Package 'CompareCausalNetworks'

June 24, 2015

Type Package

Title Interface to Diverse Estimation Methods of Causal Networks

Version 0.1.1

Date 2015-06-23

Author Christina Heinze <heinze@stat.math.ethz.ch>,
Nicolai Meinshausen <meinshausen@stat.math.ethz.ch>

Depends R (>= 3.1.0)

Imports Matrix

Maintainer Christina Heinze <heinze@stat.math.ethz.ch>

Description Unified interface for the estimation of causal networks,
including the methods 'backShift' (from package 'backShift'),
'bivariateANM' (bivariate additive noise model),
'bivariateCAM' (bivariate causal additive model),
'CAM' (causal additive model) (from package 'CAM'),
'hiddenICP' (invariant causal prediction with hidden variables),
'ICP' (invariant causal prediction) (from package 'InvariantCausalPrediction'),
'GES' (greedy equivalence search),
'GIES' (greedy interventional equivalence search),
'LINGAM', 'PC' (PC Algorithm),
'RFCI' (really fast causal inference)
(all from package 'pcalg') and regression.

License GPL

LazyData true

Suggests pcalg, InvariantCausalPrediction, glmnet, backShift, CAM,
kernlab, mgcv, testthat

URL https://github.com/christinaheinze/CompareCausalNetworks

BugReports https://github.com/christinaheinze/CompareCausalNetworks/issues

NeedsCompilation no

Repository CRAN

Date/Publication 2015-06-24 01:00:33

R topics documented:

`CompareCausalNetworks-package` *Compare estimates of causal graphs using a unified interface to various methods*

Description

Provides a unified interface to various causal graph estimation methods.

Details

Package:	CompareCausalNetworks
Type:	Package
Version:	0.1.0
Date:	2015-06-23
License:	GPL

The causal graphs can be estimated with function `getParents` and a stability-selection version is available at `getParentsStable`.

The supported methods are provided through the packages listed in `Suggests`. Thus, to use a particular method the corresponding package needs to be installed on your machine. To run the examples, most of these packages need to be installed.

Author(s)

Christina Heinze <heinze@stat.math.ethz.ch>, Nicolai Meinshausen <meinshausen@stat.math.ethz.ch>

`getParents` *Estimate the connectivity matrix of a causal graph*

Description

Estimates the connectivity matrix of a directed causal graph, using various possible methods. Supported methods at the moment are backShift, bivariateANM, bivariateCAM, CAM, hiddenICP, ICP, GES, GIES, LINGAM, PC, regression and RFCI.

Usage

```
getParents(X, environment = NULL, interventions = NULL,
  parentsOf = 1:ncol(X), method = c("ICP", "hiddenICP", "backShift", "pc",
  "LINGAM", "ges", "gies", "CAM", "rfci", "regression", "bivariateANM",
  "bivariateCAM")[1], alpha = 0.1, variableSelMat = NULL,
  excludeTargetInterventions = TRUE, onlyObservationalData = FALSE,
  indexObservationalData = 1, returnAsList = FALSE, pointConf = FALSE,
  setOptions = list(), directed = TRUE, verbose = FALSE)
```

Arguments

X	A (nxp)-data matrix with n observations of p variables.
environment	An optional vector of length n, where the entry for observation i is an index for the environment in which observation i took place (Simplest case: entries 1 for observational data and entries 2 for interventional data of unspecified type. Encoding for observational data can we changed with indexObservationalData). Is required for methods ICP, hiddenICP and backShift.
interventions	A optional list of length n. The entry for observation i is a numeric vector that specifies the variables on which interventions happened for observation i (a scalar if an intervention happened on just one variable and numeric(0) if no intervention occured for this observation). Is used for methods gies and CAM and will generate the vector environment if the latter is set to NULL. (However, this might generate too many different environments for some data sets, so a hand-picked vector environment is preferable). Is also used for ICP and hiddenICP to exclude interventions on the target variable of interest.
parentsOf	The variables for which we would like to estimate the parents. Default are all variables.
method	A string that specfies the method to use. The methods pc (PC-algorithm), LINGAM (LINGAM), ges (Greedy equivalence search), gies (Greedy interventional equivalence search) and rfci (Really fast causal inference) are imported from the package "pcalg" and are documented there in more detail, including the additional options that can be supplied via setOptions. The method CAM (Causal additive models) is documented in the package "CAM" and the methods ICP (Invariant causal prediction), hiddenICP (Invariant causal prediction with hidden variables) are from the package "InvariantCausalPrediction". The method backShift comes from the package "backShift". Finally, the methods bivariateANM and bivariateCAM are for now implemented internally but will hopefully be part of another package at some point in the near future.
alpha	The level at which tests are done. This leads to confidence intervals for ICP and hiddenICP and is used internally for pc and rfci.
variableSelMat	An optional logical matrix of dimension (pxp). An entry TRUE for entry (i,j) says that variable i should be considered as a potential parent for variable j and vice versa for FALSE. If the default value of NULL is used, all variables will be considered, but this can be very slow, especially for methods pc, ges, gies, rfci and CAM.

