# Package 'MixtureMissing'

October 12, 2022

**Type** Package

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

**Version** 1.0.2

**Description** Implementation of robust model based cluster analysis with missing data. The models used are: Multivariate Contaminated Normal Mixtures (MCNM), Multivariate Student's t Mixtures (MtM), and Multivariate Normal Mixtures (MNM) for data sets with missing values at random.
See ``Model-Based Clustering and Outlier Detection with Missing Data'' by Hung Tong and Cristina Tortora (2022) <doi:10.1007/s11634-021-00476-1>.

**Imports** ContaminatedMixt (>= 1.3.4.1), mvtnorm (>= 1.1-2), mnormt (>= 2.0.2), cluster (>= 2.1.2), rootSolve (>= 1.8.2.2), MASS (>= 7.3)

**Suggests** mice (>= 3.10.0)

**License** GPL (>= 2)

**Encoding** UTF-8

**LazyData** true

**Repository** CRAN

**RoxygenNote** 7.1.2

**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** 2022-01-30 23:00:04 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

CNM *Contaminated Normal Mixture (CNM)*

## Description

Carries out model-based clustering using a contaminated normal mixture (CNM) for complete univariate data set.

## Usage

```
CNM(
  X,
  G,
  max_iter = 20,
  epsilon = 0.01,
 init_method = c("kmedoids", "kmeans", "hierarchical", "manual", "soft", "hard"),
  equal_prop = FALSE,
  unit_var = FALSE,
  eta_min = 1.001,
  show_progress = TRUE,
  manual_clusters = NULL
)
```

## Arguments

| | |
|---|---|
| X | A vector of $n$ observations. |
| G | The number of clusters. |
| 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", "manual", "soft", "hard". When "manual" is chosen, a vector `manual_clusters` of length $n$ must be specified. |
| equal_prop | (optional) A logical value indicating whether mixing proportions should be equal at initialization of the EM algorithm; FALSE by default. |
| unit_var | (optional) A logical value indicating whether variance should be initialized as 1; FALSE by default. |
| eta_min | (optional) A numeric value close to 1 to the right specifying the minimum value of eta; 1.001 by default. |
| show_progress | (optional) A logical value indicating whether the fitting progress should be displayed; TRUE by default. |

`manual_clusters`
: 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.

## Value

An object of class `MixtureMissing` with:

`pi` Mixing proportions.

`mu` Component means.

`sigma` Component variances.

`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 outlying 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 ouliers.

`data` The original data set.

`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_lik` The final value of likelihood.

`final_loglik` The final value of log-likelihood.

`lik` All the values of 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.

## Examples

```
set.seed(1234)

mod <- CNM(iris$Sepal.Length, G = 3, init_method = 'kmedoids', max_iter = 30)

plot(mod)
summary(mod)
```

---

`cnm_close_100` *A Mixture of Two Close Contaminated Normal Distributions - 100 Observations*

---

## Description

A simulated mixture of two close contaminated normal distributions. Refer to Punzo and McNicholas (2016) for more information about the underlying distribution that generates this data set.

## Usage

```
cnm_close_100
```

## Format

A matrix with 100 rows and 3 variables. The first two variables make the bivariate data, while the last variable refers to cluster memberships. The first 30 rows belong to cluster 1, and the last 70 rows belong to cluster 2

**d1** variable 1.

**d2** variable 2.

**cluster** cluster memberships

## Source

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

## cnm_close_500 *A Mixture of Two Close Contaminated Normal Distributions - 500 Observations*

### Description

A simulated mixture of two close contaminated normal distributions. Refer to Punzo and McNicholas (2016) for more information about the underlying distribution that generates this data set.

### Usage

```
cnm_close_500
```

### Format

A matrix with 500 rows and 3 variables. The first two variables make the bivariate data, while the last variable refers to cluster memberships. The first 150 rows belong to cluster 1, and the last 350 rows belong to cluster 2

**d1** variable 1.

**d2** variable 2.

**cluster** cluster memberships

### Source

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

## cnm_far_100 *A Mixture of Two Far Contaminated Normal Distributions - 100 Observations*

### Description

A simulated mixture of two far contaminated normal distributions. Refer to Punzo and McNicholas (2016) for more information about the underlying distribution that generates this data set.

