Package ‘robFitConGraph’

February 8, 2019

Type Package

Title Graph-Constrained Robust Covariance Estimation

Version 0.1.0

Author Stuart Watt, Daniel Vogel

Maintainer Stuart Watt <s.watt.15@aberdeen.ac.uk>

Description
Contains a single function named robFitConGraph() which includes two algorithms for robust estimation of scatter matrices subject to zero-constraints in its inverse. The methodology is described in Vogel & Tyler (2014) <doi:10.1093/biomet/asu041>. See robFitConGraph() function documentation for further details.

License GPL-3

Encoding UTF-8

LazyData true

RoxygenNote 6.1.1

LinkingTo Rcpp

Imports Rcpp, mvtnorm, MASS

NeedsCompilation yes

Repository CRAN

Date/Publication 2019-02-08 12:30:03 UTC

R topics documented:

robFitConGraph-package

A package for robust estimation of scatter matrices subject to zero-constraints in its inverse.

Description

robFitConGraph contains one function named after itself: robFitConGraph.

robFitConGraph

Graph-constrained robust scatter estimation.

Description

The function computes a robust estimate of a scatter matrix subject to zero-constraints in its inverse. The methodology is described in Vogel & Tyler (2014).

Usage

```
robFitConGraph(X, amat, df, tol = 1e-04, plug.in = TRUE,
  direct = FALSE)
```

Arguments

`X`	A data matrix with n rows and p columns, representing n observations and p variables. Elements of `X` must be numeric and n must be at least $p + 1$.
`amat`	A p times p matrix representing the adjacency matrix of a graphical model. `amat` must be symetric with numerical entries `0` or `1`. The entries on the diagonal are irrelevant, they may be anything.
`df`	the degrees of freedom of the t-distribution used (see Details below).
`tol`	tolerance for numerical convergence. Iteration stops if the maximal element-wise difference between two successive matrices is less than `tol`. Must be at least `10e-14`. Default is `10e-5`.
`plug.in`	logical. The function offers two types of estimates: the plug-in M-estimator and the direct M-estimator. If `plug.in` is `TRUE`, the plug-in estimate is computed. If `FALSE`, the direct M-estimator is computed. The plug-in estimator is faster, but has higher variance. Default is `TRUE`.
`direct`	logical. If `TRUE`, the direct estimate is computed, otherwise the plug-in estimate. Default is `FALSE`. In case of conflicting specifications of plug-in and direct, `plug.in` overrides `direct`.

Details

The function robFitConGraph implements the methodology of Vogel & Tyler (2014). Two types of estimates based on maximum likelihood estimation for the t-distribution are proposed: the direct estimate and the plug-in estimate. The direct estimate is referred to as graphical M-estimator in Vogel & Tyler (2014).

The plug-in estimate is two algorithms performed sequentially: First an unconstrained t-maximum likelihood estimate of scatter is computed (the same as cov.trob from MASS). This is then plugged into the Gaussian graphical model fitting routine (the same as fitConGraph from ggm). Specifically the algorithm 17.1 from Hastie, Tibshirani, Friedman (2009) is used.

The direct estimate is the actual maximum-likelihood estimator within the elliptical graphical model based on the elliptical t-distribution. The algorithm is an iteratively-reweighted least-squares algorithm, where the Gaussian graphical model fitting procedure is nested into the t-estimation iteration. The direct estimate therefore takes longer to compute, but the estimator has a better statistical efficiency for small sample sizes. Both estimators are asymptotically equivalent. The estimates tend to be very close to each other for large sample sizes.

Although robFitConGraph combines the functionality of fitConGraph and cov.trob and contains both as special cases, it uses only the latter function. The algorithms are largely implemented in C++.

Input and output of robFitConGraph are similar to fitConGraph from the package ggm. Some notable differences:

- fitConGraph takes as input the unconstrained covariance matrix, robFitConGraph takes the actual data.
- fitConGraph returns the deviance (test statistic) and the degrees of freedom r. The degrees of freedom r are the number of sub-diagonal 1-entries in the adjacency matrix. The deviance is compared to a chi-square distribution with r degrees of freedom to assess the model fit. These degrees of freedom r are unrelated to the the parameter df, which refers to the degrees of freedom of the t-distribution. The function robFitConGraph does return the deviance, but no degrees of freedom. The deviance must be divided by a constant (σ_1 in Vogel & Tyler, 2014) before comparing it to the χ^2_r-distribution.

