Package 'RobustIV'

July 21, 2025

Type Package

Title Robust Instrumental Variable Methods in Linear Models

Version 0.2.5

Description Inference for the treatment effect with possibly invalid instrumental variables via TSHT('Guo et al.' (2016) <doi:10.48550/arXiv.1603.05224>) and SearchingSampling('Guo' (2021) <doi:10.48550/arXiv.2104.06911>), which are effective for both low- and high-dimensional covariates and instrumental variables; test of endogeneity in high dimensions ('Guo et al.' (2016) <doi:10.48550/arXiv.1609.06713>).

License GPL-3

Encoding UTF-8

LazyData true

RoxygenNote 7.2.2

URL https://github.com/zijguo/RobustIV

Imports glmnet, MASS, Matrix, igraph, intervals, CVXR

NeedsCompilation no

Depends R (>= 2.10)

Author Taehyeon Koo [aut],
Zhenyu Wang [aut],
Hyunseung Kang [ctb],
Dylan Small [ctb],
Zijian Guo [aut, cre, cph]

Maintainer Zijian Guo <zijguo@stat.rutgers.edu>

Repository CRAN

Date/Publication 2022-12-20 01:10:02 UTC

Contents

`endo.test` *Endogeneity test in high dimensions*

Description

Conduct the endogeneity test with high dimensional and possibly invalid instrumental variables.

Usage

```
endo.test(
  Y,
  D,
  Z,
  X,
  intercept = TRUE,
  invalid = FALSE,
  method = c("Fast.DeLasso", "DeLasso", "OLS"),
  voting = c("MP", "MaxClique"),
  alpha = 0.05,
  tuning.1st = NULL,
  tuning.2nd = NULL
)
```

Arguments

`Y`	The outcome observation, a vector of length n.
`D`	The treatment observation, a vector of length n.
`Z`	The instrument observation of dimension $n \times p_z$.
`X`	The covariates observation of dimension $n \times p_x$.
`intercept`	Whether the intercept is included. (default = `TRUE`)
`invalid`	If `TRUE`, the method is robust to the presence of possibly invalid IVs; If `FALSE`, the method assumes all IVs to be valid. (default = `FALSE`)
`method`	The method used to estimate the reduced form parameters. `"OLS"` stands for ordinary least squares, `"DeLasso"` stands for the debiased Lasso estimator, and `"Fast.DeLasso"` stands for the debiased Lasso estimator with fast algorithm. (default = `"Fast.DeLasso"`)
`voting`	The voting option used to estimate valid IVs. `'MP'` stnads for majority and plurality voting, `'MaxClique'` stands for maximum clique in the IV voting matrix. (default = `'MP'`)
`alpha`	The significance level for the confidence interval. (default = `0.05`)
`tuning.1st`	The tuning parameter used in the 1st stage to select relevant instruments. If `NULL`, it will be generated data-dependently, see Details. (default=`NULL`)
`tuning.2nd`	The tuning parameter used in the 2nd stage to select valid instruments. If `NULL`, it will be generated data-dependently, see Details. (default=`NULL`)

Details

When `voting = MaxClique` and there are multiple maximum cliques, the null hypothesis is rejected if one of maximum clique rejects the null. As for tuning parameter in the 1st stage and 2nd stage, if do not specify, for method "OLS" we adopt $\sqrt{\log n}$ for both tuning parameters, and for other methods we adopt $\max\left(\sqrt{2.01\log p_z}, \sqrt{\log n}\right)$ for both tuning parameters.

Value

`endo.test` returns an object of class "endotest", which is a list containing the following components:

`Q`	The test statistic.
`Sigma12`	The estimated covaraince of the regression errors.
`SHat`	The set of selected relevant IVs.
`VHat`	The set of selected vaild IVs.
`p.value`	The p-value of the endogeneity test.
`check`	The indicator that $H_0 : \Sigma_{12} = 0$ is rejected.

References

Guo, Z., Kang, H., Tony Cai, T. and Small, D.S. (2018), Testing endogeneity with high dimensional covariates, *Journal of Econometrics*, Elsevier, vol. 207(1), pages 175-187.

