.
final_loglik	The final value of log-likelihood.
loglik	All the values of log-likelihood.
AIC	Akaike information criterion.
BIC	Bayesian information criterion.
KIC	Kullback information criterion.
KICc	Corrected Kullback information criterion.
AIC3	Modified AIC.
CAIC	Bozdogan's consistent AIC.
AICc	Small-sample version of AIC.
ent	Entropy.
ICL	Integrated Completed Likelihood criterion.
AWE	Approximate weight of evidence.
CLC	Classification likelihood criterion.
init_method	The initialization method used in model fitting.

References

Browne, R. P. and McNicholas, P. D. (2015). A mixture of generalized hyperbolic distributions. *Canadian Journal of Statistics*, 43(2):176–198.

Wei, Y., Tang, Y., and McNicholas, P. D. (2019). Mixtures of generalized hyperbolic distributions and mixtures of skew-*t* distributions for model-based clustering with incomplete data. *Computational Statistics & Data Analysis*, 130:18–41.

Examples

```
data('bankruptcy')

#++++ With no missing values ++++#
```

```
X <- bankruptcy[, 2:3]
mod <- select_mixture(X, G = 2, model = c('CN', 'GH', 'St'), criterion = 'BIC', max_iter = 10)

summary(mod)
plot(mod)

#++++ With missing values ++++#

set.seed(1234)

X <- hide_values(bankruptcy[, 2:3], prop_cases = 0.1)
mod <- select_mixture(X, G = 2, model = c('CN', 'GH', 'St'), criterion = 'BIC', max_iter = 10)

summary(mod)
plot(mod)
```

`summary.MixtureMissing` *Summary for Mixture Missing*

Description

Summarizes main information regarding a `MixtureMissing` object.

Usage

```
## S3 method for class 'MixtureMissing'
summary(object, ...)
```

Arguments

`object`	A `MixtureMissing` object.
`...`	Arguments to be passed to methods, such as graphical parameters.

Details

Information includes the model used to fit the data set, initialization method, clustering table, total outliers, outliers per cluster, mixing proportions, component means and variances, final log-likelihood value, information criteria.

Value

No return value, called to summarize the fitted model's results

Examples

```
#++++ With no missing values ++++#

X <- auto[, c('horsepower', 'highway_mpg', 'price')]
mod <- MCNM(X, G = 2, init_method = 'kmedoids', max_iter = 10)
# summary(mod)

#++++ With missing values ++++#

X <- auto[, c('normalized_losses', 'horsepower', 'highway_mpg', 'price')]
mod <- MCNM(X, G = 2, init_method = 'kmedoids', max_iter = 10)
summary(mod)
```

`UScost` *US Cost of Living Indices in 2019 Data Set*

Description

The data set contains the 2019 cost of living indices of 50 states in five different categories: grocery, housing, transportation, utilities, and miscellaneous (Washington DC is not included). The indices are calculated by first determining the average cost of living in the United States to be used as a baseline set at 100. States are then measured against this baseline. For example, a state with a cost of living index of 200 is twice as expensive as the national average.

Usage

```
UScost
```

Format

A data frame with 50 rows and 7 variables. There are no missing values

Abbr State abbreviation.

State State name.

Grocery Grocery index.

Housing Housing index.

Utilities Utilities index

Transportation Transporation index.

Misc Miscellaneous index

Source

https://worldpopulationreview.com

Index