Package 'diffee'

October 13, 2022

Version 1.1.0

Date 2018-07-03

Title Fast and Scalable Learning of Sparse Changes in High-Dimensional
Gaussian Graphical Model Structure

Author Beilun Wang [aut, cre],
Yanjun Qi [aut],
Zhaoyang Wang [aut]

Maintainer Beilun Wang <bw4mw@virginia.edu>

Depends R (>= 3.0.0), igraph, pcaPP

Description This is an R implementation of Fast and Scalable Learning of Sparse Changes in High-Dimensional Gaussian Graphical Model Structure (DIFFEE). The DIFFEE algorithm can be used to fast estimate the differential network between two related datasets. For instance, it can identify differential gene network from datasets of case and control. By performing data-driven network inference from two high-dimensional data sets, this tool can help users effectively translate two aggregated data blocks into knowledge of the changes among entities between two Gaussian Graphical Model. Please run demo(diffeeDemo) to learn the basic functions provided by this package. For further details, please read the original paper: Beilun Wang, Arshdeep Sekhon, Yanjun Qi (2018) <arXiv:1710.11223>.

License GPL-2

URL https://github.com/QData/DIFFEE

BugReports https://github.com/QData/DIFFEE

RoxygenNote 6.0.1

NeedsCompilation no

Repository CRAN

Date/Publication 2018-07-03 19:50:03 UTC

R topics documented:

diffee-package	*estimating DIFFerential networks via an Elementary Estimator under a high-dimensional situation*

Description

This is an R implementation of Fast and Scalable Learning of Sparse Changes in High-Dimensional Gaussian Graphical Model Structure (DIFFEE). The DIFFEE algorithm can be used to fast estimate the differetial network between two related datasets. For instance, it can identify differential gene network from datasets of case and control. By performing data-driven network inference from two high-dimensional data sets, this tool can help users effectively translate two aggregated data blocks into knowledge of the changes among entities between two Gaussian Graphical Model. Please run demo(diffeeDemo) to learn the basic functions provided by this package. For further details, please read the original paper: Beilun Wang, Arshdeep Sekhon, Yanjun Qi (2018) <arXiv:1710.11223>.

Details

Package:	diffee
Type:	Package
Version:	1.0.0
Date:	2018-03-05
License:	GPL (>= 2)

We focus on the problem of estimating the change in the dependency structures of two p-dimensional Gaussian Graphical models (GGMs). Previous studies for sparse change estimation in GGMs involve expensive and difficult non-smooth optimization. We propose a novel method, DIFFEE for estimating DIFFerential networks via an Elementary Estimator under a high-dimensional situation. DIFFEE is solved through a faster and closed form solution that enables it to work in large-scale settings. We conduct a rigorous statistical analysis showing that surprisingly DIFFEE achieves the same asymptotic convergence rates as the state-of-the-art estimators that are much more difficult to compute. Our experimental results on multiple synthetic datasets and one real-world data about brain connectivity show strong performance improvements over baselines, as well as significant computational benefits.

Author(s)

Beilun Wang, Zhaoyang Wang

Maintainer: Beilun Wang - bw4mw at virginia dot edu

References

Beilun Wang, Arshdeep Sekhon, Yanjun Qi (2018). Fast and Scalable Learning of Sparse Changes in High-Dimensional Gaussian Graphical Model Structure. <arXiv:1710.11223>

Examples

```
## Not run:
data(exampleData)
result = diffee(exampleData[[1]], exampleData[[2]], 0.45)
plot.diffee(result)

## End(Not run)
```

---

cancer   *Microarray data set for breast cancer*

Description

et al's paper. It concerns one hundred thirty-three patients with stage I–III breast cancer. Patients were treated with chemotherapy prior to surgery. Patient response to the treatment can be classified as either a pathologic complete response (pCR) or residual disease (not-pCR). Hess *et al* developed and tested a reliable multigene predictor for treatment response on this data set, composed by a set of 26 genes having a high predictive value.

Usage

```
data(cancer)
```

Format

a list of two objects: dataframe with 133 observations of 26 features and factors indicating whether each sample (out of 133) is of type "not" or type "pcr"

Details

The dataset splits into 2 parts (pCR and not pCR), on which network inference algorithms should be applied independently or in the multitask framework: only individuals from the same classes should be consider as independent and identically distributed.