### Usage

```
cnm_far_100
```

## Format

A matrix with 100 rows and 3 variables. The first two variables make the bivariate data, while the last variable refers to cluster memberships. The first 30 rows belong to cluster 1, and the last 70 rows belong to cluster 2

**d1** variable 1.

**d2** variable 2.

**cluster** cluster memberships

## Source

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

---

`cnm_far_500` *A Mixture of Two Far Contaminated Normal Distributions - 500 Observations*

---

## Description

A simulated mixture of two far contaminated normal distributions. Refer to Punzo and McNicholas (2016) for more information about the underlying distribution that generates this data set.

## Usage

```
cnm_far_500
```

## Format

A matrix with 500 rows and 3 variables. The first two variables make the bivariate data, while the last variable refers to cluster memberships. The first 150 rows belong to cluster 1, and the last 350 rows belong to cluster 2

**d1** variable 1.

**d2** variable 2.

**cluster** cluster memberships

## Source

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

`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, CNM, MtM, and tM 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*

---

## 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", "soft", "hard"),
  manual_clusters = 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. Alternately, `X` can be a vector of $n$ observations. |
| `G` | The number of clusters. |
| `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", "soft", "hard". When "manual" is chosen, a vector `manual_clusters` of length $n$ must be specified. |
| `manual_clusters` | 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. |

## Details

Available heuristic methods include k-medoids clustering, k-means clustering, hierarchical clustering, soft and hard clustering. Alternately, the user can also enter pre-specified cluster memberships, making other initialization methods possible.

## Value

A list with the following slots:

| | |
|---|---|
| `z` | Mapping probabilities in the form of an $n$ by $G$ matrix. |
| `clusters` | An numeric vector with values from 1 to $G$ indicating initial cluster memberships. |
| `pi` | Component mixing proportions. |
| `mu` | If `X` is a matrix or data frame, `mu` is an $G$ by $d$ matrix where each row is the component mean vector. If `X` is a vector, `mu` is a vector of $G$ component means. |
| `sigma` | If `X` is a matrix or data frame, `sigma` is a $G$-dimensional array where each $d$ by $d$ matrix is the component covariance matrix. If `X` is a vector, `sigma` is a vector of $G$ component variances. |

## 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 ++++#

set.seed(1234)

init <- initialize_clusters(iris[1:4], G = 3)
init <- initialize_clusters(iris[1:4], G = 3, init_method = 'kmeans')
init <- initialize_clusters(iris[1:4], G = 3, init_method = 'hierarchical')
init <- initialize_clusters(iris[1:4], G = 3, init_method = 'soft')
init <- initialize_clusters(iris[1:4], G = 3, init_method = 'hard')

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

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

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

init$z
init$pi
init$mu
init$sigma

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

MCNM *Multivariate Contaminated Normal Mixture (MCNM)*

## 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. When using this function, 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", "soft", "hard"),
  equal_prop = FALSE,
  identity_cov = FALSE,
  eta_min = 1.001,
  show_progress = TRUE,
  manual_clusters = 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. |
| G | The number of clusters. |
| 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", "manual", "soft", "hard". When "manual" is chosen, a vector manual_clusters of length $n$ must be specified. |
| equal_prop | (optional) A logical value indicating whether mixing proportions should be equal at initialization of the EM algorithm; FALSE by default. |
| identity_cov | (optional) A logical value indicating whether covariance matrices should be initialized as identity matrices; FALSE by default. |
| eta_min | (optional) A numeric value close to 1 to the right specifying the minimum value of eta; 1.001 by default. |

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

`manual_clusters`
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.

## Value

An object of class `MixtureMissing` with:

| | |
|---|---|
| `model` | The model used to fit the data set |
| `pi` | Mixing proportions. |
| `mu` | Component mean vectors. |
| `sigma` | Component covariance matrices. |
| `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 outlying 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 ouliers. |
| `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_lik` | The final value of likelihood. |
| `final_loglik` | The final value of log-likelihood. |
| `lik` | All the values of 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('nm_5_noise_close_100')

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

X <- nm_5_noise_close_100[, 1:2]
mod <- MCNM(X, G = 2, init_method = 'kmedoids', max_iter = 10)

summary(mod)
plot(mod)

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

set.seed(1234)

X <- hide_values(nm_5_noise_close_100[, 1:2], prop_cases = 0.1)
mod <- MCNM(X, G = 2, init_method = 'kmedoids', max_iter = 10)

summary(mod)
plot(mod)
```

---

# MNM *Multivariate Normal Mixture (MNM)*

---

## Description

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

## Usage

