Package 'SHIP'

February 19, 2015

Type Package

Title SHrinkage covariance Incorporating Prior knowledge

Version 1.0.2

Author Monika Jelizarow and Vincent Guillemot

Maintainer Vincent Guillemot <vincent.j.guillemot@gmail.com>

Description The SHIP-package allows the estimation of various types of shrinkage covariance matrices. These types differ in terms of the so-called covariance target (to be chosen by the user), the highly structured matrix which the standard unbiased sample covariance matrix is shrunken towards and which optionally incorporates prior biological knowledge extracted from the database KEGG. The shrinkage intensity is obtained via an analytical procedure.

License GPL (>= 2)

NeedsCompilation no

Repository CRAN

Date/Publication 2013-12-23 15:17:34

R topics documented:

SHIP-package *SHrinkage covariance Incorporating Prior knowledge*

Description

The SHIP-package implements the shrinkage estimator of a covariance matrix given any covariance target, such as described by Schaefer and Strimmer in 2005. In addition, it proposes several targets based on biological knowledge extracted from the public database KEGG.

Details

To use the shrinkage estimator, one should just have at hand a data set in the form of a $n \times p$ matrix, and a covariance target.

If one wishes to use the proposed targets, the data set should be compatible with KEGG, i.e. it should be possible to extract for each gene the pathways it belongs to. This information, for example, can be found in libraries such as hgu133plus2.db.

Author(s)

Monika Jelizarow and Vincent Guillemot

References

- J. Schaefer and K. Strimmer, 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statist. Appl. Genet. Mol. Biol. 4:32.
- M. Jelizarow, V. Guillemot, A. Tenenhaus, K. Strimmer, A.-L. Boulesteix, 2010. Over-optimism in bioinformatics: an illustration. Bioinformatics. Accepted.

Examples

```
# A short example on a toy dataset
# require(SHIP)

data(expl)
attach(expl)

sig1 <- shrink.estim(x,targetD(x))
sig2 <- shrink.estim(x,targetF(x))
sig3 <- shrink.estim(x,targetCor(x,genegroups))
sig4 <- shrink.estim(x,targetG(x,genegroups))

paste(sig1[[2]],collapse=" ")
paste(sig2[[2]],collapse=" ")
paste(sig3[[2]],collapse=" ")
paste(sig4[[2]],collapse=" ")

## Not run:
# Example on how to get the gene groups lists
```

```
require(hgu95av2.db)
# e.g. we have some interesting gene names :
vec <- c("MYC","ID2","PTGER4","ATF4","FGFR1","MET","HLA-DRB6")
# we then want to convert them into Probe Sets
symb <- as.list(hgu95av2SYMBOL)
pbsets <- names(symb[unlist(sapply(vec,function(x,l) which(l==x)[1],symb))])
# Probe Sets which are themselves converted into a gene groups list
genegroups <- as.list(hgu95av2PATH)[pbsets]

## End(Not run)
```

`build.target` *Creating a covariance target, optionally by using information from KEGG pathways.*

Description

The function build.target() is a wrapper function to build the various types of covariance targets (D,F,G,Gpos,Gstar,cor).

Usage

```
build.target(x, genegroups = NULL, type)
```

Arguments

`x`	An $n \times p$ matrix.
`genegroups`	List of the groups each gene belongs to: each entry of the list is dedicated to a gene (identified the same way as in `x`). Each item of the list is thus a vector of pathway IDs.
`type`	Character string specifying the wished target: "D" for a diagonal target, "cor" for a correlation target, "G", "Gpos" and "Gstar" for a G-type target (see Jelizarow et al, 2010) and "F" for a F-target.

Value

A $p \times p$ target covariance matrix of a certain type.

Author(s)

Vincent Guillemot

References

M. Jelizarow, V. Guillemot, A. Tenenhaus, K. Strimmer, A.-L. Boulesteix, 2010. Over-optimism in bioinformatics: an illustration. Bioinformatics. Accepted.

See Also

targetCor, targetD, targetF, targetG, targetGpos, targetGstar,.

Examples

```
# Simulate dataset
x <- matrix(rnorm(20*30),20,30)
# Try different targets
build.target(x,type="D")
```

---

expl *Small example extracted from a microarray data set.*

Description

The microarray data set is the study on the prostate cancer by Singh et al. The collection of the microarray is hgu95av2, and the gene groups are thus given by the information in the hgu95av2.db Bioconductor library (see Carslon et al.).

