Package ‘GGMM’

March 19, 2019

Type Package
Title Mixture Gaussian Graphical Models
Version 1.0.1
Date 2019-03-17
Depends R (>= 3.0.2)
Imports mvtorm, equSA, huge
Description The Gaussian graphical model is a widely used tool for learning gene regulatory networks with high-dimensional gene expression data. For many real problems, the data are heterogeneous, which may contain some subgroups or come from different resources. This package provide a Gaussian Graphical Mixture Model (GGMM) for the heterogeneous data. You can refer to Jia, B. and Liang, F. (2018) at <arXiv:1805.02547> for detail.
License GPL-2
LazyLoad true
NeedsCompilation yes
Repository CRAN
Date/Publication 2019-03-19 14:43:22 UTC
RoxygenNote 6.0.1
Author Bochao Jia [aut, ctb, cre, cph],
Faming Liang [ctb]
Maintainer Bochao Jia <jbc409@gmail.com>

R topics documented:

<table>
<thead>
<tr>
<th>Package</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>GGMM-package</td>
<td>2</td>
</tr>
<tr>
<td>breast</td>
<td>3</td>
</tr>
<tr>
<td>BRGM</td>
<td>3</td>
</tr>
<tr>
<td>GGMM</td>
<td>5</td>
</tr>
<tr>
<td>SimHetDat</td>
<td>6</td>
</tr>
</tbody>
</table>

Index 8
Description

The Gaussian graphical model is a widely used tool for learning gene regulatory networks with high-dimensional gene expression data. For many real problems, the data are heterogeneous, which may contain some subgroups or come from different resources. This package provides a Gaussian Graphical Mixture Model (GGMM) for the heterogeneous data.

Details

Package: GGMM
Type: Package
Version: 1.0.1
Date: 2019-03-17
License: GPL-2

This package illustrates the use of the Gaussian Graphical Mixture Model in two parts:
The first part is to apply the GGMM to estimate network structures using high-dimensional heterogeneous data with a simulated dataset SimHetDat(n, p, . . .) and our proposed method GGMMDat(data, . . .).

The second part is to apply the GGMM to learn a real data example BRGM(breast, . . .), i.e. to learn a common gene regulatory network with heterogeneous gene expression data of breast cancer. The real data example are from The Cancer Genome Atlas (TCGA) with code data(breast).

Author(s)

Bochao Jia, Faming Liang Maintainer: Bochao Jia <jbc409@ufl.edu>

References


Examples

```r
library(GGMM)
library(huge)
result <- SimHetDat(n = 100, p = 200, M = 2, mu = 0.5, type = "band")
Graph <- GGMM(result$data, result$A, M = 2, iteration = 30, warm = 20)
```

breast

*Example dataset for learning gene regulatory network.*

Description

Breast cancer is one of the most prevalent types of cancer which can be classified into four molecular subtypes, namely, luminal A, basal-like, HER2-enriched, and luminal B, based on their tumor expression profiles (Haque et al, 2012). In this study, we aim to construct a single gene regulatory network across the four subtypes to discover the overall gene regulation mechanism in breast cancer. It should be used in `BRGM(data, ...)`.

Usage

```r
data(breast)
```

Format

- **breast**
  - a `n x p` breast cancer gene expression data.

References


BRGM

*Learning gene regulatory networks for breast cancer.*

Description

Gaussian Graphical Mixture Models for learning gene regulatory network with multiple subtypes of breast cancer dataset.

Usage

```r
BRGM(data, M=3, alpha1 = 0.05, alpha2 = 0.02, alpha3 = 0.2, iteration = 30, warm = 20)
```
Arguments

- **data**: A nxp matrix of breast cancer expression data.
- **M**: The number of heterogeneous groups, default of 3 based on the BIC scores.
- **alpha1**: The significance level of correlation screening in the \( \psi \)-learning algorithm, see R package `equSA` for detail. In general, a high significance level of correlation screening will lead to a slightly large separator set, which reduces the risk of missing important variables in the conditioning set. In general, including a few false variables in the conditioning set will not hurt much the accuracy of the \( \psi \)-partial correlation coefficient, the default value is 0.05.
- **alpha2**: The significance level of \( \psi \)-partial correlation coefficient screening for estimating the adjacency matrix, see `equSA`, the default value is 0.02.
- **alpha3**: The significance level of integrative \( \psi \)-partial correlation coefficient screening for estimating the adjacency matrix of GGMM method, the default value is 0.2.
- **iteration**: The number of total iterations, the default value is 30.
- **warm**: The number of burn-in iterations, the default value is 20.