```
MNM(
  X,
  G,
  max_iter = 20,
  epsilon = 0.01,
 init_method = c("kmedoids", "kmeans", "hierarchical", "manual", "soft", "hard"),
  equal_prop = FALSE,
  identity_cov = FALSE,
  show_progress = TRUE,
  manual_clusters = 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. |
| G | The number of clusters. |
| 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", "manual", "soft", "hard". When "manual" is chosen, a vector manual_clusters of length $n$ must be specified. |
| equal_prop | (optional) A logical value indicating whether mixing proportions should be equal at initialization of the EM algorithm; FALSE by default. |
| identity_cov | (optional) A logical value indicating whether covariance matrices should be initialized as identity matrices; FALSE by default. |
| show_progress | (optional) A logical value indicating whether the fitting progress should be displayed; TRUE by default. |
| manual_clusters | 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. |

## Value

An object of class MixtureMissing with:

| | |
|---|---|
| model | The model used to fit the data set |
| pi | Mixing proportions. |
| mu | Component mean vectors. |
| sigma | Component covariance matrices. |

| | |
|---|---|
| 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. |
| 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_lik | The final value of likelihood. |
| final_loglik | The final value of log-likelihood. |
| lik | All the values of 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

Wolfe, J. H. (1965). A computer program for the maximum likelihood analysis of types. Technical report, NAVAL PERSONNEL RESEARCH ACTIVITY SAN DIEGO United States.

Ghahramani, Z. and Jordan, M. I. (1995). Learning from incomplete data.

## Examples

```
data('nm_5_noise_close_100')

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

X <- nm_5_noise_close_100[, 1:2]
mod <- MNM(X, G = 2, init_method = 'kmedoids', max_iter = 10)
```

```
summary(mod)
plot(mod)

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

set.seed(1234)

X <- hide_values(nm_5_noise_close_100[, 1:2], prop_cases = 0.1)
mod <- MNM(X, G = 2, init_method = 'kmedoids', max_iter = 10)

summary(mod)
plot(mod)
```

---

MtM *Multivariate* t *Mixture (M*t*M)*

---

## Description

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

## Usage

```
MtM(
  X,
  G,
  max_iter = 20,
  epsilon = 0.01,
 init_method = c("kmedoids", "kmeans", "hierarchical", "manual", "soft", "hard"),
  equal_prop = FALSE,
  identity_cov = FALSE,
  df0 = rep(10, G),
  outlier_cutoff = 0.95,
  show_progress = TRUE,
  manual_clusters = 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. |
| G | The number of clusters. |
| 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", "manual", "soft", "hard". When "manual" is chosen, a vector manual_clusters of length $n$ must be specified. |
| equal_prop | (optional) A logical value indicating whether mixing proportions should be equal at initialization of the EM algorithm; FALSE by default. |
| identity_cov | (optional) A logical value indicating whether covariance matrices should be initialized as identity matrices; FALSE by default. |
| df0 | (optional) Starting value of component degrees of freedom; 10 by default. |
| outlier_cutoff | (optional) A number between 0 and 1 indicating the percentile cutoff used for outlier detection. |
| show_progress | (optional) A logical value indicating whether the fitting progress should be displayed; TRUE by default. |
| manual_clusters | 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. |

## Value

An object of class MixtureMissing with:

| | |
|---|---|
| model | The model used to fit the data set |
| pi | Mixing proportions. |
| mu | Component mean vectors. |
| sigma | Component covariance matrices. |
| df | Component degrees of freedom. |
| 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 ouliers. |
| 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_lik | The final value of likelihood. |

| | |
|---|---|
| final_loglik | The final value of log-likelihood. |
| lik | All the values of 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

Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the *t* distribution. *Statistics and computing, 10*(4):339-348.

Wang, H., Zhang, Q., Luo, B., and Wei, S. (2004). Robust mixture modelling using multivariatet-distribution with missing information. *Pattern Recognition Letters, 25*(6):701-710.

## Examples

```
data('nm_5_noise_close_100')

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

X <- nm_5_noise_close_100[, 1:2]
mod <- MtM(X, G = 2, init_method = 'kmedoids', max_iter = 10)

summary(mod)
plot(mod)

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

set.seed(1234)