Usage

```
data(expl)
```

Details

The dataset is a list containing:

- a 102×100 matrix x of 100 genes randomly chosen from the data set of Singh et al.,
- a list 'genegroups' containing 100 vectors of KEGG pathway IDs (which each gene belongs to).

Source

- M. Carlson, S. Falcon, H. Pages, N. Li. hgu95av2.db: Affymetrix Human Genome U95 Set annotation data (chip hgu95av2). R package version 2.2.12.
- D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub, W. R. Sellers, 2002. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell, Department of Adult Oncology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA., 1, 203-209.

Examples

```
data(expl)
dim(expl$x)
length(expl$genegroups)
```

`shrink.estim` *Shrinkage estimator of the covariance matrix, given a data set and a covariance target.*

Description

The shrinkage estimator is computed independently of the target's nature.

Usage

```
shrink.estim(x, tar)
```

Arguments

`x`	A $n \times p$ matrix (the data set) .
`tar`	A $p \times p$ matrix (the covariance target).

Value

A $p \times p$ shrinkage covariance matrix and the estimated λ.

Author(s)

Monika Jelizarow and Vincent Guillemot

References

J. Schaefer and K. Strimmer, 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statist. Appl. Genet. Mol. Biol. 4:32.

Examples

```
# Simulate dataset
x <- matrix(rnorm(20*30),20,30)
# Try different targets
shrink.estim(x,tar=build.target(x,type="D"))
```

`targetCor` *Computation of the target Cor.*

Description

The $p \times p$ target Cor is computed from the $n \times p$ data matrix. It it a modified version of target G. In particular, it tests the correlations (with a significance level of 0.05) and sets the non-significant correlations to zero before the mean correlation $\bar{r}$ is computed.

Usage

```
targetCor(x, genegroups)
```

Arguments

`x`	A $n \times p$ data matrix.
`genegroups`	A list of genes obtained using the database KEGG, where each entry itself is a list of pathway names this genes belongs to. If a gene does not belong to any gene functional group, the entry is NA.

Value

A $p \times p$ matrix.

Author(s)

Monika Jelizarow and Vincent Guillemot

References

J. Schaefer and K. Strimmer, 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statist. Appl. Genet. Mol. Biol. 4:32.

See Also

`targetCor`, `targetF`, `targetG`, `targetGstar`, `targetGpos`.

Examples

```
# A short example on a toy dataset
# require(SHIP)
data(expl)
attach(expl)
tar <- targetCor(x,genegroups)
which(tar[upper.tri(tar)]!=0) # not many non zero coefficients !
```

`targetD` *Computation of the diagonal target D ('diagonal, unequal variances').*

Description

The $p \times p$ diagonal target D is computed from the $n \times p$ data matrix. It is defined as follows $(i, j = 1, ..., p)$:

$$t_{ij} = \begin{cases} s_{ii} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

where s_{ij} denotes the entry of the unbiased covariance matrix in row i, column j.

Usage

```
targetD(x, genegroups)
```

Arguments

`x`	A $n \times p$ data matrix.
`genegroups`	The genegroups are not used for this target.

Value

A $p \times p$ diagonal matrix.

Author(s)

Monika Jelizarow and Vincent Guillemot

References

J. Schaefer and K. Strimmer, 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statist. Appl. Genet. Mol. Biol. 4:32.

See Also

`targetCor`, `targetF`, `targetG`, `targetGstar`, `targetGpos`.

Examples

```
x <- matrix(rnorm(10*30),10,30)
tar <- targetD(x,NULL)
```

`targetF` *Computation of target F ('constant correlation model').*

Description

The $p \times p$ target F is computed from the $n \times p$ data matrix. It is defined as follows ($i, j = 1, ..., p$):

$$t_{ij} = \begin{cases} s_{ii} & \text{if } i = j \\ \bar{r}\sqrt{s_{ii}s_{jj}} & \text{if } i \neq j \end{cases}$$

where $\bar{r}$ is the average of sample correlations and s_{ij} denotes the entry of the unbiased covariance matrix in row i, column j.

Usage

```
targetF(x, genegroups)
```

Arguments

`x`	A $n \times p$ data matrix.
`genegroups`	The genegroups are not used for this target.

Value

A $p \times p$ matrix.

Author(s)

Monika Jelizarow and Vincent Guillemot

References

J. Schaefer and K. Strimmer, 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statist. Appl. Genet. Mol. Biol. 4:32.

See Also

`targetCor`, `targetF`, `targetG`, `targetGstar`, `targetGpos`.