excludeTargetInterventions
: When looking for parents of variable k in 1,...,p, set to TRUE if observations where an intervention on variable k occured should be excluded. Default is TRUE.

onlyObservationalData
: If set to TRUE, only observational data is used. It will take the index in environment specified by indexObservationalData. If environment is NULL, all observations are used. Default is FALSE.

indexObservationalData
: Index in environment that encodes observational data. Default is 1.

returnAsList
: If set to TRUE, will return a list, where entry k is a list containing the estimated parents of variable k. The option directed will be ignored if set to TRUE. Default is FALSE.

pointConf
: If TRUE, numerical estimates will be returned if possible. For methods ICP and hiddenICP, these are the values in the individual confidence intervals (at chosen level alpha) that are closest to 0; for other methods these are point estimates. Some methods do not return numerical point estimates; for these the output will remain binary 0/1 (no-edge/edge). Default is FALSE.

setOptions
: A list that can take method-specific options; see the individual documentations of the methods for more options and their possible values.

directed
: If TRUE, an edge will be returned if and only if an edge has been detected to be directed (ie entry will be set to 0 for entry (j,k) if both j->k and k-> j are estimated). Ignored if not the whole graph is estimated or if returnAsList is TRUE.

verbose
: If TRUE, detailed output is provided.

Value

If option returnAsList is FALSE, a sparse matrix, where a 0 entry in position (j,k) corresponds to an estimate of "no edge" j -> k, while an entry 1 corresponds to an estimated egde. If option pointConf is TRUE, the 1 entries will be replaced by numerical values that are either point estimates of the causal coefficients or confidence bounds (see above). If option returnAsList is TRUE, a list will be returned. The k-th entry in the list is the numeric vector with the indices of the estimated parents of node k.

Author(s)

Christina Heinze <heinze@stat.math.ethz.ch>, Nicolai Meinshausen <meinshausen@stat.math.ethz.ch>

See Also

getParentsStable for stability selection-based estimation of the causal graph.

Examples

```
## load the backShift package for data generation and plotting functionality
if(!requireNamespace("backShift", quietly = TRUE))
 stop("The package 'backShift' is needed for the examples to
```

```
 work. Please install it.", call. = FALSE)

require(backShift)

# Simulate data with connectivity matrix A with assumptions
# 1) hidden variables present
# 2) precise location of interventions is assumed unknown
# 3) different environments can be distinguished

## simulate data
myseed <- 1

# sample size n
n <- 10000

# p=3 predictor variables and connectivity matrix A
p <- 3
labels <- c("1", "2", "3")
A <- diag(p)*0
A[1,2] <- 0.8
A[2,3] <- 0.8
A[3,1] <- -0.4

# divide data in 10 different environments
G <- 10

# simulate
simResult <- simulateInterventions(n, p, A, G, intervMultiplier = 3,
             noiseMult = 1, nonGauss = TRUE, hiddenVars = TRUE,
             knownInterventions = FALSE, fracVarInt = NULL, simulateObs = TRUE,
             seed = myseed)
X <- simResult$X
environment <- simResult$environment

## apply all  methods given in vector 'methods'
## (using all data pooled for pc/LINGAM/rfci/ges -- can be changed with option
## 'onlyObservationalData=TRUE')

methods <- c("backShift", "LINGAM") #c("pc", "rfci", "ges")

# select whether you want to run stability selection
stability <- FALSE

# arrange graphical output into a rectangular grid
sq <- ceiling(sqrt(length(methods)+1))
par(mfrow=c(ceiling((length(methods)+1)/sq),sq))

## plot and print true graph
cat("\n true graph is  ------  \n" )
print(A)
plotGraphEdgeAttr(A, plotStabSelec = FALSE, labels = labels, thres.point = 0,
 main = "TRUE GRAPH")
```

```
## loop over all methods and compute and print/plot estimate
for (method in methods){
  cat("\n result for method", method,"  ------  \n" )

  if(!stability){
    # Option 1): use this estimator as a point estimate
    Ahat <- getParents(X, environment, method=method, alpha=0.1, pointConf = TRUE)
  }else{
    # Option 2): use a stability selection based estimator
    # with expected number of false positives bounded by EV=2
    Ahat <- getParentsStable(X, environment, EV=2, method=method, alpha=0.1)
  }

  # print and plot estimate (point estimate thresholded if numerical estimates
  # are returned)
  print(Ahat)
  if(!stability)
    plotGraphEdgeAttr(Ahat, plotStabSelec = FALSE, labels = labels,
     thres.point = 0.05,
     main=paste("POINT ESTIMATE FOR METHOD\n", toupper(method)))
  else
    plotGraphEdgeAttr(Ahat, plotStabSelec = TRUE, labels = labels,
     thres.point = 0, main = paste("STABILITY SELECTION
     ESTIMATE\n FOR METHOD", toupper(method)))
 }
```

---

getParentsStable *Estimate the connectivity matrix of a causal graph using stability selection.*

Description

Estimates the connectivity matrix of a directed causal graph, using various possible methods. Supported methods at the moment are backShift, bivariateANM, bivariateCAM, CAM, hiddenICP, ICP, GES, GIES, LINGAM, PC, regression and RFCI. Uses stability selection to select an appropriate sparseness.