Value

- **adj**: A pxp Estimated adjacency matrix for network construction.
- **label**: The estimated group indices for each observation.
- **bic**: The BIC scores for determining the number of groups \( M \).

Author(s)

Bochao Jia<jbc409@ufl.edu> and Faming Liang

References


Examples

```r
library(GGMM)
library(huge)
data(breast)
## long time ##
M <- 3
```
GGMM

Learning high-dimensional Gaussian Graphical Models with Heterogeneous Data.

Description

Gaussian Graphical Mixture Models for learning high-dimensional network with simulated heterogeneous data.

Usage

GGMM(data, A, M, alpha1 = 0.1, alpha2 = 0.05, alpha3 = 0.05, iteration = 30, warm = 20)

Arguments

data          n x p mixture Gaussian distributed dataset.
A             p x p true adjacency matrix for evaluating the performance.
M             The number of heterogeneous groups.
alpha1        The significance level of correlation screening in the ψ-learning algorithm, see R package equSA for detail. In general, a high significance level of correlation screening will lead to a slightly large separator set, which reduces the risk of missing important variables in the conditioning set. In general, including a few false variables in the conditioning set will not hurt much the accuracy of the ψ-partial correlation coefficient, the default value is 0.1.
alpha2        The significance level of ψ-partial correlation coefficient screening for estimating the adjacency matrix. The default value is 0.05.
alpha3        The significance level of integrative ψ-partial correlation coefficient screening for estimating the adjacency matrix of GGMM method, the default value is 0.05.
iteration     The number of total iterations, the default value is 30.
warm          The number of burn-in iterations, the default value is 20.

Value

RecPre        The output of Recall and Precision values of our proposed method.
Adj           p x p Estimated adjacency matrix.
label         The estimated group indices for each observation.
BIC           The BIC scores for determining the number of groups M.
**Author(s)**

Bochao Jia<jbc409@ufl.edu> and Faming Liang

**References**


**Examples**

```r
library(GGMM)
library(huge)
result <- SimHetDat(n = 100, p = 200, M = 3, mu = 0.5, type = "band")
Est <- GGM(result$data, result$A, M = 3, iteration = 30, warm = 20)
## plot network by our estimated adjacency matrix.
huge.plot(Est$Adj)
## plot the Recall-Precision curve
plot(Est$RecPre[,1], Est$RecPre[,2], type="l", xlab="Recall", ylab="Precision")
```

**SimHetDat**  
*Simulate Heterogeneous Data for Gaussian Graphical Models*

**Description**

Simulate Heterogeneous data with a band structure, which can be used in GGMM(data,...) for estimating the structure of the Gaussian graphical network.

**Usage**

SimHetDat(n = 100, p = 200, M = 3, mu = 0.3, type = "band")
SimHetDat

Arguments

n
Number of observations for each group, default of 100.

p
Number of covariates for each observation, default of 200.

M
Number of latent groups for the simulated dataset choose 2 or 3, default of 3.

mu
The mean difference among groups. If \( M = 3 \), the mean of three groups are \(-mu, 0, mu\), respectively. If \( M = 2 \), the mean of two groups are \(0, mu\), respectively.

type
type="band" which denotes the band structure, with precision matrix

\[
C_{i,j} = \begin{cases} 
0.5, & \text{if } |j-i| = 1, i = 2, \ldots, (p-1), \\
0.25, & \text{if } |j-i| = 2, i = 3, \ldots, (p-2), \\
1, & \text{if } i = j, i = 1, \ldots, p, \\
0, & \text{otherwise}.
\end{cases}
\]

Value

data
\( n \times p \) Heterogeneous Gaussian distributed data.

A
\( p \times p \) adjacency matrix used for generating data.

label
The group indices for each observation.

Author(s)

Bochao Jia<jbc409@ufl.edu> and Faming Liang

References


Examples

library(GGMM)
SimHetDat(n = 100, p = 200, M = 3, mu = 0.5, type = "band")
Index

*Topic **BRGM**
   BRGM, 3
*Topic **GGMM**
   GGMM, 5
*Topic **SimHetDat**
   SimHetDat, 6
*Topic **datasets**
   breast, 3
*Topic **package**
   GGMM-package, 2

breast, 3
BRGM, 3

GGMM, 5
GGMM-package, 2

SimHetDat, 6