Package 'IVGC'

March 14, 2019

Type Package

Title Two-step Estimation Using IV Regression and Graph-constrained
Regularization Method

Version 1.0.1

Date 2019-03-13

Author Xu Liu [aut,cre],
Bin Gao [aut],
Hongzhe Li [aut],
Yuehua Cui [aut],
Yiming Liu [ctb]

Maintainer Xu Liu <liu.xu@shufe.edu>

Depends R (>= 3.1.0)

Imports glmnet

Description Efficient two-step procedure to achieve variable selection and estimation under an IV regression framework by incorporating gene network information. The method is applicable in both low and high dimensional situations and the cyclical coordinate descent algorithm is used to solve them.

License GPL-2

LazyData true

NeedsCompilation yes

Repository GitHub

URL https://github.com/xliusufe/IVGC

Encoding UTF-8

Archs i386, x64

R topics documented:

IVGC-package   *Two-step Estimation Using IV Regression and Graph-constrained Regularization Method*

Description

Efficient two-step procedure to achieve variable selection and estimation under an IV regression framework by incorporating gene network information. The method is applicable in both low and high dimensional situations and the cyclical coordinate descent algorithm is used to solve them.

Details

Package:	IVGC
Type:	Package
Version:	1.0.1
Date:	2019-03-13
License:	GPL-2

References

Gao, B., Liu, X., Li, H. and Cui, Y. (2019). Integrative Analysis of Genetical Genomics Data Incorporating Network Structures. Manuscript.

---

main   *Fit an IV regression model with lasso and graph-constrained regularization*

Description

This function is a two-step procedure to achieve variable selection and estimation under an IV regression framework by incorporating gene regulatory network structures. Because of high correlations between coefficients, in the first step, the LASSO-type regression is applied to estimate the coefficient matrix. And in the second step, two penalties (LASSO and graph-constrained penalties) are applied to encourage the shrinking coefficients of a pair of connected variable in a network, thus achieving a grouping effect.

Usage

```
main(y, x, S, G,is.graph="IVGC",nlambda=50,nalpha=50,nfold=5,epsilon=1e-6)
```

Arguments

y	The response, a vector of size n
x	The expression data dimension $n \times p$
S	The SNP variable with dimension $n \times q$

G	The graph structure of X with dimension $p \times p$
is.graph	The method including "IVGC", "IV", "both". "IVGC" denotes IV regression incorporating network structures (Gao et al. 2019); "IV" denotes IV regression without network structures; "both" outputs the results by "IVGC" and "IV". Default is "IVGC".
nlambda	The length of tuning parameter λ which is selected automatically. Default is 50.
nalpha	The length of tuning parameter α which is selected automatically. Default is 50.
nfold	The number of folds for the tuning selection by CV. Default is 5.
epsilon	The perturbation when the proposed algorithm is used. Default is epsilon=1e-6.

Details

The simulated data set is saved in simulatedData.RData, which is list with four values Y, X, gene (S) and graph (G).

Value

beta_ivgc	the beta estimator by the proposed method
beta_iv	the beta estimator by IV method without graph structure
Gamma_hat	the Gamma estimator by grouped lasso

Author(s)

Gao Bin, Xu Liu (liu.xu@sufe.edu.cn), Hongzhe Li, and Yuehua Cui

References

Gao, B., Liu, X., Li, H. and Cui, Y. (2019+) Integrative Analysis of Genetical Genomics Data Incorporating Network Structures. Manuscript.

Examples

```
data(simulatedData)
y <- simulatedData$Y
x <- simulatedData$X
S <- simulatedData$gene
G <- simulatedData$graph
fit <- main(y, x, S, G)
```

---

simulatedData   *Genetical Genomics Data Incorporating Network Structures*

Description

This is a simulated data on gene expression. Certain correlation structures on genes are also imposed to show the impact of network structures on variable selection and estimation.

Usage

```
data("simulatedData")
```

Details

We simulated a total of p variables of gene expression and considered three group structures on coefficients with five variables in each group. Within each group, the variables are correlated. The strength of the correlation was controlled by the number of effective SNPs they had in common. The simulated dataset is saved in simulatedData.RData, which is a list with four values Y, X, gene (S) and graph (G). For this data, we set SNPnum=3, $p = q = 100$, $n = 600$ and $\beta_k = 0.5$ for $k = 1, 2, ..., 10$, where SNPnum is the number of common SNPs to achieve different correlation strengths.

Y: the response, a vector of size 600

X: the expression data dimension 600×100

gene: the SNP variable with dimension 600×100

graph: the graph structure of X with dimension 100×100

References

Gao, B., Liu, X., Li, H. and Cui, Y. (2019+) Integrative Analysis of Genetical Genomics Data Incorporating Network Structures. Manuscript.

Examples

```
data(simulatedData)
y <- simulatedData$Y
x <- simulatedData$X
S <- simulatedData$gene
G <- simulatedData$graph
```

Index