Package ‘heteromixgm’

August 1, 2024

Type Package

Title Copula Graphical Models for Heterogeneous Mixed Data

Imports Matrix, igraph, parallel, tmvtnorm, glasso, BDgraph, methods, stats, utils, MASS

Version 2.0.1

Maintainer Sjoerd Hermes <sjoerd.hermes@wur.nl>

Description A multi-core R package that allows for the statistical modeling of multi-group multivariate mixed data using Gaussian graphical models. Combining the Gaussian copula framework with the fused graphical lasso penalty, the 'heteromixgm' package can handle a wide variety of datasets found in various sciences. The package also includes an option to perform model selection using the AIC, BIC and EBIC information criteria, a function that plots partial correlation graphs based on the selected precision matrices, as well as simulate mixed heterogeneous data for exploratory or simulation purposes and one multi-group multivariate mixed agricultural dataset pertaining to maize yields. The package implements the methodological developments found in Hermes et al. (2024) <doi:10.1080/10618600.2023.2289545>.

License GPL-3

Encoding UTF-8

LazyData true

Depends R (>= 3.10)

NeedsCompilation no

Author Sjoerd Hermes [aut, cre], Joost van Heerwaarden [ctb], Pariya Behrouzi [ctb]

Repository CRAN

Date/Publication 2024-08-01 11:00:02 UTC

Contents

data_sim ....................................................... 2
heteromixgm ................................................... 3
initialize ..................................................... 4
Description

Simulate mixed multi-group data.

Usage

data_sim(network, n, p, K, ncat, rho, gamma_g = NULL, gamma_o, gamma_b = NULL, gamma_p = NULL, prob = NULL, nclass = NULL)

Arguments

- `network`: Type of network, either "circle", "Random", "Cluster", "Scale-free", "AR1" or "AR2".
- `n`: Number of observations.
- `p`: Number of variables.
- `K`: Number of groups.
- `ncat`: Number of categories for ordinal variables.
- `rho`: Dissimilarity parameter inducing dissimilarity between the K datasets.
- `gamma_g`: Proportion of Gaussian variables in the data.
- `gamma_o`: Proportion of ordinal variables in the data.
- `gamma_b`: Proportion of binomial variables in the data.
- `gamma_p`: Proportion of Poisson variables in the data.
- `prob`: Edge occurrence probability in random graph.
- `nclass`: Number of clusters in cluster graph.

Value

- `z`: A list of \( K \) \( n \) by \( p \) matrices representing the latent Gaussian transformed (observed) data.
- `theta`: A list of \( K \) \( n \) by \( p \) matrices representing the precision matrices corresponding to the latent Gaussian (unobserved) data.

Author(s)

Sjoerd Hermes, Joost van Heerwaarden and Pariya Behrouzi
Maintainer: Sjoerd Hermes <sjoerd.hermes@wur.nl>
heteromixgm

References


Examples

data_sim(network = "Random", n = 10, p = 50, K = 3, ncat = 6, rho = 0.25, 
gamma_o = 0.5, gamma_b = 0.1, gamma_p = 0.2, prob = 0.05)

Description

This function implements either the Gibbs or approximation method within the Gaussian copula graphical model to estimate the conditional expectation for the data that not follow Gaussianity assumption (e.g. ordinal, discrete, continuous non-Gaussian, or mixed dataset).

Usage

heteromixgm(X, method, lambda1, lambda2, ncores)

Arguments

- **X**: A list containing $K n_k \times p$ matrices ($K$ is the number of groups, $n_k$ is the sample size for group $k$ and $p$ is the number of variables)
- **method**: Choice between "Gibbs" and "Approximate" indicating which method to use.
- **lambda1**: Vector containing values (in $[0,1]$) for the sparsity penalization of each $\Theta^k$.
- **lambda2**: Vector containing values (in $[0,1]$) for the similarity penalization between the $\Theta^k$.
- **ncores**: Number of cores to be used during parallel computing.

Value

- **Z**: New transformation of the data based on given or default Sigma.
- **ES**: Expectation of covariance matrix (diagonal scaled to 1) of the Gaussian copula graphical model.
- **Sigma**: The covariance matrix of the latent variable given the data.
- **Theta**: The inverse covariance matrix of the latent variable given the data.
- **loglik**: Value of the Log likelihood under the estimated parameters.

Author(s)

Sjoerd Hermes, Joost van Heerwaarden and Pariya Behrouzi
Maintainer: Sjoerd Hermes <sjoerd.hermes@wur.nl>
References


Examples

data(maize)
l1 <- c(0.4)
l2 <- c(0,0.1)
ncores <- 1
est <- heteromixgm(maize, "Approximate", l1, l2, ncores)

Description

Initialize parameters to be used in the approximate method algorithm.

Usage

initialize(y, ncores)

Arguments

y Data.
ncores Number of cores to be used during parallel computing.

