Package 'modgo'

September 11, 2024

Type Package

Title Mock Data Generation

Version 1.0.1

Date 2024-08-21

Maintainer Georgios Koliopanos <george.koliopanos@cardio-care.ch>

Description Generation of synthetic data from a real dataset using the combination of rank normal inverse transformation with the calculation of correlation matrix <doi:10.1055/a-2048-7692>. Completely artificial data may be generated through the use of Generalized Lambda Distribution and Generalized Poisson Distribution <doi:10.1201/9781420038040>. Quantitative, binary, ordinal categorical, and survival data may be simulated. Functionalities are offered to generate synthetic data sets according to user's needs.

Encoding UTF-8

RoxygenNote 7.3.2

Suggests knitr, rmarkdown

VignetteBuilder knitr

License GPL-3

Depends R (>= 4.1)

Imports ggplot2 (>= 3.4.0), patchwork (>= 1.1.2), wesanderson (>= 0.3.6.9000), Matrix (>= 1.6.1.1), ggcorrplot (>= 0.1.4.1), gridExtra (>= 2.3), psych (>= 2.2.9), GLDEX (>= 2.0.0.9.2), MASS (>= 7.3), gp (>= 1.0), stats, utils, survival

NeedsCompilation no

Author Andreas Ziegler [aut], Francisco Miguel Echevarria [aut], Georgios Koliopanos [cre]

Repository CRAN

Date/Publication 2024-09-11 16:20:02 UTC

Contents

`checkArguments` *Check Arguments*

Description

Check that the arguments are following the corresponding conditions

Usage

```
checkArguments(
  data = NULL,
  ties_method = "max",
  variables = colnames(data),
  bin_variables = NULL,
  categ_variables = NULL,
  count_variables = NULL,
  n_samples = nrow(data),
  sigma = NULL,
  nrep = 100,
  noise_mu = FALSE,
  pertr_vec = NULL,
  change_cov = NULL,
  change_amount = 0,
  seed = 1,
  thresh_var = NULL,
  thresh_force = FALSE,
  var_prop = NULL,
  var_infl = NULL,
```

```
  infl_cov_stable = FALSE,
  tol = 1e-06,
  stop_sim = FALSE,
  new_mean_sd = NULL,
  multi_sugg_prop = NULL,
  generalized_mode = FALSE,
  generalized_mode_model = NULL,
  generalized_mode_lmbds = NULL
)
```

Arguments

data	a data frame containing the data whose characteristics are to be mimicked during the data simulation.
ties_method	Method on how to deal with equal values during rank transformation. Acceptable input:"max","average","min". This parameter is passed by rbi_normal_transform to the parameter ties.method of rank.
variables	a vector of which variables you want to transform. Default:colnames(data)
bin_variables	a character vector listing the binary variables.
categ_variables	a character vector listing the ordinal categorical variables.
count_variables	a character vector listing the count as a sub sub category of categorical variables. Count variables should be part of categorical variables vector. Count variables are treated differently when using gldex to simulate them.
n_samples	Number of rows of each simulated data set. Default is the number of rows of data.
sigma	a covariance matrix of NxN (N= number of variables) provided by the user to bypass the covariance matrix calculations
nrep	number of repetitions.
noise_mu	Logical value if you want to apply noise to multivariate mean. Default: FALSE
pertr_vec	A named vector.Vector's names are the continuous variables that the user want to perturb. Variance of simulated data set mimic original data's variance.
change_cov	change the covariance of a specific pair of variables.
change_amount	the amount of change in the covariance of a specific pair of variables.
seed	A numeric value specifying the random seed. If seed = NA, no random seed is set.
thresh_var	A data frame that contains the thresholds(left and right) of specified variables (1st column: variable names, 2nd column: Left thresholds, 3rd column: Right thresholds)
thresh_force	A logical value indicating if you want to force threshold in case the proportion of samples that can surpass the threshold are less than 10%
var_prop	A named vector that provides a proportion of value=1 for a specific binary variable(=name of the vector) that will be the proportion of this value in the simulated data sets.[this may increase execution time drastically]

`var_infl` A named vector.Vector's names are the continuous variables that the user want to perturb and increase their variance

`infl_cov_stable`
Logical value. If TRUE,perturbation is applied to original data set and simulations values mimic the perturbed original data set.Covariance matrix used for simulation = original data's correlations. If FALSE, perturbation is applied to the simulated data sets.