References

J.A. Mejia, D. Booser, R.L. Theriault, U. Buzdar, P.J. Dempsey, R. Rouzier, N. Sneige, J.S. Ross, T. Vidaurre, H.L. Gomez, G.N. Hortobagyi, and L. Pustzai (2006). Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with Paclitaxel and Fluorouracil, Doxorubicin, and Cyclophosphamide in breast cancer, *Journal of Clinical Oncology*, vol. 24(26), pp. 4236–4244.

---

diffee   *Fast and Scalable Learning of Sparse Changes in High-Dimensional*
                                    Gaussian Graphical Model Structure

---

Description

Estimate DIFFerential networks via an Elementary Estimator under a high-dimensional situation. Please run demo(diffee) to learn the basic functions provided by this package. For further details, please read the original paper: Beilun Wang, Arshdeep Sekhon, Yanjun Qi (2018) <arXiv:1710.11223>.

Usage

```
diffee(C, D, lambda = 0.05, covType = "cov", thre = "soft")
```

Arguments

C	A input matrix for the 'control' group. It can be data matrix or covariance matrix. If C is a symmetric matrix, the matrices are assumed to be covariance matrix. More details at <https://github.com/QData/DIFFEE>
D	A input matrix for the 'disease' group. It can be data matrix or covariance matrix. If D is a symmetric matrix, the matrices are assumed to be covariance matrix. More details at <https://github.com/QData/DIFFEE>
lambda	A positive number. The hyperparameter controls the sparsity level of the matrices. The λ_n in the following section: Details.
covType	A parameter to decide which Graphical model we choose to estimate from the input data.
	If covType = "cov", it means that we estimate multiple sparse Gaussian Graphical models. This option assumes that we calculate (when input X represents data directly) or use (when X elements are symmetric representing covariance matrices) the sample covariance matrices as input to the simule algorithm.
	If covType = "kendall", it means that we estimate multiple nonparanormal Graphical models. This option assumes that we calculate (when input X represents data directly) or use (when X elements are symmetric representing correlation matrices) the kendall's tau correlation matrices as input to the simule algorithm.
thre	A parameter to decide which threshold function to use for T_v. If thre = "soft", it means that we choose soft-threshold function as T_v. If thre = "hard", it means that we choose hard-threshold function as T_v.

Details

The DIFFEE algorithm is a fast and scalable Learning algorithm of Sparse Changes in High-Dimensional Gaussian Graphical Model Structure. It solves the following equation:

$$\min_{\Delta} ||\Delta||_1$$

Subject to :

$$([T_v(\hat{\Sigma}_d)]^{-1} - [T_v(\hat{\Sigma}_c)]^{-1})||_\infty \leq \lambda_n$$

Please also see the equation (2.11) in our paper. The λ_n is the hyperparameter controlling the sparsity level of the matrix and it is the lambda in our function. For further details, please see our paper: Beilun Wang, Arshdeep Sekhon, Yanjun Qi (2018) <arXiv:1710.11223>.

Value

diffNet   A matrix of the estimated sparse changes between two Gaussian Graphical Models

Author(s)

Beilun Wang

References

Beilun Wang, Arshdeep Sekhon, Yanjun Qi (2018). Fast and Scalable Learning of Sparse Changes in High-Dimensional Gaussian Graphical Model Structure. <arXiv:1710.11223>

Examples

```
## Not run:
data(exampleData)
result = diffee(exampleData[[1]], exampleData[[2]], 0.45)
plot.diffee(result)

## End(Not run)
```

---

exampleData   *A simulated toy dataset that includes 2 data matrices (from 2 related tasks).*

Description

A simulated toy dataset that includes 2 data matrices (from 2 related tasks). Each data matrix is about 100 features observed in 200 samples. The two data matrices are about exactly the same set of 100 features. This multi-task dataset is generated from two related random graphs. Please run demo(diffee) to learn the basic functions provided by this package. For further details, please read the original paper: <http://link.springer.com/article/10.1007/s10994-017-5635-7>.

Usage

```
data(exampleData)
```

Format

The format is: List of 2 matrices $: num [1:200, 1:100] -0.0982 -0.2417 -1.704 0.4 ... ..- attr(*, "dimnames")=List of 2 .. ..$: NULL .. ..$: NULL $: num [1:200, 1:100] -0.161 0.41 0.17 0. ... ..- attr(*, "dimnames")=List of 2 .. ..$: NULL .. ..$: NULL

| exampleData500 | *A simulated toy dataset that includes 2 data matrices (from 2 related tasks).* |

Description

A simulated toy dataset that includes 2 data matrices (from 2 related tasks). Each data matrix is about 500 features observed in 1000 samples. The two data matrices are about exactly the same set of 500 features. This multi-task dataset is generated from two related random graphs. Please run demo(diffee) to learn the basic functions provided by this package. For further details, please read the original paper: <http://link.springer.com/article/10.1007/s10994-017-5635-7>.