Examples

```
# A short example on a toy dataset
# require(SHIP)
data(expl)
attach(expl)
tar <- targetF(x,NULL)
which(tar[upper.tri(tar)]!=0) # many non zero coefficients !
```

`targetG` *Computation of target G ('knowledge-based constant correlation model').*

Description

The $p \times p$ target G is computed from the $n \times p$ data matrix. It is defined as follows ($i, j = 1, ..., p$):

$$t_{ij} = \begin{cases} s_{ii} & \text{if } i = j \\ \bar{r}\sqrt{s_{ii}s_{jj}} & \text{if } i \neq j, i \sim j \\ 0 & \text{otherwise} \end{cases}$$

where $\bar{r}$ is the average of sample correlations and s_{ij} denotes the entry of the unbiased covariance matrix in row i, column j. The notation $i \sim j$ means that genes i and j are connected, i.e. genes i and j are in the same gene functional group.

Usage

```
targetG(x, genegroups)
```

Arguments

`x`	A $n \times p$ data matrix.
`genegroups`	A list of genes obtained using the database KEGG, where each entry itself is a list of pathway names this genes belongs to. If a gene does not belong to any gene functional group, the entry is NA.

Value

A $p \times p$ matrix.

Author(s)

Monika Jelizarow and Vincent Guillemot

References

- J. Schaefer and K. Strimmer, 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statist. Appl. Genet. Mol. Biol. 4:32.
- M. Jelizarow, V. Guillemot, A. Tenenhaus, K. Strimmer, A.-L. Boulesteix, 2010. Over-optimism in bioinformatics: an illustration. Bioinformatics. Accepted.

See Also

`targetCor`, `targetF`, `targetG`, `targetGstar`, `targetGpos`.

Examples

```
# A short example on a toy dataset
# require(SHIP)
data(expl)
attach(expl)
tar <- targetG(x,genegroups)
which(tar[upper.tri(tar)]!=0) # not many non zero coefficients !
```

`targetGpos` *Computation of the target Gpos.*

Description

The $p \times p$ target Gpos is computed from the $n \times p$ data matrix. It it a modified version of target G. In particular, it completely ignores negative correlations and computes the mean correlation $\bar{r}$ using the positive ones only.

Usage

```
targetGpos(x, genegroups)
```

Arguments

`x`	A $n \times p$ data matrix.
`genegroups`	A list of genes obtained using the database KEGG, where each entry itself is a list of pathway names this genes belongs to. If a gene does not belong to any gene functional group, the entry is NA.

Value

A $p \times p$ matrix.

Author(s)

Monika Jelizarow and Vincent Guillemot

References

- J. Schaefer and K. Strimmer, 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statist. Appl. Genet. Mol. Biol. 4:32.
- M. Jelizarow, V. Guillemot, A. Tenenhaus, K. Strimmer, A.-L. Boulesteix, 2010. Over-optimism in bioinformatics: an illustration. Bioinformatics. Accepted.

See Also

`targetCor`, `targetF`, `targetG`, `targetGstar`, `targetGpos`.

Examples

```
# A short example on a toy dataset
# require(SHIP)
data(expl)
attach(expl)
tar <- targetGpos(x,genegroups)
which(tar[upper.tri(tar)]!=0) # not many non zero coefficients !
```

`targetGstar` *Computation of the target Gstar.*

Description

The $p \times p$ target Gstar is computed from the $n \times p$ data matrix. It it a modified version of target G. In particular, it involves two parameters for the correlation (a positive and a negative one) instead of the single parameter $\bar{r}$ in order to account for negatively correlated genes within the same pathway

Usage

```
targetGstar(x, genegroups)
```

Arguments

`x`	A $n \times p$ data matrix.
`genegroups`	A list of genes obtained using the database KEGG, where each entry itself is a list of pathway names this genes belongs to. If a gene does not belong to any gene functional group, the entry is NA.

Value

A $p \times p$ matrix.

Author(s)

Monika Jelizarow and Vincent Guillemot

References

- J. Schaefer and K. Strimmer, 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statist. Appl. Genet. Mol. Biol. 4:32.
- M. Jelizarow, V. Guillemot, A. Tenenhaus, K. Strimmer, A.-L. Boulesteix, 2010. Over-optimism in bioinformatics: an illustration. Bioinformatics. Accepted.

See Also

`targetCor`, `targetF`, `targetG`, `targetGstar`, `targetGpos`.

Examples

```
# A short example on a toy dataset
# require(SHIP)
data(expl)
attach(expl)
tar <- targetGstar(x,genegroups)
which(tar[upper.tri(tar)]!=0) # not many non zero coefficients !
```

Index