`tol` A numeric value that set up tolerance(relative to largest variance) for numerical lack of positive-definiteness in Sigma

`stop_sim` A logical value indicating if the analysis should stop before simulation and produce only the correlation matrix

`new_mean_sd` A matrix that contains two columns named "Mean" and "SD" that the user specifies desired Means and Standard Deviations in the simulated data sets for specific continues variables. The variables must be declared as ROWNAMES in the matrix

`multi_sugg_prop`
A named vector that provides a proportion of value=1 for specific binary variables(=name of the vector) that will be the close to the proportion of this value in the simulated data sets.

`generalized_mode`
A logical value indicating if you want to use generalized distribution to simulate your data

`generalized_mode_model`
A matrix that contains two columns named "Variable" and "Model". This matrix can be used only if a generalized_mode_model argument is provided. It specifies what model should be used for each Variable. Model values should be "RMFMKL", "RPRS", "STAR" or a combination of them, e.g. "RMFMKL-RPRS" or "STAR-STAR", in case the use wants a bimodal simulation. The user can select Generalised Poisson model for poisson variables, but this model cannot be included in bimodal simulation.

`generalized_mode_lmbds`
A matrix that contains lmbds values for each of the variables of the data set to be used for either Generalized Lambda Distribution Generalized Poisson Distribution or setting up thresholds

Value

No value, called for checking arguments of `modgo`

Author(s)

Francisco M. Ojeda, George Koliopanos

Cleveland *Cleveland Dataset ('Cleveland')*

Description

Rows: samples (303) x Columns: Variables (11)

Usage

```
data("Cleveland")
```

Format

A data frame

Details

Cleveland Clinic Heart Disease Data set from the University of California in Irvine (UCI) machine learning data repository

Dua, Dheeru, and Casey Graff. 2017. "UCI Machine Learning Repository." University of California, Irvine,School of Information; Computer Sciences. http://archive.ics.uci.edu/ml

Selected 11 variables and impute missing values Imputation method is described in the Supplementary file 1 of the modgo paper

References

Detrano, R. et al. (1989) "International application of a new probability algorithm for the diagnosis of coronary artery disease," *The American Journal of Cardiology*, **64**(5), 304-310.

Examples

```
data("Cleveland", package="modgo")
```

corr_plots *Plots correlation matrix of original and simulated data*

Description

Produces a graphical display of the correlation matrix of the original dataset, a single simulated dataset and also of the average of the correlation matrices across all simulations for an object returned by modgo.

Usage

```
corr_plots(
  Modgo_obj,
  sim_dataset = 1,
  variables = colnames(Modgo_obj[["simulated_data"]][[1]])
)
```

Arguments

Modgo_obj	An object returned by modgo.
sim_dataset	Number indicating the simulated dataset in Modgo_obj to be used in plots.
variables	A character vector indicating the columns in the data to be used in plots.

Value

A patchwork object created by wrap_plots depicting correlation matrices.

Author(s)

Francisco M. Ojeda, George Koliopanos

Examples

```
data("Cleveland",package="modgo")

test_modgo <- modgo(data = Cleveland,
     bin_variables = c("CAD","HighFastBloodSugar","Sex","ExInducedAngina"),
     categ_variables =c("Chestpaintype"))

corr_plots(test_modgo)
```

---

distr_plots *Plots distribution of original and simulated data*

Description

Produces a graphical display of the distribution of the variables of the original dataset and a single simulated dataset for an object returned by modgo.

Usage

```
distr_plots(
  Modgo_obj,
  variables = colnames(Modgo_obj[["original_data"]]),
  sim_dataset = 1,
  wespalette = "Cavalcanti1",
  text_size = 12
)
```

Arguments

Modgo_obj	An object returned by modgo.
variables	A character vector indicating the columns in the data to be used in plots.
sim_dataset	Number indicating the simulated dataset in Modgo_obj to be used in plots.
wespalette	a name of the selected wesanderson color pallet
text_size	a number for the size of the annotation text

Details

For continuous variables box-and-whisker plots are displayed, while categorical variables bar charts are produced.

