Package 'CovSelHigh'

July 3, 2017

Version 1.1.1

Author Jenny Häggström

Maintainer Jenny Häggström <jenny.haggstrom@umu.se>

Depends R (>= 2.14.0)

Imports bnlearn,MASS, bindata, Matching, doRNG, glmnet, randomForest,foreach,xtable, doParallel, bartMachine, tmle

Title Model-Free Covariate Selection in High Dimensions

Description Model-free selection of covariates in high dimensions under unconfoundedness for situations where the parameter of interest is an average causal effect. This package is based on model-free backward elimination algorithms proposed in de Luna, Waernbaum and Richardson (2011) <DOI:10.1093/biomet/asr041> and VanderWeele and Shpitser (2011) <DOI:10.1111/j.1541-0420.2011.01619.x>. Confounder selection can be performed via either Markov/Bayesian networks, random forests or LASSO.

License GPL-3

Encoding UTF-8

NeedsCompilation no

Repository CRAN

Date/Publication 2017-07-03 09:35:40 UTC

R topics documented:

cov.sel.high *Model-Free Covariate Selection in High Dimensions*

Description

Model-free selection of covariates in high dimensions under unconfoundedness for situations where the parameter of interest is an average causal effect. This package is based on model-free backward elimination algorithms proposed in de Luna, Waernbaum and Richardson (2011) and VanderWeele and Shpitser (2011). Confounder selection can be performed via either Markov/Bayesian networks, random forests or LASSO.

Usage

```
cov.sel.high(T=NULL, Y=NULL, X=NULL,type=c("mmpc","mmhc","rf","lasso"),
                  betahat=TRUE, parallel=FALSE, Simulate=TRUE,N=NULL, Setting=1,
                    rep=1, Models=c("Linear", "Nonlinear", "Binary"),
                    alpha=0.05, mmhc_score=c("aic","bic"))
```

Arguments

T	A vector, containing 0 and 1, indicating a binary treatment variable.
Y	A vector of observed outcomes.
X	A matrix or data frame containing columns of covariates. The covariates may be a mix of continuous, unordered discrete (to be specified in the data frame using factor), and ordered discrete (to be specified in the data frame using ordered).
type	The type of method used for selection. The networks algorithms are "mmpc" for min-max parents and children (Markov network) and "mmhc" for max-min hill climbing (Bayesian network). Other available methods are random forests, "rf", and LASSO, "lasso".
betahat	If betahat=TRUE the average treatment effect for each selected subset and the full covariate vector is estimated using propensity score matching (PSM) via the function Match and using targeted maximum likelihood estimation (TMLE) via the function tmle.
parallel	If parallel=TRUE and there is a registered parallel backend then the computation will be parallelized. Default is parallel=FALSE.
Simulate	If data is to be simulated according to one of the designs in Häggström (2017) then Simulate should be set to TRUE.
N	If Simulate=TRUE, N is the number of observations to be simulated.
Setting	If Simulate=TRUE, Setting is the simulation setting to be used. Unconfoundedness holds given X if Setting=1. M-bias given X if Setting=2.
rep	If Simulate=TRUE, rep is the number of replications to be simulated.
Models	If Simulate=TRUE, Models is the type of outcome models to be used, options are "Linear", "Nonlinear" and "Binary".

alpha	A numeric value, the target nominal type I error rate (tuning parameter) for "mmpc" and "mmhc".
mmhc_score	The score to use for "mmhc".

Details

See Häggström (2017).

Value

cov.sel.high returns a list with the following content:

X.T	The set of covariates targeting the subset containing all causes of T.
Q.0	The set of covariates targeting the subset of X.T which is also associated with Y given T=0, the response in the control group.
Q.1	The set of covariates targeting the subset of X.T which is also associated with Y given T=1, the response in the treatment group.
Q	Union of Q.0 and Q.1.
X.0	The set of covariates targeting the subset containing all causes of Y given T=0.
X.1	The set of covariates targeting the subset containing all causes of Y given T=1.
X.Y	Union of X.0 and X.1.
Z.0	The set of covariates targeting the subset of X.0 which is also associated with T.
Z.1	The set of covariates targeting the subset of X.1 which is also associated with T.
Z	Union of Z.0 and Z.1.
X.TY	Union of X.T and X.Y, the set of covariates targeting the subset containing all causes of T and Y.
cardinalities	The cardinalities of each selected subset.
est_psm	The PSM estimate of the average causal effect, for the full covariate vector and each selected subset.
se_psm	The Abadie-Imbens standard error for the PSM estimate of the average causal effect, for the full covariate vector and each selected subset.
est_tmle	The TMLE estimate of the average causal effect, for the full covariate vector and each selected subset.
se_psm	The influence-curve based standard error for the TMLE estimate of the average causal effect, for the full covariate vector and each selected subset.
N	The number of observations.
Setting	The Setting used.
rep	The number of replications.
Models	Models used.
type	type used.
alpha	alpha used.
mmhc_score	score used.
varnames	Variable names of the used data.

Note

Depending on the method type specified `cov.sel.high` calls one of the functions `mmpc`, `mmhc`, `randomForest`, `cv.glmnet` and, if betahat=TRUE, `Match` and `tmle`, therefore the packages `bnlearn`, `randomForest`, `glmnet`, `Matching` and `tmle` are required.

Author(s)

Jenny Häggström, <jenny.haggstrom@umu.se>

References

de Luna, X., I. Waernbaum, and T. S. Richardson (2011). Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika 98*. 861-875

Häggström, J. (2017). Data-Driven Confounder Selection via Markov and Bayesian Networks. *ArXiv e-prints*.

