HomUHet Vignette

Type Package

Title Identifying and Separating Homogeneous and Heterogeneous Predictors

Version: 0.0.0.9000

Date: 2020-01-22

Depends tidyverse, glmnet, 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],Shili Lin [aut], Lo-Bin Chang [aut]

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. HomUHet defines a biomarker as having a homogeneous effect if its effects are an non-zero constant across studies, and heterogeneous if the effects are not all equal. However, identifying the biological heterogeneity and homogeneity in the effects of predictors across data sets 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. Since HomUHet is using a two step procedure, it requires special standardization where each predictor needs to be standardized within each study prior to applying HomUHet and future work will be done for cases (e.g. the study-specific sample variances of a predictor are not assumed to be constant across studies) where this standardization may not apply.

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.
- HomUHet.sim.beta simulate the coefficient matrix for all predictors that can be used in HomUHet.sim.

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 (data, solution_path_plot = FALSE)

Arguments

data is the data frame containing containing observations concatenated from all studies where the first column is the study label containing integers that indicate the study to which the observation belongs, the second column is the response variable and the following columns are the predictors including both genetic and non-genetic variables. All studies should use the same set of predictors. The predictors are also expected to be standardized such that the sample mean is 0 and sample variance is 1 for each predictor within each study. The response variable is limited to be a continuous variable. Below shows an example of the input data frame.

```
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
data(HomUHet_data)
HomUHet_data[1:5,1:4]
#>     study_label         y     Pred_V1    Pred_V2
#> X             1  31.70315 -0.64638049  0.9449718
#> X.1           1  -3.51426  0.62634070 -1.2253529
#> X.2           1 -11.52535 -0.63947104 -1.0142125
#> X.3           1 -36.64011 -0.03219034  0.2569333
#> X.4           1  32.92215  0.21230205  0.1268277
# input data of 4 studies, same set of 500 predictors for each study.
#The first column is the study label, the second column is the response variable y,
#and Pred_V1 and Pred_V2 are the first 2 predictors.

unique(HomUHet_data$study_label)
#> [1] 1 2 3 4
# study_label contains 4 different integers labeling the studies.
```

```
x1=HomUHet_data[HomUHet_data$study_label==1,]
# extracting observations from study 1
(colMeans(x1[,-(1:2)]))[1:5] # checking if the means are 0. showing the first 5 predictors here.
#>       Pred_V1       Pred_V2       Pred_V3       Pred_V4       Pred_V5
#> -3.491401e-17 -7.910627e-18 -1.741921e-17  1.808863e-17  1.569169e-18
(apply(x1[,-(1:2)],2,sd))[1:5] # checking if the variances are 1. showing first 5 here.
#> Pred_V1 Pred_V2 Pred_V3 Pred_V4 Pred_V5
#>       1       1       1       1       1
```

In this sample data, the first column contains the study labels. For example, study_label=1 indicates that row 1 is an observation from study 1. The second column is the response variable, y. Starting from the third column are the predictors, which are standardized.

solution_path_plot TRUE if outputting solution path plots is desired.