Value

A ggplot object depicting distribution of different variables.

Author(s)

Andreas Ziegler, Francisco M. Ojeda, George Koliopanos

Examples

```
data("Cleveland",package="modgo")
test_modgo <- modgo(data = Cleveland,
     bin_variables = c("CAD","HighFastBloodSugar","Sex","ExInducedAngina"),
     categ_variables =c("Chestpaintype"))

distr_plots(test_modgo)
```

---

generalizedMatrix *Generalized Lambda and Poisson preparation*

Description

Prepare the four moments matrix for GLD and GPD

Usage

```
generalizedMatrix(
  data,
  variables = colnames(data),
  bin_variables = NULL,
  generalized_mode_model = NULL,
  multi_sugg_prop = NULL
)
```

Arguments

data	a data frame with original variables.
variables	a vector of which variables you want to transform. Default:colnames(data)
bin_variables	a character vector listing the binary variables.
generalized_mode_model	A matrix that contains two columns named "Variables" and "Model". This matrix can be used only if a generalized_mode_model argument is provided. It specifies what model should be used for each Variable. Model values should be "RMFMKL", "RPRS", "STAR" or a combination of them, e.g. "RMFMKL-RPRS" or "STAR-STAR", in case the use wants a bimodal simulation. The user can select Generalized Poisson model for poisson variables, but this model cannot be included in bimodal simulation
multi_sugg_prop	A named vector that provides a proportion of value=1 for specific binary variables(=name of the vector) that will be the close to the proportion of this value in the simulated data sets

Value

A numeric matrix with the four moments for each distribution and a number that corresponds to a GLD model

Author(s)

Francisco M. Ojeda, George Koliopanos

Examples

```
data("Cleveland",package="modgo")
Variables <- c("Age","STDepression")
Model <- c("rprs", "star-rmfmkl")
model_matrix <- cbind(Variables,
                      Model)
test_modgo <- generalizedMatrix(data = Cleveland,
     generalized_mode_model = model_matrix,
     bin_variables = c("CAD","HighFastBloodSugar","Sex","ExInducedAngina"))
```

---

general_transform_inv *Inverse gldex transformation*

Description

Inverse transforms z values of a vector to simulated values driven by the original dataset using Generalized Lambda and Generalized Poisson percentile functions

Usage

```
general_transform_inv(x, data = NULL, n_samples, lmbds)
```

Arguments

x	a vector of z values
data	a data frame with original variables.
n_samples	number of samples you need to produce.
lmbds	a vector with generalized lambdas values

Value

A numeric vector with inverse transformed values

Author(s)

Andreas Ziegler, Francisco M. Ojeda, George Koliopanos

Examples

```
data("Cleveland",package="modgo")
test_rank <- rbi_normal_transform(Cleveland[,1])
test_generalized_lmbds <- generalizedMatrix(Cleveland,
                  bin_variables = c("Sex", "HighFastBloodSugar", "CAD"))
test_inv_rank <- general_transform_inv(x = test_rank,
                  data = Cleveland[,1],
                  n_samples = 100,
                  lmbds = test_generalized_lmbds[,1])
```

---

generate_simulated_data

Generate new data set by using previous correlation matrix

Description

This function is used internally by modgo. It conducts the computation of the correlation matrix of the transformed variables, which are assumed to follow a multivariate normal distribution.