Value

ES Expectation of covariance matrices (diagonal scaled to 1) of the Gaussian copula graphical model.
Z New transformation of the data based on given or default Sigma.
lower_upper Lower and upper truncation points for the truncated normal distribution.

Author(s)

Sjoerd Hermes, Joost van Heerwaarden and Pariya Behrouzi
Maintainer: Sjoerd Hermes <sjoerd.hermes@wur.nl>

References

Examples

```r
y <- list(matrix(runif(25), 5, 5), matrix(runif(25), 5, 5), matrix(runif(25), 5, 5))
ncores <- 1
initialize(y, ncores)
```

Description

Calculates lower and upper bands for each data point, using a set of cut-points which is obtained from the Gaussian copula.

Usage

```r
lower.upper(y)
```

Arguments

- **y**: An \((n_k \times p)\) matrix corresponding to the data matrix (\(n_k\) is the sample size for group \(k\) and \(p\) is the number of variables).

Value

- **lower**: A \(n_k\) by \(p\) matrix representing the lower band for each data point.
- **upper**: A \(n_k\) by \(p\) matrix representing the upper band for each data point.

Author(s)

Sjoerd Hermes, Joost van Heerwaarden and Pariya Behrouzi
Maintainer: Sjoerd Hermes <sjoerd.hermes@wur.nl>

References


Examples

```r
y <- list(matrix(runif(25), 5, 5), matrix(runif(25), 5, 5), matrix(runif(25), 5, 5))
lower.upper(y[[1]])
```
Maize data

Description

This is a dataset consisting of maize yields, environmental and management variables measured across 2 groups. The groups pertain to different seasons (2010 and 2013) for farms in Pawe Ethiopia.

Usage

```r
data("maize")
```

Format

The format is: List of 2

Details

Contains a subset of data used in the Hermes et al. (2024) paper, which is a subset of data used in the Vasco Silva et al. (forthcoming) paper.

Source


References


Examples

```r
data(maize)
```
modselect

Description

Model selection using the AIC, BIC and eBIC.

Usage

modselect(est, X, l1, l2, gamma)

Arguments

- `est`: Estimates of model obtained from cgmm() function
- `X`: A list of $K$ $n_k$ by $p$ data matrices.
- `l1`: Vector containing l1 penalty values.
- `l2`: Vector containing l2 penalty values.
- `gamma`: EBIC gamma parameter.

Value

- `selectmat`: Matrix containing the "optimal" l1 and l2 values for each information criterion.
- `theta_aic`: Estimated precision matrices using the AIC for model selection.
- `theta_bic`: Estimated precision matrices using the BIC for model selection.
- `theta_ebic`: Estimated precision matrices using the EBIC for model selection.

Author(s)

Sjoerd Hermes, Joost van Heerwaarden and Pariya Behrouzi
Maintainer: Sjoerd Hermes <sjoerd.hermes@wur.nl>

References


Examples

```r
X <- list(matrix(runif(25), 5, 5), matrix(runif(25), 5, 5), matrix(runif(25), 5, 5))
l1 <- c(0.4)
l2 <- c(0.1)
gamma <- 0.5
ncores <- 1
est <- heteromixgm(X, "Approximate", l1, l2, ncores)
modselect(est, X, l1, l2, gamma)
```
plot_pcorgraph  

Plot partial correlation graphs

Description
Plots all $K$ partial correlation graphs based on the $\Theta$ selected using one of the information criteria.

Usage
plot_pcorgraph(Theta, pos_clr, neg_clr, plot_layout, label_cex)

Arguments
- **Theta**: List of $K$ selected $\Theta$
- **pos_clr**: Color, hexadecimal color allowed, representing the positive partial correlations in the plotted graphs.
- **neg_clr**: Color, hexadecimal color allowed, representing the negative partial correlations in the plotted graphs.
- **plot_layout**: Number of rows and columns for the plot layout.
- **label_cex**: Size of the vertex labels in the plotted graphs.

Value
There is no return value. The function only shows plots in the graphics output device.

Author(s)
Sjoerd Hermes, Joost van Heerwaarden and Pariya Behrouzi
Maintainer: Sjoerd Hermes <sjoerd.hermes@wur.nl>

References

Examples
```r
temp <- data_sim(network = "Random", n = 100, p = 20, K = 4, ncat = 6, rho = 0.25,
                 gamma_o = 0.5, gamma_b = 0.1, gamma_p = 0.2, prob = 0.05)
X <- temp$z
l1 <- c(0.1)
l2 <- c(0,0.1)
gamma <- 0.5
ncores <- 1
est <- heteromixgm(X, "Approximate", l1, l2, ncores)
temp = modselect(est, X, l1, l2, gamma)
plot_pcorgraph(temp$theta_aic, "green", "red", c(2,2), 4.5)
```
Index

* datasets
  maize, 6
data_sim, 2
heteromixgm, 3
initialize, 4
lower.upper, 5
maize, 6
modselect, 7
plot_pcorgraph, 8