2.2 Output values from HomUHet

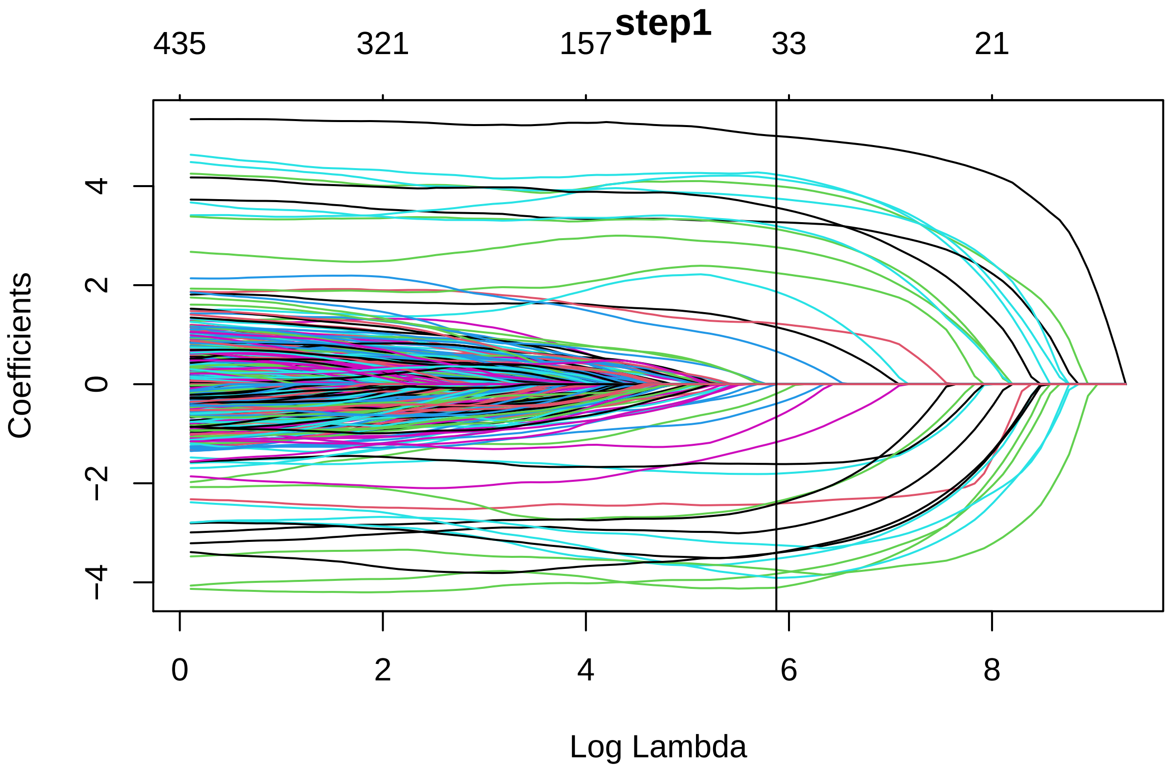
Values the outputs are two solution path plots (if checked TRUE in solution_path_plot) and a list containing the following elements

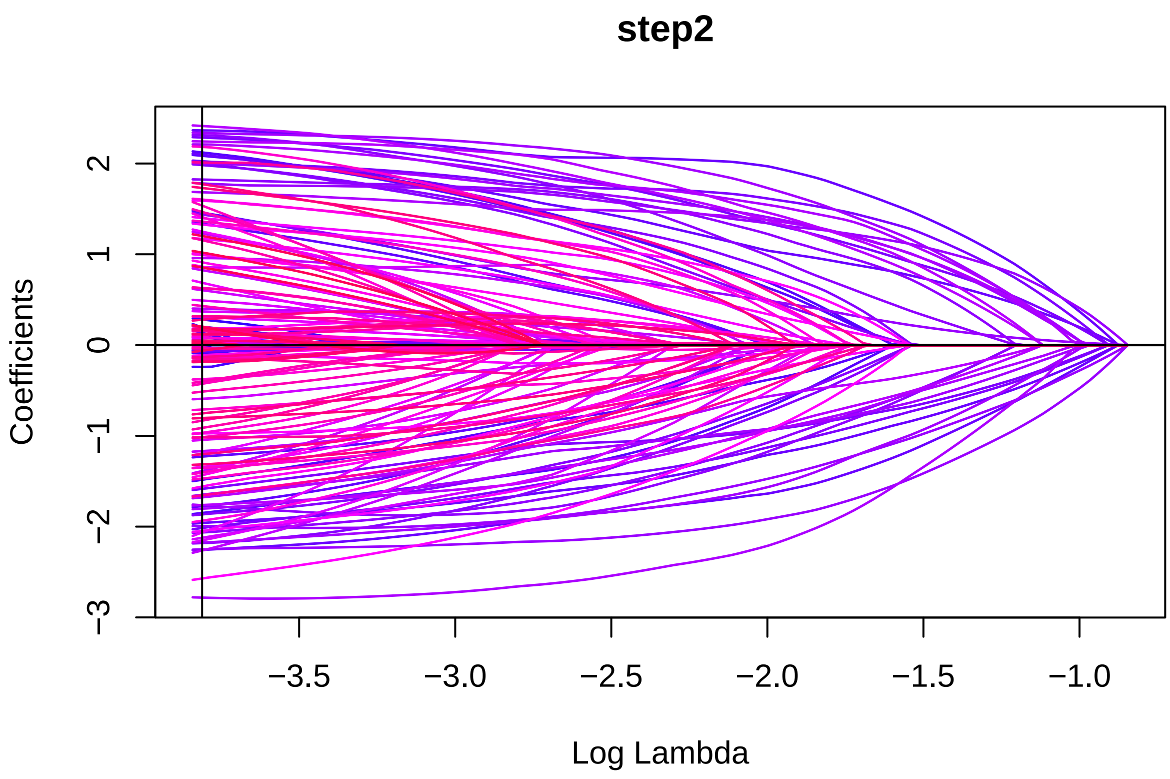
Homo is a vector containing the names (if provided in the input data) or column numbers of identified homogeneous predictors.