Usage

```
generate_simulated_data(
  data,
  df_sim,
  variables,
  bin_variables,
  categ_variables,
  count_variables,
  n_samples,
  generalized_mode,
  generalized_mode_lmbds,
  multi_sugg_prop,
  pertr_vec,
  var_infl,
  infl_cov_stable
)
```

Arguments

`data`	a data frame with original variables.
`df_sim`	a data frame with simulated values.
`variables`	variables a character vector indicating which columns of `data` should be used.
`bin_variables`	a character vector listing the binary variables.
`categ_variables`	a character vector listing the ordinal categorical variables.
`count_variables`	a character vector listing the count as a sub sub category of categorical variables. Count variables should be part of categorical variables vector. Count variables are treated differently when using gldex to simulate them.
`n_samples`	Number of rows of each simulated data set. Default is the number of rows of `data`.
`generalized_mode`	A logical value indicating if generalized lambda/poisson distributions or set up thresholds will be used to generate the simulated values
`generalized_mode_lmbds`	A matrix that contains lmbds values for each of the variables of the data set to be used for either Generalized Lambda Distribution Generalized Poisson Distribution or setting up thresholds
`multi_sugg_prop`	A named vector that provides a proportion of value=1 for specific binary variables(=name of the vector) that will be the close to the proportion of this value in the simulated data sets.
`pertr_vec`	A named vector.Vector's names are the continuous variables that the user want to perturb. Variance of simulated data set mimic original data's variance.
`var_infl`	A named vector.Vector's names are the continuous variables that the user want to perturb and increase their variance

infl_cov_stable
: Logical value. If TRUE,perturbation is applied to original data set and simulations values mimic the perturbed original data set.Covariance matrix used for simulation = original data's correlations. If FALSE, perturbation is applied to the simulated data sets.

Value

A data frame with simulated values

Author(s)

Francisco M. Ojeda, George Koliopanos

Inverse_transformation_variables

Inverse transform variables

Description

This function is used internally by modgo. It transforms all variables to their original scale.

Usage

```
Inverse_transformation_variables(
  data,
  df_sim,
  variables,
  bin_variables,
  categ_variables,
  count_variables,
  n_samples,
  generalized_mode,
  generalized_mode_lmbds
)
```

Arguments

data	a data frame with original variables.
df_sim	data frame with transformed variables.
variables	variables a character vector indicating which columns of data should be used.
bin_variables	a character vector listing the binary variables.
categ_variables	a character vector listing the ordinal categorical variables.

`count_variables`
: a character vector listing the count as a sub sub category of categorical variables. Count variables should be part of categorical variables vector. Count variables are treated differently when using gldex to simulate them.

`n_samples`
: Number of rows of each simulated data set. Default is the number of rows of `data`.

`generalized_mode`
: A logical value indicating if generalized lambda/poisson distributions or set up thresholds will be used to generate the simulated values

`generalized_mode_lmbds`
: A matrix that contains lambdas values for each of the variables of the data set to be used for either Generalized Lambda Distribution Generalized Poisson Distribution or setting up thresholds

Value

A data frame with all inverse transformed values.

Author(s)

Francisco M. Ojeda, George Koliopanos

`modgo` *MOck Data GeneratiOn*

Description

`modgo` Create mock dataset from a real one by using ranked based inverse normal transformation. Data with perturbed characteristics can be generated.

Usage

```
modgo(
  data,
  ties_method = "max",
  variables = colnames(data),
  bin_variables = NULL,
  categ_variables = NULL,
  count_variables = NULL,
  n_samples = nrow(data),
  sigma = NULL,
  nrep = 100,
  noise_mu = FALSE,
  pertr_vec = NULL,
  change_cov = NULL,
  change_amount = 0,
  seed = 1,
```

```
  thresh_var = NULL,
  thresh_force = FALSE,
  var_prop = NULL,
  var_infl = NULL,
  infl_cov_stable = FALSE,
  tol = 1e-06,
  stop_sim = FALSE,
  new_mean_sd = NULL,
  multi_sugg_prop = NULL,
  generalized_mode = FALSE,
  generalized_mode_model = NULL,
  generalized_mode_lmbds = NULL
)
```