Usage

```
getParentsStable(X, environment, interventions = NULL, EV = 1,
  nodewise = TRUE, threshold = 0.75, nsim = 100,
  sampleSettings = 1/sqrt(2), sampleObservations = 1/sqrt(2),
  parentsOf = 1:ncol(X), method = c("ICP", "hiddenICP", "backShift", "pc",
  "LINGAM", "ges", "gies", "CAM", "rfci", "regression", "bivariateANM",
  "bivariateCAM")[1], alpha = 0.1, variableSelMat = NULL,
  excludeTargetInterventions = TRUE, onlyObservationalData = FALSE,
  indexObservationalData = NULL, setOptions = list(), verbose = FALSE)
```

Arguments

X	A (nxp)-data matrix with n observations of p variables.
environment	An optional vector of length n, where the entry for observation i is an index for the environment in which observation i took place (simplest case entries 1 for observational data and entries 2 for interventional data of unspecified type). Is required for methods ICP, hiddenICP, backShift.
interventions	A optional list of length n. The entry for observation i is a numeric vector that specifies the variables on which interventions happened for observation i (a scalar if an intervention happened on just one variable and numeric(0) if no intervention occured for this observation). Is used for method gies but will generate the vector environment if this is set to NULL (even though it might generate too many different environments for some data so a hand-picked vector environment is preferable). Is also used for ICP and hiddenICP to exclude interventions on the target variable of interest.
EV	A bound on the expected number of falsely selected edges.
nodewise	If FALSE, stability selection retains for each subsample the largest overall entries in the connectivity matrix. If TRUE, values are ordered row- and node-wise first and then the largest entries in each row and column are retained. Error control is valid (under exchangeability assumption) in both cases. The latter setting TRUE is perhaps more robust and is the default.
threshold	The empirical selection frequency in (0.5,1) under subsampling that needs to be surpassed for an edge to be selected.
nsim	The number of resamples for stability selection.
sampleSettings	The fraction of different environments to resample in each resampling (at least two different environments will be selected so the argument is without effect if there are just two different environments in total).
sampleObservations	The fraction of samples to resample in each environment.
parentsOf	The variables for which we would like to estimate the parents. Default are all variables.
method	A string that specfies the method to use. The methods pc (PC-algorithm), LINGAM (LINGAM), ges (Greedy equivalence search), gies (Greedy interventional equivalence search) and rfci (Really fast causal inference) are imported from the package "pcalg" and are documented there in more detail, including the additional options that can be supplied via setOptions. The method CAM (Causal additive models) is documented in the package "CAM" and the methods ICP (Invariant causal prediction), hiddenICP (Invariant causal prediction with hidden variables) are from the package "InvariantCausalPrediction". The method backShift comes from the package "backShift". Finally, the methods bivariateANM and bivariateCAM are for now implemented internally but will hopefully be part of another package at some point in the near future.
alpha	The level at which tests are done. This leads to confidence intervals for ICP and hiddenICP and is used internally for pc and rfci.
variableSelMat	An optional logical matrix of dimension (pxp). An entry TRUE for entry (i,j) says that variable i should be considered as a potential parent for variable j and

vice versa for FALSE. If the default value of NULL is used, all variables will be considered, but this can be very slow, especially for methods pc, ges, gies, rfci and CAM.

excludeTargetInterventions
: When looking for parents of variable k in 1,...,p, set to TRUE if observations where an intervention on variable k occured should be excluded. Default is TRUE.

onlyObservationalData
: If set to TRUE, only observational data is used. It will take the index in environment specified by indexObservationalData. If environment is NULL, all observations are used. Default is FALSE.

indexObservationalData
: Index in environment that encodes observational data. Default is 1.

setOptions
: A list that can take method-specific options; see the individual documentations of the methods for more options and their possible values.

verbose
: If TRUE, detailed output is provided.

Value

A sparse matrix, where a 0 entry in (j,k) corresponds to an estimate of 'no edge' j -> parentsOf[k]. Entries between 0 and 100 give the selection percentage of this edge over all resamples (set to 0 if below critical threshold) and all non-zero values are considered as selected edges.

Author(s)

Nicolai Meinshausen <meinshausen@stat.math.ethz.ch>, Christina Heinze <heinze@stat.math.ethz.ch>

References

Stability selection (2010): N. Meinshausen and P. Buhlmann, Journal of the Royal Statistical Society: Series B, 72, 417-473

See Also

getParents for the underlying point-estimate of the causal graph.

---

simData_unknownShiftInterventions

Data from a causal cyclic model with shift interventions

Description

A dataset to run the tests.

Usage

```
simData_unknownShiftInterventions
```

Format

A list created by `simulateInterventions`. All inputs are contained in the list element `configs`. For details see the help page of `simulateInterventions`.

Index