HomUHet Vignette

Type Package

Title Identifying and Separating Homogeneous and Heterogeneous Predictors

Version: 0.0.0.9000

Date: 2020-01-22

Depends tidyverse, glmnet, HDeconometrics, gglasso

Suggests knitr, rmarkdown

Description This package contains functions to identify and separate predictors with homogeneous or heterogeneous effects across ultra high-dimensional datasets.

VignetteBuilder knitr

Encoding UTF-8

RoxygenNote 7.1.1

Author Pei S. Yang [aut, cre]

Maintainer Pei S. Yang <yang.1736@osu.edu>

Repository Git Hub

1.1 Introduction

The presence of biological homogeneity and heterogeneity is a phenomenon that many researchers seek to understand. Often times, researchers want to identify biomarkers having heterogeneous or homogeneous effects on an outcome. However, identifying the biological heterogeneity and homogeneity in the effects of predictors while maintaining computational efficiency and accounting for dependency structure within correlated data is challenging and urgently needed.

HomUHet is developed to address the problem by using individual level data to fit penalized linear regression models.

In this package, we provide the following functions

- HomUHet fits penalized linear regression models in two steps and performs variable selection, which provides the names of identified predictors with homogeneous or heterogeneous effects, as well as their estimated coefficients.
- HomUHet.sim simulate multiple data sets with correlated predictors, homogeneous and heterogeneous effects of predictors and generate the response variable.

The usage of the package will be illustrated in the following sections.

2 HomUHet

HomUHet can be used to identify and separates predictors with homogeneous or heterogeneous effect across multiple data sets through fitting penalized linear regression models in two steps. Applicable to Gaussian response variable and very high dimensional data where the number of predictors could be larger than the number of observations per data set.

2.1 Input data for HomUHet

HomUHet<-function(x,y,sid, solution_path_plot=FALSE)

Arguments

x is the the predictor matrix. a matrix of $N \times J$ containing observations from all studies for all predictors, where N is the number of total observations of all data sets and J is the number of predictor. Each column contains the observation of that predictor in all of the studies. Before combining predictors from all data sets and supply them as **x**, the predictors need to be standardized such that the sample mean is 0 and sample variance is 1 for each predictor within each study. In order for this standardization to apply, the distributions of each predictor in all studies are assumed to be the same across studies (but the distributions are allowed to be different across predictors). In the case when the predictors include both covariates and SNPs, it is advised that the covariates are to be standardized as described but the SNPs should be left out of the standardization. Below shows an example of the input **x** matrix.

```
library(HomUHet)
#> Loading required package: glmnet
#> Loading required package: Matrix
#> Loaded glmnet 4.1
#> Loading required package: gglasso
#> Loading required package: dplyr
#>
#> Attaching package: 'dplyr'
#> The following objects are masked from 'package:stats':
#>
#>     filter, lag
#> The following objects are masked from 'package:base':
#>
#>     intersect, setdiff, setequal, union
mydata=HomUHet.sim(Pred_type="SNP",J=500,K=4,level="h",
                   rho=0.5,sigma=2,nlower=50,nupper=300)
# simulating data of 4 studies with varying sample sizes, 500 predictors for each study.
x=mydata$x # the standardized predictor matrix
y=mydata$y # the response variable
sid=mydata$sid # the study labels
x1=x[sid==1,] # extracting observations from study 1
(colMeans(x1))[1:5] # checking if the means are 0. showing first 5 here.
#> [1]  2.186803e-17  2.901719e-17 -1.042937e-16 -6.823246e-17  3.742797e-17
(apply(x1,2,sd))[1:5] # checking if the variances are 1. showing first 5 here.
#> [1] 1 1 1 1 1
```

y is the response variable following gaussian distribution. a vector of N observations for the response variable should be supplied here.

sid the study labels for each observation in x and y. The sid argument needs to be an N length vector containing integers that are the same for observations in the same study and different for different studies.

Therefore each entry will index the study from which the corresponding observation comes. For example, if the data contains 4 studies, then we would expect 4 unique integers in **sid**:

```
unique(sid)
#> [1] 1 2 3 4
length(unique(sid)) # 4 in this case
#> [1] 4
```

solution_path_plot TRUE if outputting solution path plots is desired

2.2 Output values from HomUHet

Values

Homo is a vector containing the names or indexes of identified homogeneous predictors.

Heter is a vector containing the names or indexes of identified heterogeneous predictors.

coefficients a matrix containing the estimated coefficients for the identified homogeneous or heterogeneous predictors.

```
fit=HomUHet(x=x,y=y,sid=sid)
fit$Homo # the identified homogeneous predictors
#>  [1]   1   2   3   4   5 106 107 108 109 110
fit$Heter # the identified heterogeneous predictors
#>  [1] 211 213 215 217 219 221 223 225 227 229 231 233 235 237 239 241 243 245 247
#> [20] 249 251 253 255 257 259 260 261 263 265 267 269
fit$coefficients[1:2,] # the estimated coefficients
#>      [,1]        [,2]               [,3]               [,4]
#> [1,] "predictor" "study1"           "study2"           "study3"
#> [2,] "1"         "7.09084942052278" "7.09084942052278" "7.09084942052278"
#>      [,5]               [,6]
#> [1,] "study4"           "type"
#> [2,] "7.09084942052278" "Homogeneous"
fit$coefficients[15:17,] # the estimated coefficients
#>      [,1]  [,2]                [,3]                [,4]
#> [1,] "217" "-8.47428487731835" "7.05957345052421"  "6.13738742416928"
#> [2,] "219" "-6.32607893241284" "5.33678595560104"  "6.44713753666949"
#> [3,] "221" "6.98663813965076"  "-6.51358613357776" "6.57648505437682"
#>      [,5]                [,6]
#> [1,] "-9.08671963195884" "Heterogeneous"
#> [2,] "-4.75774351731787" "Heterogeneous"
#> [3,] "-5.57937661275832" "Heterogeneous"
```

3 Simulating multiple data sets

HomUHet.sim<-function(Pred_type=c("Con","SNP"),J=1400,K=c(4,10),level=c("l","m","h"), rho=0.5,sigma=2,nlower=50,n

Arguments

J is the number of predictors. J should be at least 300 for this function to work.

K is the number of studies. choose between 4 and 10.

level represents the level of heterogeneity in the effects of predictors. "l" stands for low, "m" stands for medium, and "h" stands for high.

rho should be a number between 0 and 1. This controls the degree of correlation between predictors

sigma should be a positive number. This controls the added noise to the simulated response variable

nlower sets the lower bound of the K sample sizes. The sample size for each study is picked from a uniform distribution with lower bound equal to nlower and upper bound equal to nupper

nupper sets the upper bound of the K sample sizes

Values

x is a matrix of predictors

y is a vector of response variable which is from gaussian distribution.

sid is a vector of study labels taking on integers.