Arguments

data	a data frame containing the data whose characteristics are to be mimicked during the data simulation.
ties_method	Method on how to deal with equal values during rank transformation. Acceptable input:"max","average","min". This parameter is passed by rbi_normal_transform to the parameter ties.method of rank.
variables	a vector of which variables you want to transform. Default:colnames(data)
bin_variables	a character vector listing the binary variables.
categ_variables	a character vector listing the ordinal categorical variables.
count_variables	a character vector listing the count as a sub sub category of categorical variables. Count variables should be part of categorical variables vector. Count variables are treated differently when using gldex to simulate them.
n_samples	Number of rows of each simulated data set. Default is the number of rows of data.
sigma	a covariance matrix of NxN (N= number of variables) provided by the user to bypass the covariance matrix calculations
nrep	number of repetitions.
noise_mu	Logical value if you want to apply noise to multivariate mean. Default: FALSE
pertr_vec	A named vector.Vector's names are the continuous variables that the user want to perturb. Variance of simulated data set mimic original data's variance.
change_cov	change the covariance of a specific pair of variables.
change_amount	the amount of change in the covariance of a specific pair of variables.
seed	A numeric value specifying the random seed. If seed = NA, no random seed is set.
thresh_var	A data frame that contains the thresholds(left and right) of specified variables (1st column: variable names, 2nd column: Left thresholds, 3rd column: Right thresholds)

`thresh_force` A logical value indicating if you want to force threshold in case the proportion of samples that can surpass the threshold are less than 10%

`var_prop` A named vector that provides a proportion of value=1 for a specific binary variable(=name of the vector) that will be the proportion of this value in the simulated data sets.[this may increase execution time drastically]

`var_infl` A named vector.Vector's names are the continuous variables that the user want to perturb and increase their variance

`infl_cov_stable`
Logical value. If TRUE,perturbation is applied to original data set and simulations values mimic the perturbed original data set.Covariance matrix used for simulation = original data's correlations. If FALSE, perturbation is applied to the simulated data sets.

`tol` A numeric value that set up tolerance(relative to largest variance) for numerical lack of positive-definiteness in Sigma

`stop_sim` A logical value indicating if the analysis should stop before simulation and produce only the correlation matrix

`new_mean_sd` A matrix that contains two columns named "Mean" and "SD" that the user specifies desired Means and Standard Deviations in the simulated data sets for specific continues variables. The variables must be declared as ROWNAMES in the matrix

`multi_sugg_prop`
A named vector that provides a proportion of value=1 for specific binary variables(=name of the vector) that will be the close to the proportion of this value in the simulated data sets.

`generalized_mode`
A logical value indicating if generalized lambda/poisson distributions or set up thresholds will be used to generate the simulated values

`generalized_mode_model`
A matrix that contains two columns named "Variable" and "Model". This matrix can be used only if a generalized_mode_model argument is provided. It specifies what model should be used for each Variable. Model values should be "rmfmkl", "rprs", "star" or a combination of them, e.g. "rmfmkl-rprs" or "star-star", in case the use wants a bimodal simulation. The user can select Generalised Poisson model for poisson variables, but this model cannot be included in bimodal simulation

`generalized_mode_lmbds`
A matrix that contains lambdas values for each of the variables of the data set to be used for either Generalized Lambda Distribution Generalized Poisson Distribution or setting up thresholds

Details

Simulated data is generated based on available data. The simulated data mimics the characteristics of the original data. The algorithm used is based on the ranked based inverse normal transformation (Koliopanos et al. (2023)).

Value

A list with the following components:

`simulated_data` A list of data frames containing the simulated data.

`original_data` A data frame with the input data.

`correlations` a list of correlation matrices. The ith element is the correlation matrix for the ith simulated dataset. The `(repn + 1)`the (last) element of the list is the average of the correlation matrices.

`bin_variables` character vector listing the binary variables

`categ_variables`
a character vector listing the ordinal categorical variables

`covariance_matrix`
Covariance matrix used when generating observations from a multivariate normal distribution.

`seed` Random seed used.

`samples_produced`
Number of rows of each simulated dataset.

`sim_dataset_number`
Number of simulated datasets produced.

A list with the following components:

`simulated_data` A list of data frames containing the simulated data.

`original_data` A data frame with the input data.

`correlations` a list of correlation matrices. The ith element is the correlation matrix for the ith simulated dataset. The `(repn + 1)`the (last) element of the list is the average of the correlation matrices.