Examples

```
n = 500; L = 11; s = 3; k = 10; px = 10;
beta = 1; gamma = c(rep(1,k),rep(0,L-k))
phi<-(1/px)*seq(1,px)+0.5; psi<-(1/px)*seq(1,px)+1
epsilonSigma = matrix(c(1,0.8,0.8,1),2,2)
Z = matrix(rnorm(n*L),n,L)
X = matrix(rnorm(n*px),n,px)
epsilon = MASS::mvrnorm(n,rep(0,2),epsilonSigma)
D =  0.5 + Z %*% gamma + X %*% psi + epsilon[,1]
Y = -0.5 + Z %*% c(rep(1,s),rep(0,L-s)) + D * beta + X %*% phi + epsilon[,2]
endo.test.model <- endo.test(Y,D,Z,X,invalid = TRUE)
summary(endo.test.model)
```

`lineardata` *lineardata*

Description

Psuedo data provided by Youjin Lee, which is generated mimicing the structure of Framingham Heart Study data.

Usage

```
data(lineardata)
```

Format

A data.frame with 1445 observations on 12 variables:

- **Y:** The globulin level.
- **D:** The LDL-C level.
- **Z.1:** SNP genotypes.
- **Z.2:** SNP genotypes.
- **Z.3:** SNP genotypes.
- **Z.4:** SNP genotypes.
- **Z.5:** SNP genotypes.
- **Z.6:** SNP genotypes.
- **Z.7:** SNP genotypes.
- **Z.8:** SNP genotypes.
- **age:** the age of the subject.
- **sex:** the sex of the subject.

Source

The Framingham Heart Study data supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University.

Examples

```
data(lineardata)
```

SearchingSampling *Searching-Sampling*

Description

Construct Searching and Sampling confidence intervals for the causal effect, which provides the robust inference of the treatment effect in the presence of invalid instrumental variables in both low-dimensional and high-dimensional settings. It is robust to the mistakes in separating valid and invalid instruments.

Usage

```
SearchingSampling(
  Y,
  D,
  Z,
  X = NULL,
  intercept = TRUE,
  method = c("OLS", "DeLasso", "Fast.DeLasso"),
  robust = TRUE,
  Sampling = TRUE,
  alpha = 0.05,
  CI.init = NULL,
  a = 0.6,
  rho = NULL,
  M = 1000,
  prop = 0.1,
  filtering = TRUE,
  tuning.1st = NULL,
  tuning.2nd = NULL
)
```

Arguments

Y	The outcome observation, a vector of length n.
D	The treatment observation, a vector of length n.
Z	The instrument observation of dimension $n \times p_z$.
X	The covariates observation of dimension $n \times p_x$.
intercept	Whether the intercept is included. (default = TRUE)
method	The method used to estimate the reduced form parameters. "OLS" stands for ordinary least squares, "DeLasso" stands for the debiased Lasso estimator, and "Fast.DeLasso" stands for the debiased Lasso estimator with fast algorithm. (default = "OLS")
robust	If TRUE, the method is robust to heteroskedastic errors. If FALSE, the method assumes homoskedastic errors. (default = TRUE)
Sampling	If TRUE, use the proposed sampling method; else use the proposed searching method. (default=TRUE)
alpha	The significance level (default=0.05)
CI.init	An initial range for beta. If NULL, it will be generated automatically. (default=NULL)
a	The grid size for constructing beta grids. (default=0.6)
rho	The shrinkage parameter for the sampling method. (default=NULL)
M	The resampling size for the sampling method. (default = 1000)
prop	The proportion of non-empty intervals used for the sampling method. (default=0.1)

filtering	Filtering the resampled data or not. (default=TRUE)
tuning.1st	The tuning parameter used in the 1st stage to select relevant instruments. If NULL, it will be generated data-dependently, see Details. (default=NULL)
tuning.2nd	The tuning parameter used in the 2nd stage to select valid instruments. If NULL, it will be generated data-dependently, see Details. (default=NULL)

Details

When robust = TRUE, the method will be input as 'OLS'. For rho, M, prop, and filtering, they are required only for Sampling = TRUE. As for tuning parameter in the 1st stage and 2nd stage, if do not specify, for method "OLS" we adopt $\sqrt{\log n}$ for both tuning parameters, and for other methods we adopt $\max\left(\sqrt{2.01\log p_z}, \sqrt{\log n}\right)$ for both tuning parameters.

Value

SearchingSampling returns an object of class "SS", which is a list containing the following components:

ci	1-alpha confidence interval for beta.
SHat	The set of selected relevant IVs.
VHat	The initial set of selected relevant and valid IVs.
check	The indicator that the plurality rule is satisfied.