X <- hide_values(nm_5_noise_close_100[, 1:2], prop_cases = 0.1)
mod <- MtM(X, G = 2, init_method = 'kmedoids', max_iter = 10)

summary(mod)
plot(mod)
```

NM *Normal Mixture (NM)*

## Description

Carries out model-based clustering using a normal mixture (NM) for complete univariate data set.

## Usage

```
NM(
  X,
  G,
  max_iter = 20,
  epsilon = 0.01,
 init_method = c("kmedoids", "kmeans", "hierarchical", "manual", "soft", "hard"),
  equal_prop = TRUE,
  unit_var = FALSE,
  show_progress = TRUE,
  manual_clusters = NULL
)
```

## Arguments

| | |
|---|---|
| X | A vector of $n$ observations. |
| G | The number of clusters. |
| 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", "manual", "soft", "hard". When "manual" is chosen, a vector manual_clusters of length $n$ must be specified. |
| equal_prop | (optional) A logical value indicating whether mixing proportions should be equal at initialization of the EM algorithm; FALSE by default. |
| unit_var | (optional) A logical value indicating whether variance should be initialized as 1; FALSE by default. |
| show_progress | (optional) A logical value indicating whether the fitting progress should be displayed; TRUE by default. |
| manual_clusters | 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. |

## Value

An object of class `MixtureMissing` with:

| | |
|---|---|
| `pi` | Mixing proportions. |
| `mu` | Component means. |
| `sigma` | Component variances. |
| `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. |
| `data` | The original data set. |
| `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_lik` | The final value of likelihood. |
| `final_loglik` | The final value of log-likelihood. |
| `lik` | All the values of 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

Wolfe, J. H. (1965). A computer program for the maximum likelihood analysis of types. Technical report, NAVAL PERSONNEL RESEARCH ACTIVITY SAN DIEGO United States.

## Examples

```
set.seed(1234)

mod <- NM(iris$Sepal.Length, G = 3, init_method = 'kmedoids', max_iter = 30)

plot(mod)
summary(mod)
```

---

`nm_1_noise_close_100` *A Mixture of Two Close Normal Distributions with 1 by High Atypical Points - 100 Observations*

---

## Description

A simulated mixture of two close normal distributions with 1 substituted by high atypical points. Refer to Punzo and McNicholas (2016) for more information about the underlying distribution that generates this data set.

## Usage

```
nm_1_noise_close_100
```

## Format

A matrix with 100 rows and 4 variables. The first two variables make the bivariate data; the last variable refers to cluster memberships; and the last variable refers to outlier indication (1 means outlier, 0 otherwise). The first 30 rows belong to cluster 1, and the last 70 rows belong to cluster 2

**d1** variable 1.

**d2** variable 2.

**cluster** cluster memberships

**outlier** outlier indication

## Source

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

## nm_1_noise_close_500 *A Mixture of Two Close Normal Distributions with 1 by High Atypical Points - 500 Observations*

### Description

A simulated mixture of two close normal distributions with 1 substituted by high atypical points. Refer to Punzo and McNicholas (2016) for more information about the underlying distribution that generates this data set.

### Usage

```
nm_1_noise_close_500
```

### Format

A matrix with 500 rows and 4 variables. The first two variables make the bivariate data; the last variable refers to cluster memberships; and the last variable refers to outlier indication (1 means outlier, 0 otherwise). The first 150 rows belong to cluster 1, and the last 350 rows belong to cluster 2

**d1** variable 1.

**d2** variable 2.

**cluster** cluster memberships

**outlier** outlier indication

### Source

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

## nm_1_noise_far_100 *A Mixture of Two Far Normal Distributions with 1 by High Atypical Points - 100 Observations*

### Description

A simulated mixture of two far normal distributions with 1 substituted by high atypical points. Refer to Punzo and McNicholas (2016) for more information about the underlying distribution that generates this data set.

### Usage

```
nm_1_noise_far_100
```

## Format

A matrix with 100 rows and 3 variables. The first two variables make the bivariate data, while the last variable refers to cluster memberships. The first 30 rows belong to cluster 1, and the last 70 rows belong to cluster 2

**d1** variable 1.

**d2** variable 2.

**cluster** cluster memberships

## Source

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

---

`nm_1_noise_far_500` *A Mixture of Two Far Normal Distributions with 1 by High Atypical Points - 500 Observations*

---

## Description

A simulated mixture of two far normal distributions with 1 substituted by high atypical points. Refer to Punzo and McNicholas (2016) for more information about the underlying distribution that generates this data set.

## Usage