`bin_variables` character vector listing the binary variables

`categ_variables`
a character vector listing the ordinal categorical variables

`covariance_matrix`
Covariance matrix used when generating observations from a multivariate normal distribution.

`seed` Random seed used.

`samples_produced`
Number of rows of each simulated dataset.

`sim_dataset_number`
Number of simulated datasets produced.

Author(s)

Francisco M. Ojeda, George Koliopanos

References

Koliopanos, G. and Ojeda, F. and Ziegler Andreas (2023), "A simple-to-use R package for mimicking study data by simulations," *Methods Inf Med*.

Examples

```
data("Cleveland",package="modgo")
test_modgo <- modgo(data = Cleveland,
     bin_variables = c("CAD","HighFastBloodSugar","Sex","ExInducedAngina"),
     categ_variables =c("Chestpaintype"))
```

---

modgo_survival *MOck Data GeneratiOn*

Description

modgo_survival Create mock dataset from a real one by using Generalized Lambdas Distributions and by seperating the data set in 2 based in the event status.

Usage

```
modgo_survival(
  data,
  event_variable = NULL,
  time_variable = NULL,
  surv_method = 1,
  ties_method = "max",
  variables = colnames(data),
  bin_variables = NULL,
  categ_variables = NULL,
  count_variables = NULL,
  n_samples = nrow(data),
  sigma = NULL,
  nrep = 100,
  noise_mu = FALSE,
  pertr_vec = NULL,
  change_cov = NULL,
  change_amount = 0,
  seed = 1,
  thresh_var = NULL,
  thresh_force = FALSE,
  var_prop = NULL,
  var_infl = NULL,
  infl_cov_stable = FALSE,
  tol = 1e-06,
  stop_sim = FALSE,
  new_mean_sd = NULL,
  multi_sugg_prop = NULL,
  generalized_mode = TRUE,
  generalized_mode_model = NULL,
  generalized_mode_model_event = "rprs",
  generalized_mode_model_no_event = "rprs",
```

```
  generalized_mode_lmbds = NULL
)
```

Arguments

data	a data frame containing the data whose characteristics are to be mimicked during the data simulation.
event_variable	a character string listing the event variable.
time_variable	a character string listing the time variable.
surv_method	A numeric value that indicates which one of the 2 survival methods will be used. First method(surv_method = 1): Event and no event data sets are using different covariance matrices for the simulation. Second method(surv_method = 2): Event and no event data sets are using the same covariance matrix for the simulation
ties_method	Method on how to deal with equal values during rank transformation. Acceptable input:"max","average","min". This parameter is passed by rbi_normal_transform to the parameter ties.method of rank.
variables	a vector of which variables you want to transform. Default:colnames(data)
bin_variables	a character vector listing the binary variables.
categ_variables	a character vector listing the ordinal categorical variables.
count_variables	a character vector listing the count as a sub sub category of categorical variables. Count variables should be part of categorical variables vector. Count variables are treated differently when using gldex to simulate them.
n_samples	Number of rows of each simulated data set. Default is the number of rows of data.
sigma	a covariance matrix of NxN (N= number of variables) provided by the user to bypass the covariance matrix calculations
nrep	number of repetitions.
noise_mu	Logical value if you want to apply noise to multivariate mean. Default: FALSE
pertr_vec	A named vector.Vector's names are the continuous variables that the user want to perturb. Variance of simulated data set mimic original data's variance.
change_cov	change the covariance of a specific pair of variables.
change_amount	the amount of change in the covariance of a specific pair of variables.
seed	A numeric value specifying the random seed. If seed = NA, no random seed is set.
thresh_var	A data frame that contains the thresholds(left and right) of specified variables (1st column: variable names, 2nd column: Left thresholds, 3rd column: Right thresholds)
thresh_force	A logical value indicating if you want to force threshold in case the proportion of samples that can surpass the threshold are less than 10%

`var_prop` A named vector that provides a proportion of value=1 for a specific binary variable(=name of the vector) that will be the proportion of this value in the simulated data sets.[this may increase execution time drastically]

`var_infl` A named vector.Vector's names are the continuous variables that the user want to perturb and increase their variance

`infl_cov_stable`
Logical value. If TRUE,perturbation is applied to original data set and simulations values mimic the perturbed original data set.Covariance matrix used for simulation = original data's correlations. If FALSE, perturbation is applied to the simulated data sets.

