Package 'HGMND'

October 12, 2022

Type Package

Title Heterogeneous Graphical Model for Non-Negative Data

Version 0.1.0

Description Graphical model is an informative and powerful tool to explore the conditional dependence relationships among variables. The traditional Gaussian graphical model and its extensions either have a Gaussian assumption on the data distribution or assume the data are homogeneous. However, there are data with complex distributions violating these two assumptions. For example, the air pollutant concentration records are non-negative and, hence, non-Gaussian. Moreover, due to climate changes, distributions of these concentration records in different months of a year can be far different, which means it is uncertain whether datasets from different months are homogeneous. Methods with a Gaussian or homogeneous assumption may incorrectly model the conditional dependence relationships among variables. Therefore, we propose a heterogeneous graphical model for non-negative data (HGMND) to simultaneously cluster multiple datasets and estimate the conditional dependence matrix of variables from a non-Gaussian and non-negative exponential family in each cluster.

License GPL-3

Encoding UTF-8

LazyData true

Imports genscore

Depends R (>= 3.6.0)

NeedsCompilation no

Author Jiaqi Zhang [aut, cre],
Xinyan Fan [aut],
Yang Li [aut]

Maintainer Jiaqi Zhang <boarzhang@gmail.com>

Repository CRAN

Date/Publication 2021-04-19 09:00:02 UTC

R topics documented:

`getCluster` *Get the cluster structure of the HGMND estimate*

Description

After estimating the conditional dependence matrices of the multiple datasets using the HGMND method, the cluster structure can be revealed by comparison of these matrices.

Usage

```
getCluster(est.HGMND, method = "F", tol = 1e-5)
```

Arguments

`est.HGMND`	a list, the result of the function `HGMND` with class "est.HGMND".
`method`	the method of evaluating the difference of two conditional dependence matrices. The function `norm` from the `base` package is used to calculate the matrix norm of the element-wise difference of two matrices. It must be chosen from `"O"`, `"I"`, `"F"`, `"M"`, `"2"`, corresponding to the same settings in the function `norm`. Default to `"F"`, the Frobenius norm.
`tol`	tolerance in evaluating the difference of two conditional dependence matrices. If the calculated difference is no larger than `tol`, they are regarded as in one cluster. Default to `1e-5`.

Value

the function `getCluster` returns the clustering structure of the multiple conditional dependence matrices.

`mat.comapre`	a matrix of 0 or 1. If the element on the ith row and jth column of the matrix is 1, the ith and the jth conditional dependence matrices are in the same cluster, 0 otherwise.
`est.cluster`	a vector with length same as the number of conditional dependence matrices indicating the cluster label of each matrix.

Examples

```
# This is an example of HGMND with simulated data
data(HGMND_SimuData)
h              <- genscore::get_h_hp("mcp", 1, 5)
HGMND_SimuData <- lapply(HGMND_SimuData, function(x) scale(x, center = FALSE))
mat.chain      <- diag(length(HGMND_SimuData))
diag(mat.chain[-nrow(mat.chain), -1]) <- 1

result <- HGMND(x        = HGMND_SimuData,
                setting  = "gaussian",
```

```
                h        = h,
                centered = FALSE,
                mat.adj  = mat.chain,
                lambda1  = 0.086,
                lambda2  = 3.6,
                gamma    = 1,
                tol      = 1e-3,
                silent   = TRUE)
Theta       <- result[["Theta"]]
res.cluster <- getCluster(result)
```

HGMND *Heterogeneous Graphical Model for Non-Negative Data*

Description

The HGMND is the main function to estimate the conditional dependence matrices of variables from different datasets.

Usage

```
HGMND(x,
     setting,
     h,
     centered,
     mat.adj,
     lambda1,
     lambda2,
     gamma   = 1,
     maxit   = 200,
     tol     = 1e-5,
     silent  = TRUE)
```

Arguments

x	a list of data matrices sharing the same variables in their columns.
setting	a string that indicates the data distribution, must be chosen from "gaussian", "gamma", "exp".
h	the function h(x) used in the h-generalized score matching loss, which returns a list containing hx = h(x) and its derivative hpx = hp(x), where x is the data matrix. See details for more information.
centered	logical, if centered = TRUE, the data distribution is assumed centered with $\eta = 0$.
mat.adj	the adjacency matrix of the network among the multiple datasets, containing only 0s and 1s. Only the upper-triangle of mat.adj is used.
lambda1	the non-negative tuning parameter which controls the sparsity level of the estimation.

lambda2	the non-negative tuning parameter which controls the homogeneity level of the estimation.
gamma	the step size parameter in ADMM. Default to 1.
maxit	maximum number of iterations. Default to 200.
tol	tolerance in the convergence criterion. Default to 1e-5.
silent	logical, if silent = FALSE, the prime and dual feasibility and the time used in each ADMM iteration will show on the console.

Details

h can be generated by function get_h_hp in package genscore. See more details in Yu S., Lin, L. & Gilks, W. (2020). genscore: Generalized Score Matching Estimators. R package version 1.0.2. https://CRAN.R-project.org/package=genscore and Yu, S., Drton, M., & Shojaie, A. (2019). Generalized Score Matching for Non-Negative Data. J. Mach. Learn. Res., 20, 76-1.

Suppose we have M datasets, and we demand the network among them to be connected and have $M - 1$ edges, hence acyclic. This is sufficient for computational feasibility, which however does not prevent our method from being applicable to diverse network structures.

Value

The HGMND method returns the estimated conditional dependence matrix of each dataset.

Theta	the 3-dimensional array containing the estimation of the multiple conditional dependence matrices. The 3rd dimension represents different datasets.
M	an integer, the number of datasets.
P	an integer, dimension of the random vector of interest.

References

Yu, S., Drton, M., & Shojaie, A. (2019). Generalized Score Matching for Non-Negative Data. J. Mach. Learn. Res., 20, 76-1.

Yu S., Lin, L. & Gilks, W. (2020). genscore: Generalized Score Matching Estimators. R package version 1.0.2. https://CRAN.R-project.org/package=genscore.

Examples

```
# This is an example of HGMND with simulated data
data(HGMND_SimuData)
h              <- genscore::get_h_hp("mcp", 1, 5)
HGMND_SimuData <- lapply(HGMND_SimuData, function(x) scale(x, center = FALSE))
mat.chain      <- diag(length(HGMND_SimuData))
diag(mat.chain[-nrow(mat.chain), -1]) <- 1

result <- HGMND(x        = HGMND_SimuData,
                setting  = "gaussian",
                h        = h,
                centered = FALSE,
                mat.adj  = mat.chain,
```

```
                lambda1  = 0.086,
                lambda2  = 3.6,
                gamma    = 1,
                tol      = 1e-3,
                silent  = TRUE)
Theta       <- result[["Theta"]]
```

HGMND_SimuData *An example of simulated data for HGMND*

Description

The dataset HGMND_SimuData contains 20 data matrices from two clusters. The first 10 matrices belong to the first cluster and the last 10 ones belong to the other. Data in the same cluster are from the same non-centered truncated Gaussian distribution.

Usage

```
HGMND_SimuData
```

Format

A list of length 20.

Index