Value

List with 5 elements:

Shat	p x p scatter matrix estimate
mu	numerical p-vector (robust location estimate)
em.it	integer. Number of iterations of the t-MLE computation.
ips.it	integer. In the case of the plug-in estimate, this is the number of iterations of the Gaussian graphical model fitting procedure (Algorithm 17.1) in Hastie et al 2004). In the case of the direct estimate, the Gaussian graphical model fitting is executed em.it times and the average number of iterations is returned.
dev	numerical. Value of the deviance test statistic D_n as defined in Vogel & Tyler (2014, p. 866 bottom). Comparing the model fitted against the full model.

Author(s)

Stuart Watt, Daniel Vogel

References

Vogel, D., Tyler, D. E. (2014): Robust estimators for nondecomposable elliptical graphical models, *Biometrika*, 101, 865-882

Hastie, T., Tibshirani, R. and Friedman, J. (2004). *The elements of statistical learning*. New York: Springer.

See Also

`fitConGraph` from package `ggm` for non-robust graph-constrained covariance estimation

`cov.trob` from package `MASS` for unconstrained p times p t-MLE scatter matrix

Examples

```
# --- build a graphical model ---

chordless.p.cycle <- function(rho,p){
  M <- diag(1,p)
  for (i in 1:(p-1)) M[i,i+1] <- M[i+1,i] <- -rho
  M[1,p] <- M[p,1] <- -rho
  return(M)
}
p <- 7                              # number of variables
rho <- 0.4                          # partial correlation
PCM <- chordless.p.cycle(rho,p)     # partial correlation matrix
SM <- cov2cor(solve(PCM))           # shape matrix (i.e covariance matrix up to scale)
model <- abs(sign(PCM))             # adjacency matrix of the chordless-7-cycle
# > model
#      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
# [1,]    1    1    0    0    0    0    1
# [2,]    1    1    1    0    0    0    0
# [3,]    0    1    1    1    0    0    0
# [4,]    0    0    1    1    1    0    0
# [5,]    0    0    0    1    1    1    0
# [6,]    0    0    0    0    1    1    1
# [7,]    1    0    0    0    0    1    1

# This is the cordless-7-cycle (p.872 Figure 1 (a) in Vogel & Tyler, 2014).
# All non-zero partial correlations are 0.4.
# The true covariance is (up to scale) 'SM'. This matrix is constructed such
# that it has zero entries in its inverse as specified by 'model'.


# --- generate data from the graphical model ---

n <- 50            # number of observations
```

```
df.data <- 3       # degrees of freedom
library(mvtnorm)   # for rmvt function
set.seed(918273)   # for reproducability
X <- rmvt(n=n,sigma=SM,df=df.data)

# X contains a data set of size n = 50 and dimension p = 7, sampled from the
# elliptical t-distribution with 3 degrees of freedom and shape matrix 'SM'


# --- compare estimates ---

# We compute three scatter estimates:

# 1) the direct graph-constrained t-MLE estimator:
S1 <- robFitConGraph(X, amat=model, df=df.data, plug.in=FALSE, direct=TRUE)$Shat
round(S1,d=2)

# 2) the plug-in graph-constrained t-MLE estimator:
S2 <- robFitConGraph(X, amat=model, df=df.data, plug.in=TRUE, direct=FALSE)$Shat
round(S2,d=2)

# 3) the saple covariance matrix:
round(cov(X),d=2)

# S1 and S2 are very similar. In Vogel & Tyler, 2014, it is shown that they
# are asymptotically equivalent as n goes to infinity.
# The sample covariance matrix substantially differs from S1 and S2. Note that
# S1 and S2 just estimate a multiple of the true covariance matrix (similarly
# SM is just proportional to the true covariance matrix). Therefore, consider
# correlation estimates based on the various scatter estimators:

# the true correlation matrix:
round(cov2cor(SM),d=2)

# sample correlations:
round(cov2cor(cov(X)),d=2)

# robust correlations based on the direct graph-constrained t-MLE:
round(cov2cor(S1),d=2)

# robust correlations based on the plug.in graph-constrained t-MLE:
round(cov2cor(S2),d=2)

# The correlation estimates based on S1 and S2 are close to the true
# correlations, whereas the sample correlations, again, differ strongly.
# Note: sample correlations are not asymptotically normal at the t3 distribution.
```

Index