`tol` A numeric value that set up tolerance(relative to largest variance) for numerical lack of positive-definiteness in Sigma

`stop_sim` A logical value indicating if the analysis should stop before simulation and produce only the correlation matrix

`new_mean_sd` A matrix that contains two columns named "Mean" and "SD" that the user specifies desired Means and Standard Deviations in the simulated data sets for specific continues variables. The variables must be declared as ROWNAMES in the matrix

`multi_sugg_prop`
A named vector that provides a proportion of value=1 for specific binary variables(=name of the vector) that will be the close to the proportion of this value in the simulated data sets.

`generalized_mode`
A logical value indicating if generalized lambda/poisson distributions or set up thresholds will be used to generate the simulated values

`generalized_mode_model`
A matrix that contains two columns named "Variable" and "Model". This matrix can be used only if a generalized_mode_model argument is provided. It specifies what model should be used for each Variable. Model values should be "rmfmkl", "rprs", "star" or a combination of them, e.g. "rmfmkl-rprs" or "star-star", in case the use wants a bimodal simulation. The user can select Generalised Poisson model for poisson variables, but this model cannot be included in bimodal simulation

`generalized_mode_model_event`
A matrix that contains two columns named "Variable" and "Model" and it is to be used for the event data set(event = 1). This matrix can be used only if a generalized_mode_model argument is provided. It specifies what model should be used for each Variable. Model values should be "rmfmkl", "rprs", "star" or a combination of them, e.g. "rmfmkl-rprs" or "star-star", in case the use wants a bimodal simulation. The user can select Generalised Poisson model for poisson variables, but this model cannot be included in bimodal simulation

`generalized_mode_model_no_event`
A matrix that contains two columns named "Variable" and "Model" and it is to be used for the non-event data set(event = 0). This matrix can be used only if a generalized_mode_model argument is provided. It specifies what model should be used for each Variable. Model values should be "rmfmkl", "rprs", "star" or a combination of them, e.g. "rmfmkl-rprs" or "star-star", in case the use wants a

bimodal simulation. The user can select Generalised Poisson model for poisson variables, but this model cannot be included in bimodal simulation

`generalized_mode_lmbds`
: A matrix that contains lambdas values for each of the variables of the data set to be used for either Generalized Lambda Distribution Generalized Poisson Distribution or setting up thresholds

Details

Simulated data is generated based on available data. The simulated data mimics the characteristics of the original data. The algorithm used is based on the ranked based inverse normal transformation (Koliopanos et al. (2023)).

Value

A list with the following components:

`simulated_data`
: A list of data frames containing the simulated data.

`original_data`
: A data frame with the input data.

`correlations`
: a list of correlation matrices. The ith element is the correlation matrix for the ith simulated dataset. The `(repn + 1)`the (last) element of the list is the average of the correlation matrices.

`bin_variables`
: character vector listing the binary variables

`categ_variables`
: a character vector listing the ordinal categorical variables

`covariance_matrix`
: Covariance matrix used when generating observations from a multivariate normal distribution.

`seed`
: Random seed used.

`samples_produced`
: Number of rows of each simulated dataset.

`sim_dataset_number`
: Number of simulated datasets produced.

Author(s)

Francisco M. Ojeda, George Koliopanos

Examples

```
data("cancer", package = "survival")
cancer_data <- na.omit(cancer)
cancer_data$sex <- cancer_data$sex - 1
cancer_data$status <- cancer_data$status - 1
test_surv <- modgo_survival(data = cancer_data,
                            surv_method = 1,
                            bin_variables = c("status", "sex"),
                            categ_variables = "ph.ecog",
                            event_variable = "status",
```

```
                    time_variable = "time",
                    generalized_mode_model_no_event = "rmfmkl",
                    generalized_mode_model_event = "rprs")
```

`multicenter_comb` *Modgo multi-studies*

Description

Combines modgo objects from a multiple studies to a single one in order to calculate new correlations and visualise the data

Usage

```
multicenter_comb(modgo_1, ...)
```

Arguments

`modgo_1`	a list modgo object.
`...`	multiple modgo object names.