Nagarajan, R., M. Scutari and S. Lebre. (2013) *Bayesian Networks in R with Applications in Systems Biology*. Springer, New York. ISBN 978-1461464457.

Scutari, M. (2010). Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software, 35*, 1-22. URL http://www.jstatsoft.org/v35/i03/.

Sekhon, J.S. (2011). Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R. *Journal of Statistical Software, 42*, 1-52. URL http://www.jstatsoft.org/v42/i07/.

See Also

`bnlearn-package`, `randomForest`, `cv.glmnet`, `Match` and `tmle`

Examples

```
##Use simulated data, select subsets using mmpc
ans<-cov.sel.high(type="mmpc",N=1000, rep=2, Models="Linear", betahat=FALSE, mmhc_score="aic")


##Use simulated data, select subsets using mmpc and estimate ACEs, parallell version
#library(doParallel)
#library(doRNG)
#cl <- makeCluster(4)
#registerDoParallel(cl)
#ans<-cov.sel.high(type="mmpc", parallel=TRUE,  N=500, rep=10, Models="Linear", mmhc_score="aic")
#stopCluster(cl)
```

cov.sel.high.lasso *cov.sel.high.lasso*

Description

Function called by cov.sel.high if type="lasso". Not meant to be used on its own.

Usage

```
cov.sel.high.lasso(Y, X, minscreen = 2, ...)
```

Arguments

Y	Outcome variable or treatment variable.
X	A matrix or data frame containing columns of covariates and all functions of covariates e.g. interactions that should be included in the lasso model.
minscreen	The minimum number of columns in X that should be selected.
...	Additional arguments passed on to cv.glmnet.

Details

See cv.glmnet.

Value

cov.sel.high.lasso returns a logical vector of the same length as the number of columns in X. The positions of values in the vector refers to the (functions of) covariates in the corresponding X columns. Value TRUE implies that (the function of) the covariate has a corresponding coefficient not equal to zero.

Author(s)

Jenny Häggström, <jenny.haggstrom@umu.se>

See Also

cv.glmnet

cov.sel.high.rf *cov.sel.high.rf*

Description

Function called by cov.sel.high if type="rf". Not meant to be used on its own.

Usage

```
cov.sel.high.rf(Y, X, threshold = 0.25, ntree = 1000, ...)
```

Arguments

Y	Outcome variable or treatment variable.
X	A matrix or data frame containing columns of covariates.
threshold	Variable importance threshold, see Value.
ntree	Number of trees to grow. This should not be set to too small a number, to ensure that every input row gets predicted at least a few times.
...	Additional arguments passed on to randomForest.

Details

See randomForest.

Value

cov.sel.high.rf returns a logical vector of the same length as the number of columns in X. The positions of values in the vector refers to the covariates in the corresponding X columns. Value TRUE implies that the covariate has a variable importance value of more than threshold*the largest observed variable importance value.

Author(s)

Jenny Häggström, <jenny.haggstrom@umu.se>

See Also

randomForest

`cov.sel.high.sim` *Simulate Example Data for CovSelHigh*

Description

Function used internally by cov.sel.high to simulates example data used in Häggström (2016).

Usage

```
cov.sel.high.sim(N, Setting, rep, Models)
```

Arguments

`N`	The number of observations to be simulated.
`Setting`	The simulation setting to be used. Unconfoundedness holds given X if Setting=1. M-bias given X if Setting=2.
`rep`	The number of replications to be simulated.
`Models`	The type of outcome models to be used, options are `"Linear"`, `"Nonlinear"` and `"Binary"`.

Value

`cov.sel.high` returns a list with the following content:

`dat`	A data frame with simulated data.

Note

`cov.sel.high.sim` calls the functions `rmvbin` and `mvrnorm`.

Author(s)

Jenny Häggström, <jenny.haggstrom@umu.se>

References

Häggström, J. (2017). Data-Driven Confounder Selection via Markov and Bayesian Networks. *ArXiv e-prints*.

cov.sel.high.sim.res *Summarize Simulation Results for CovSelHigh*

Description

Function used to summarize results from cov.sel.high when simulated data is used.

Usage

```
cov.sel.high.sim.res(object)
```

Arguments

object	A list returned from cov.sel.high.

Value

cov.sel.high.sim.res returns a list with the following content:

resmat	A matrix with the following columns (columns related to average causal effect estimation are only present when betahat=TRUE): XTuc Quc XYuc Zuc XTYuc SinXT SinQ SinXY SinZ SinXTY XTeqS QeqS XYeqS ZeqS XTYeqS cards betahatest_psm betahatse_psm betahat_cicov_psm ciL_psm ciU_psm betahatest_tmle betahatse_tmle betahat_cicov_tmle ciL_tmle ciU_tmle ciwidth_psm ciwidth_tmle
summary_resmat	A list with the following columns (columns related to average causal effect estimation are only present when betahat=TRUE): Subset_selection Median_cardinality Betahat_bias_psm Betahat_sd_psm Betahat_mse_psm Betahat_CI_coverage_psm Betahat_CI_width_psm Betahat_mean_lower_CI_psm Betahat_mean_upper_CI_psm Betahat_bias_psm Betahat_sd_psm Betahat_mse_psm Betahat_CI_coverage_psm Betahat_CI_width_psm Betahat_mean_lower_CI_psm Betahat_mean_upper_CI_psm
xtable1	LaTeX table summarizing the results.
xtable2	LaTeX table summarizing the results.

Author(s)

Jenny Häggström, <jenny.haggstrom@umu.se>

References

Häggström, J. (2017). Data-Driven Confounder Selection via Markov and Bayesian Networks. *ArXiv e-prints*.

Index