```
nm_1_noise_far_500
```

## Format

A matrix with 500 rows and 3 variables. The first two variables make the bivariate data, while the last variable refers to cluster memberships. The first 150 rows belong to cluster 1, and the last 350 rows belong to cluster 2

**d1** variable 1.

**d2** variable 2.

**cluster** cluster memberships

## Source

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

nm_5_noise_close_100 *A Mixture of Two Close Normal Distributions with 5 by Noise - 100 Observations*

## Description

A simulated mixture of two close normal distributions with 1 substituted by noise. Refer to Punzo and McNicholas (2016) for more information about the underlying distribution that generates this data set.

## Usage

```
nm_5_noise_close_100
```

## Format

A matrix with 100 rows and 4 variables. The first two variables make the bivariate data; the last variable refers to cluster memberships; and the last variable refers to outlier indication (1 means outlier, 0 otherwise). The first 30 rows belong to cluster 1, and the last 70 rows belong to cluster 2

**d1** variable 1.

**d2** variable 2.

**cluster** cluster memberships

**outlier** outlier indication

## Source

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

nm_5_noise_close_500 *A Mixture of Two Close Normal Distributions with 5 by Noise - 500 Observations*

## Description

A simulated mixture of two close normal distributions with 1 substituted by noise. Refer to Punzo and McNicholas (2016) for more information about the underlying distribution that generates this data set.

## Usage

```
nm_5_noise_close_500
```

## Format

A matrix with 500 rows and 4 variables. The first two variables make the bivariate data; the last variable refers to cluster memberships; and the last variable refers to outlier indication (1 means outlier, 0 otherwise). The first 150 rows belong to cluster 1, and the last 350 rows belong to cluster 2

**d1** variable 1.

**d2** variable 2.

**cluster** cluster memberships

**outlier** outlier indication

## Source

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

---

`nm_5_noise_far_100` *A Mixture of Two Far Normal Distributions with 5 by Noise - 100 Observations*

---

## Description

A simulated mixture of two far normal distributions with 1 substituted by noise. Refer to Punzo and McNicholas (2016) for more information about the underlying distribution that generates this data set.

## Usage

```
nm_5_noise_far_100
```

## Format

A matrix with 100 rows and 4 variables. The first two variables make the bivariate data; the last variable refers to cluster memberships; and the last variable refers to outlier indication (1 means outlier, 0 otherwise). The first 30 rows belong to cluster 1, and the last 70 rows belong to cluster 2

**d1** variable 1.

**d2** variable 2.

**cluster** cluster memberships

**outlier** outlier indication

## Source

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

## nm_5_noise_far_500 *A Mixture of Two Far Normal Distributions with 5 by Noise - 500 Observations*

### Description

A simulated mixture of two far normal distributions with 1 substituted by noise. Refer to Punzo and McNicholas (2016) for more information about the underlying distribution that generates this data set.

### Usage

```
nm_5_noise_far_500
```

### Format

A matrix with 500 rows and 4 variables. The first two variables make the bivariate data; the last variable refers to cluster memberships; and the last variable refers to outlier indication (1 means outlier, 0 otherwise). The first 150 rows belong to cluster 1, and the last 350 rows belong to cluster 2

**d1** variable 1.

**d2** variable 2.

**cluster** cluster memberships

**outlier** outlier indication

### Source

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

## 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 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('nm_5_noise_close_100')

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

X <- nm_5_noise_close_100[, 1:2]
mod <- MCNM(X, G = 2, init_method = 'kmedoids', max_iter = 10)
plot(mod)

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

set.seed(1234)

X <- hide_values(nm_5_noise_close_100[, 1:2], prop_cases = 0.1)
mod <- MCNM(X, G = 2, init_method = 'kmedoids', max_iter = 10)
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.

## Value

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

## Examples

```
data('nm_5_noise_close_100')

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

# X <- nm_5_noise_close_100[, 1:2]
# mod <- MCNM(X, G = 2, init_method = 'kmedoids', max_iter = 10)
# summary(mod)

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

set.seed(1234)

X <- hide_values(nm_5_noise_close_100[, 1:2], prop_cases = 0.1)
mod <- MCNM(X, G = 2, init_method = 'kmedoids', max_iter = 10)
summary(mod)
```

---

# tM t *Mixture (tM)*

---

## Description

Carries out model-based clustering using a *t* mixture (*t*M) for complete univariate data set.

## Usage