Value

A modgo object/list that consist the merging of multiple modgo objects.

Author(s)

Francisco M. Ojeda, George Koliopanos

`rbi_normal_transform` *Rank based inverse normal transformation*

Description

Applies the rank based inverse normal transformation to numeric vector.

Usage

```
rbi_normal_transform(x, ties_method = c("max", "min", "average"))
```

Arguments

`x`	a numeric vector
`ties_method`	Method on how to deal with equal values during rank transformation.Acceptable input:"max","average","min". This parameter is passed to the parameter `ties.method` of `rank`.

Details

The rank based inverse normal transformation (Beasley et al. (2009)), transforms values of a vector to ranks and then applies the quantile function of the standard normal distribution.

Value

A numeric vector with rank transformed values.

Author(s)

Andreas Ziegler, Francisco M. Ojeda, George Koliopanos

References

Beasley, T.M. and Erickson S. and Allison D.B. (2009), "Rank-based inverse normal transformations are increasingly used, but are they merited?," *Behavior genetics* **39**, 580-595.

Examples

```
data("Cleveland",package="modgo")
test_rank <- rbi_normal_transform(Cleveland[,1])
```

`rbi_normal_transform_inv`

Inverse of rank based inverse normal transformation

Description

Transforms a vector `x` using the inverse of rank based inverse normal transformation associated with a given vector `x_original`. This inverse is defined as $F_n^{-1}\Phi(x)$, where F_n^{-1} is the inverse empirical cumulative distribution function of `x_original` and Φ is the cumulative distribution function of a standard normal random variable.

Usage

```
rbi_normal_transform_inv(x, x_original)
```

Arguments

`x`	a numeric vector.
`x_original`	a numeric vector from the original dataset

Value

A numeric vector with inverse transformed values

Author(s)

Andreas Ziegler, Francisco M. Ojeda, George Koliopanos

Examples

```
data("Cleveland",package="modgo")
test_rank <- rbi_normal_transform(Cleveland[,1])
test_inv_rank <- rbi_normal_transform_inv(x = test_rank,
                                          x_original = Cleveland[,1])
```

Sigma_calculation *Calculate Sigma with the help of polychoric and polyserial functions*

Description

This function is used internally by modgo. It conducts the computation of the correlation matrix of the transformed variables, which are assumed to follow a multivariate normal distribution.

Usage

```
Sigma_calculation(data, variables, bin_variables, categ_variables, ties_method)
```

Arguments

data	a data frame with original variables.
variables	variables a character vector indicating which columns of data should be used.
bin_variables	a character vector listing the binary variables.
categ_variables	a character vector listing the ordinal categorical variables.
ties_method	Method on how to deal with equal values during rank transformation. Acceptable input:"max","average","min". This parameter is passed by rbi_normal_transform to the parameter ties.method of rank.

Value

A numeric matrix with correlation values.

Author(s)

Francisco M. Ojeda, George Koliopanos

`Sigma_transformation` *Correlation of transformed variables*

Description

This function is used internally by `modgo`. It finishes the computation of the correlation matrix of the transformed variables, which are assumed to follow a multivariate normal distribution. It computes the correlations involving at least one categorical variable. For this purpose the biserial, tetrachoric, polyserial and polychoric correlations are used.

Usage

```
Sigma_transformation(
  data,
  data_z,
  Sigma,
  variables,
  bin_variables = c(),
  categ_variables = c()
)
```

Arguments

`data`	a data frame with original variables.
`data_z`	data frame with transformed variables.
`Sigma`	A numeric square matrix.
`variables`	variables a character vector indicating which columns of `data` should be used.
`bin_variables`	a character vector listing the binary variables.
`categ_variables`	a character vector listing the ordinal categorical variables.

Value

A numeric matrix with correlation values.

Author(s)

Francisco M. Ojeda, George Koliopanos

Index