Heter is a vector containing the names (if provided in the input data) or column numbers of identified heterogeneous predictors.

coefficients a matrix containing the estimated coefficients for the identified homogeneous or heterogeneous predictors. The estimates of a homogeneous predictor are identical across studies

```
fit=HomUHet(data=HomUHet_data, solution_path_plot = TRUE)
```





```
fit$Homo # the identified homogeneous predictors (names are not provided so these are column names)
#> [1]   1   2   3   5 106 107 108 109 110
fit$Heter # the identified heterogeneous predictors
#>  [1]   4  57  58  78 211 213 215 216 217 219 221 223 225 227 229 231 233 235 237
#> [20] 239 241 243 245 247 249 251 253 255 257 259 261 263 264 265 266 267 269 306
#> [39] 353 382 491
(fit$coefficients[(fit$coefficients)$type=="Homogeneous",])[1:5,]
#>   predictor             study1             study2             study3
#> 1         1  1.63677716044454  1.63677716044454  1.63677716044454
#> 2         2  1.42743123550982  1.42743123550982  1.42743123550982
#> 3         3  2.49994447678595  2.49994447678595  2.49994447678595
#> 4         5  2.58827270793464  2.58827270793464  2.58827270793464
#> 5       106 -1.91007941075491 -1.91007941075491 -1.91007941075491
#>               study4        type
#> 1  1.63677716044454 Homogeneous
#> 2  1.42743123550982 Homogeneous
#> 3  2.49994447678595 Homogeneous
#> 4  2.58827270793464 Homogeneous
#> 5 -1.91007941075491 Homogeneous
# some of the estimated homogeneous coefficients
(fit$coefficients[(fit$coefficients)$type=="Heterogeneous",])[1:5,]
#>    predictor              study1             study2              study3
#> 10         4 0.0810195474459451  0.201365320597479 -0.0322990343255563
#> 11        57    1.3917470880287   1.38268492905552    1.39615535466056
#> 12        58 -0.088923383035408 -0.240790041685688 -0.0729758736933753
#> 13        78 -0.921922485115252 -0.955671441048591   -1.02076654230117
#> 14       211  -1.48363703487686   2.08477039300735   -1.84980429856419
#>                study4          type
#> 10  0.281616925096583 Heterogeneous
#> 11   1.37538189554755 Heterogeneous
#> 12 -0.182055722850135 Heterogeneous
#> 13 -0.863340177925524 Heterogeneous
#> 14   2.12160801369404 Heterogeneous
# some of the estimated coefficients
```

The vertical black lines in the two solution path plots mark the values of lambda used by HomUHet in two steps to generate the outputted values.

3.1 Simulating multiple data sets

HomUHet.sim <- function(Pred_type=c(“Gaussian”,“SNP”), J, K, beta=NULL, rho=0.5, sigma, level=c(“l”, “m”, “e”), nlower, nupper, allele_freq)

Arguments

Pred_type indicates which type of predictor the user wishes to simulate.

J is the number of predictors.

K is the number of studies.

beta the coefficient matrix of dimension $K \times J$ containing coefficients for the homogeneous, heterogeneous and unassociated predictors.

level represents the level of heterogeneity in the effects of predictors. “l” stands for low, “m” stands for medium, and “h” stands for high. used in a default example and ignored if beta is supplied.

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

allele_freq a J-length vector containing the allele frequencies of the J predictors. required only if Pred_type is ”SNP”.

the following shows an example where 3 studies and 12 SNP predictors are to be generated. Among the 12 predictors, the first predictor has homogeneous effects (all equal to 2) and the 11th predictor has heterogeneous effects (2, 4, 6, respectively in the 3 studies).

```
beta=cbind(c(2,2,2), matrix(0,ncol=10,nrow=3),c(2,4,6))
# constructing a beta matrix
K=nrow(beta) # the number of studies
J=ncol(beta) # the number of predictors
allele_freq=runif(J,0.05,0.5) # allele frequency
mydata=HomUHet.sim(Pred_type="SNP", J=J, K=K, beta=beta,
                      rho=0.5,sigma=2,
                      nlower=50,nupper=200, allele_freq=allele_freq)

mydata[1:2,1:5] # showing the first few lines and columns of the output data.
#>     study_label        y  Pred_V1    Pred_V2    Pred_V3
#> X             1 22.83283 2.236505 -0.5842359 -0.8493382
#> X.1           1 24.51172 2.236505  1.5392370  0.8227964
```

value The output is a data frame containing, in that order, the simulated study label, response variable and predictors.