Package 'GGMM'

May 9, 2018

Type Package

Title Mixture Gaussian Graphical Models

Version 1.0.0

Date 2018-05-07

Depends R (>= 3.0.2)

Imports mvtnorm, 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 2018-05-09 08:36:18 UTC

RoxygenNote 6.0.1

Author Bochao Jia [aut, ctb, cre, cph],
Faming Liang [ctb]

Maintainer Bochao Jia <jbc409@ufl.edu>

R topics documented:

GGMM-package *Gaussian Graphical Mixture Models*

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.

Details

Package:	GGMM
Type:	Package
Version:	1.0.0
Date:	2018-05-07
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 `GGMM(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

Liang, F., Song, Q. and Qiu, P. (2015). An Equivalent Measure of Partial Correlation Coefficients for High Dimensional Gaussian Graphical Models. J. Amer. Statist. Assoc., 110, 1248-1265.<doi:10.1080/01621459.2015.1012391>

Liang, F. and Zhang, J. (2008) Estimating FDR under general dependence using stochastic approximation. Biometrika, 95(4), 961-977.<doi:10.1093/biomet/asn036>

Liang, F., Jia, B., Xue, J., Li, Q., and Luo, Y. (2018). An Imputation Regularized Optimization Algorithm for High-Dimensional Missing Data Problems and Beyond. Submitted to Journal of the Royal Statistical Society Series B. <arXiv:1802.02251>

Jia, B., Xu, S., Xiao, G., Lamba, V., Liang, F. (2017) Inference of Genetic Networks from Next Generation Sequencing Data. Biometrics.

Jia, B. and Liang, F. (2018). Learning Gene Regulatory Networks with High-Dimensional Heterogeneous Data. Accept by ICSA Springer Book. <arXiv:1805.02547>

Examples

```
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 lu- minal 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

```
data(breast)
```

Format

breast a *nxp* breast cancer gene expression data.

References

Haque, R., Ahmed, S. A., Inzhakova, G., Shi, J., Avila, C., Polikoff, J., ... and Press, M. F. (2012). Impact of breast cancer subtypes and treatment on survival: an analysis spanning two decades. Cancer Epidemiology and Prevention Biomarkers, 21(10), 1848-1855.

BRGM *Learning gene regulatory networks for breast cancer.*

Description

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

Usage

```
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 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 0.05.
alpha2	The significance level of ψ-partial correlation coefficient screening for estimating the adjacency matrix, see **equSA**, the default value is 0.02.
alpha3	The significance level of integrative ψ-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	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

Liang, F., Song, Q. and Qiu, P. (2015). An Equivalent Measure of Partial Correlation Coefficients for High Dimensional Gaussian Graphical Models. J. Amer. Statist. Assoc., 110, 1248-1265.

Liang, F. and Zhang, J. (2008) Estimating FDR under general dependence using stochastic approximation. Biometrika, 95(4), 961-977.

Liang, F., Jia, B., Xue, J., Li, Q., and Luo, Y. (2018). An Imputation Regularized Optimization Algorithm for High-Dimensional Missing Data Problems and Beyond. Submitted to Journal of the Royal Statistical Society Series B.

Examples

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

```
Gra <- BRGM(breast, M, iteration = 30, warm = 20)
## plot gene regulatory network by our estimated adjacency matrix.
huge.plot(Gra$Adj)
```

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	nxp mixture Gaussian distributed dataset.
A	pxp 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, see **equSA**, 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	pxp 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

Liang, F., Song, Q. and Qiu, P. (2015). An Equivalent Measure of Partial Correlation Coefficients for High Dimensional Gaussian Graphical Models. J. Amer. Statist. Assoc., 110, 1248-1265.

Liang, F. and Zhang, J. (2008) Estimating FDR under general dependence using stochastic approximation. Biometrika, 95(4), 961-977.

Liang, F., Jia, B., Xue, J., Li, Q., and Luo, Y. (2018). An Imputation Regularized Optimization Algorithm for High-Dimensional Missing Data Problems and Beyond. Submitted to Journal of the Royal Statistical Society Series B.

Jia, B. and Liang, F. (2018). Learning Gene Regulatory Networks with High-Dimensional Heterogeneous Data. Accept by ICSA Springer Book.

Examples

```
library(GGMM)
library(huge)
result <- SimHetDat(n = 100, p = 200, M = 3, mu = 0.5, type = "band")
Est <- GGMM(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")
```

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

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

Value

data	*nxp* Heterogeneous Gaussian distributed data.
A	*pxp* adjacency matrix used for generating data.
label	The group indices for each observation.

Author(s)

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

References

Liang, F., Song, Q. and Qiu, P. (2015). An Equivalent Measure of Partial Correlation Coefficients for High Dimensional Gaussian Graphical Models. J. Amer. Statist. Assoc., 110, 1248-1265.

Examples

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

Index