References

Guo, Z. (2021), Causal Inference with Invalid Instruments: Post-selection Problems and A Solution Using Searching and Sampling, Preprint *arXiv:2104.06911*.

Examples

```
data("lineardata")
Y <- lineardata[,"Y"]
D <- lineardata[,"D"]
Z <- as.matrix(lineardata[,c("Z.1","Z.2","Z.3","Z.4","Z.5","Z.6","Z.7","Z.8")])
X <- as.matrix(lineardata[,c("age","sex")])
Searching.model <- SearchingSampling(Y,D,Z,X, Sampling = FALSE)
summary(Searching.model)
Sampling.model <- SearchingSampling(Y,D,Z,X)
summary(Sampling.model)
```

TSHT *Two-Stage Hard Thresholding*

Description

Perform Two-Stage Hard Thresholding method, which provides the robust inference of the treatment effect in the presence of invalid instrumental variables.

Usage

```
TSHT(
  Y,
  D,
  Z,
  X,
  intercept = TRUE,
  method = c("OLS", "DeLasso", "Fast.DeLasso"),
  voting = c("MaxClique", "MP", "Conservative"),
  robust = TRUE,
  alpha = 0.05,
  tuning.1st = NULL,
  tuning.2nd = NULL
)
```

Arguments

Y	The outcome observation, a vector of length n.
D	The treatment observation, a vector of length n.
Z	The instrument observation of dimension $n \times p_z$.
X	The covariates observation of dimension $n \times p_x$.
intercept	Whether the intercept is included. (default = TRUE)
method	The method used to estimate the reduced form parameters. "OLS" stands for ordinary least squares, "DeLasso" stands for the debiased Lasso estimator, and "Fast.DeLasso" stands for the debiased Lasso estimator with fast algorithm. (default = "OLS")
voting	The voting option used to estimate valid IVs. 'MP' stands for majority and plurality voting, 'MaxClique' stands for finding maximal clique in the IV voting matrix, and 'Conservative' stands for conservative voting procedure. Conservative voting is used to get an initial estimator of valid IVs in the Searching-Sampling method. (default= 'MaxClique').
robust	If TRUE, the method is robust to heteroskedastic errors. If FALSE, the method assumes homoskedastic errors. (default = TRUE)
alpha	The significance level for the confidence interval. (default = 0.05)
tuning.1st	The tuning parameter used in the 1st stage to select relevant instruments. If NULL, it will be generated data-dependently, see Details. (default=NULL)

`tuning.2nd` The tuning parameter used in the 2nd stage to select valid instruments. If `NULL`, it will be generated data-dependently, see Details. (default=`NULL`)

Details

When `robust = TRUE`, the `method` will be input as `'OLS'`. When `voting = MaxClique` and there are multiple maximum cliques, `betaHat`,`beta.sdHat`,`ci`, and `VHat` will be list objects where each element of list corresponds to each maximum clique. As for tuning parameter in the 1st stage and 2nd stage, if do not specify, for method "OLS" we adopt $\sqrt{\log n}$ for both tuning parameters, and for other methods we adopt $\max\left(\sqrt{2.01\log p_z}, \sqrt{\log n}\right)$ for both tuning parameters.

Value

`TSHT` returns an object of class "TSHT", which is a list containing the following components:

`betaHat`	The estimate of treatment effect.
`beta.sdHat`	The estimated standard error of `betaHat`.
`ci`	The 1-alpha confidence interval for `beta`.
`SHat`	The set of selected relevant IVs.
`VHat`	The set of selected relevant and valid IVs.
`voting.mat`	The voting matrix.
`check`	The indicator that the majority rule is satisfied.

References

Guo, Z., Kang, H., Tony Cai, T. and Small, D.S. (2018), Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting, *J. R. Stat. Soc. B*, 80: 793-815.

Examples

```
data("lineardata")
Y <- lineardata[,"Y"]
D <- lineardata[,"D"]
Z <- as.matrix(lineardata[,c("Z.1","Z.2","Z.3","Z.4","Z.5","Z.6","Z.7","Z.8")])
X <- as.matrix(lineardata[,c("age","sex")])
TSHT.model <- TSHT(Y=Y,D=D,Z=Z,X=X)
summary(TSHT.model)
```

Index