Usage

```
data(exampleData500)
```

Format

A list of 2 matrices of size [1:500, 1:500] and [1:500, 1:500]

| exampleDataGraph | *A simulated toy dataset that includes 3 igraph objects* |

Description

(first one being the shared graph and second and third being task specific 1 and 2 graphs) The graphs are generated from two related random graphs and the underlaying high dimensional gaussian distribution generates the exampleData dataset. exampleDataGraph serves as a groundtruth to compare in demo(synthetic).

Usage

```
data(exampleDataGraph)
```

Format

A list of 3 igraph objects

---

nip_37_data   *NIPS word count dataset*

Description

This NIPS Conference Papers 1987-2015 Data set is avaiable at UCI Machine Learning Repository. The original dataset is in the form of a 11463 x 5812 matrix of word counts (11463 words and 5812 conference papers) Due to the size of the original dataset, it is preprocessed and reduced to a list of two matrices (2900 x 37 and 2911 x 37) The dataset consists of two tasks (early (up to 2006) and recent (after 2006) NIPS conference papers) with 37 words

Usage

```
data(nip_37_data)
```

Format

a list of two nonnegative integer matrices (1:2900, 1:37) and (1:2911,1:37) Columns are named with year_paperid and rows are names with word name

References

'Poisson Random Fields for Dynamic Feature Models'. Perrone V., Jenkins P. A., Spano D., Teh Y. W. (2016)

---

 plot.diffee   *Plot diffee result specified by user input*

Description

This function can plot and return multiple sparse graphs distinguished by edge colors from the result generated by diffee

Usage

```
## S3 method for class 'diffee'
plot(x, graphlabel = NULL, type = "task", index = NULL,
  graphlayout = NULL, ...)
```

Arguments

x	output generated from diffee function (diffee class)
graphlabel	vertex names for the graph, there are three options: (1) NA (no label) (2) NULL (default numeric label according to the feature order) (3) a vector of labels (a vector of labels cooresponding to x) deault value is NULL
type	type of graph, there are four options: (1) "task" (graph for each task (including shared part) specified further by subID (task number)) (2) "neighbour" (zoom into nodes in the graph specified further by index (node id))
index	determines which node(s) to zoom into when parameter type is "neighbour" could either be an integer or vector of integers representing node ids (zoom into one node or multiple nodes)
graphlayout	layout for the graph (two column matrix specifying x,y coordinates of each node in graph) if not provided, igraph will use the default layout_nicely() function to present the graph
...	extra parameters passed to plot.igraph

Details

when only the diffee is provided, the function will plot all graphs with default numeric labels User can specify multiple subID and multiple index to zoom in multiple nodes on multiple graphs Each graph will include a descriptive title.

Value

a plot of graph / subgraph from diffee result specified by user input

Author(s)

Beilun Wang, Zhaoyang Wang (Author), Beilun Wang (maintainer)

Examples

```
## Not run:
data(exampleData)
result = diffee(exampleData[[1]], exampleData[[2]], 0.45)
plot.diffee(result)

## End(Not run)
```

| returngraph | *return igraph object from diffee result specified by user input* |

Description

This function can return an igraph object from diffee result for user to work with directly

Usage

```
returngraph(x, type = "task", index = NULL)
```

Arguments

x	output generated from diffee function (diffee class)
type	type of graph, there are four options: (1) "task" (graph for each task (including shared part) specified further by subID (task number)) (2) "neighbour" (zoom into nodes in the graph specified further by parameter "index" (node id)
index	determines which node(s) to zoom into when parameter type is "neighbour" could either be an integer or vector of integers representing node ids (zoom into one node or multiple nodes)

Details

the function aims to provide users the flexibility to explore and visualize the graph own their own generated from diffee

Value

an igraph object of graph / subgraph from diffee result specified by user input

Author(s)

Beilun Wang, Zhaoyang Wang (Author), Beilun Wang (maintainer)

Examples

```
## Not run:
data(exampleData)
result = diffee(exampleData[[1]], exampleData[[2]], 0.45)
graph = returngraph(result)

## End(Not run)
```

Index