```
tM(
  X,
  G,
  max_iter = 20,
  epsilon = 0.01,
  init_method = c("kmedoids", "kmeans", "hierarchical", "manual", "soft", "hard"),
  equal_prop = TRUE,
  unit_var = FALSE,
  df0 = rep(10, G),
  outlier_cutoff = 0.95,
  show_progress = TRUE,
  manual_clusters = NULL
)
```

## Arguments

| | |
|---|---|
| X | A vector of $n$ observations. |
| G | The number of clusters. |
| 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", "manual", "soft", "hard". When "manual" is chosen, a vector manual_clusters of length $n$ must be specified. |
| equal_prop | (optional) A logical value indicating whether mixing proportions should be equal at initialization of the EM algorithm; FALSE by default. |
| unit_var | (optional) A logical value indicating whether variance should be initialized as 1; FALSE by default. |
| df0 | (optional) Starting values of the degrees of freedom; 10 for all clusters by default. |
| outlier_cutoff | (optional) A percentile for outlier detection; 0.95 by default. |
| show_progress | (optional) A logical value indicating whether the fitting progress should be displayed; TRUE by default. |
| manual_clusters | 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. |

## Value

An object of class MixtureMissing with:

| | |
|---|---|
| pi | Mixing proportions. |
| mu | Component means. |
| sigma | Component variances. |
| df | Component degrees of freedom. |
| 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 ouliers. |
| 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_lik | The final value of likelihood. |
| final_loglik | The final value of log-likelihood. |
| lik | All the values of 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

Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the *t* distribution. *Statistics and computing, 10*(4):339-348.

## Examples

```
set.seed(1234)

mod <- tM(iris$Sepal.Length, G = 3, init_method = 'kmedoids', max_iter = 30)

plot(mod)
summary(mod)
```

## tm_close_100 *A Mixture of Two Close Student's t Distributions - 100 Observations*

### Description

A simulated mixture of two close Student's $t$ distributions. Refer to Punzo and McNicholas (2016) for more information about the underlying distribution that generates this data set.

### Usage

```
tm_close_100
```

### Format

A matrix with 100 rows and 3 variables. The first two variables make the bivariate data, while the last variable refers to cluster memberships. The first 30 rows belong to cluster 1, and the last 70 rows belong to cluster 2

**d1** variable 1.

**d2** variable 2.

**cluster** cluster memberships

### Source

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

## tm_close_500 *A Mixture of Two Close Student's t Distributions - 500 Observations*

### Description

A simulated mixture of two close Student's $t$ distributions. Refer to Punzo and McNicholas (2016) for more information about the underlying distribution that generates this data set.

### Usage

```
tm_close_500
```

### Format

A matrix with 500 rows and 3 variables. The first two variables make the bivariate data, while the last variable refers to cluster memberships. The first 150 rows belong to cluster 1, and the last 350 rows belong to cluster 2

**d1** variable 1.

**d2** variable 2.

**cluster** cluster memberships

## Source

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

---

`tm_far_100` *A Mixture of Two Far Student's t Distributions - 100 Observations*

---

## Description

A simulated mixture of two far Student's $t$ distributions. Refer to Punzo and McNicholas (2016) for more information about the underlying distribution that generates this data set.

## Usage

```
tm_far_100
```

## Format

A matrix with 500 rows and 4 variables. The first two variables make the bivariate data; the last variable refers to cluster memberships; and the last variable refers to outlier indication (1 means outlier, 0 otherwise). The first 30 rows belong to cluster 1, and the last 70 rows belong to cluster 2

**d1** variable 1.

**d2** variable 2.

**cluster** cluster memberships

**outlier** outlier indication

## Source

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

---

`tm_far_500` *A Mixture of Two Far Student's t Distributions - 500 Observations*

---

## Description

A simulated mixture of two far Student's $t$ distributions. Refer to Punzo and McNicholas (2016) for more information about the underlying distribution that generates this data set.

## Usage

```
tm_far_500
```

## Format

A matrix with 500 rows and 3 variables. The first two variables make the bivariate data, while the last variable refers to cluster memberships. The first 150 rows belong to cluster 1, and the last 350 rows belong to cluster 2

**d1** variable 1.

**d2** variable 2.

**cluster** cluster memberships

## Source

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

# Index