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. 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.

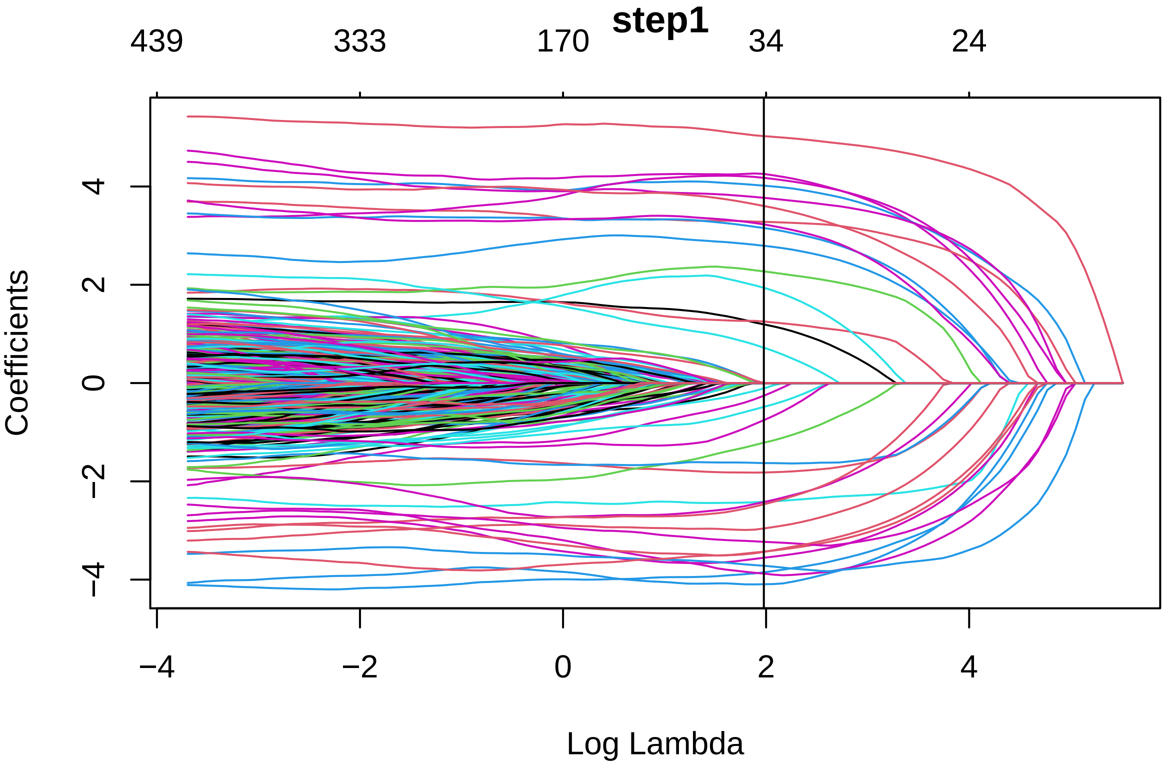
```
library(HomUHet)
#> Loading required package: glmnet
#> Loading required package: Matrix
#> Loaded glmnet 4.1
#> Loading required package: gglasso
#> Loading required package: dplyr
#>
#> Attaching package: 'dplyr'
#> The following objects are masked from 'package:stats':
#>
#>     filter, lag
#> The following objects are masked from 'package:base':
#>
#>     intersect, setdiff, setequal, union
data(HomUHet_data)
HomUHet_data[1:5,1:4]# input data of 4 studies, same set of 500 predictors for each study.
#>     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

dim(HomUHet_data)
#> [1] 821 502

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

x1=HomUHet_data[HomUHet_data$study_label==1,] # extracting observations from study 1
```

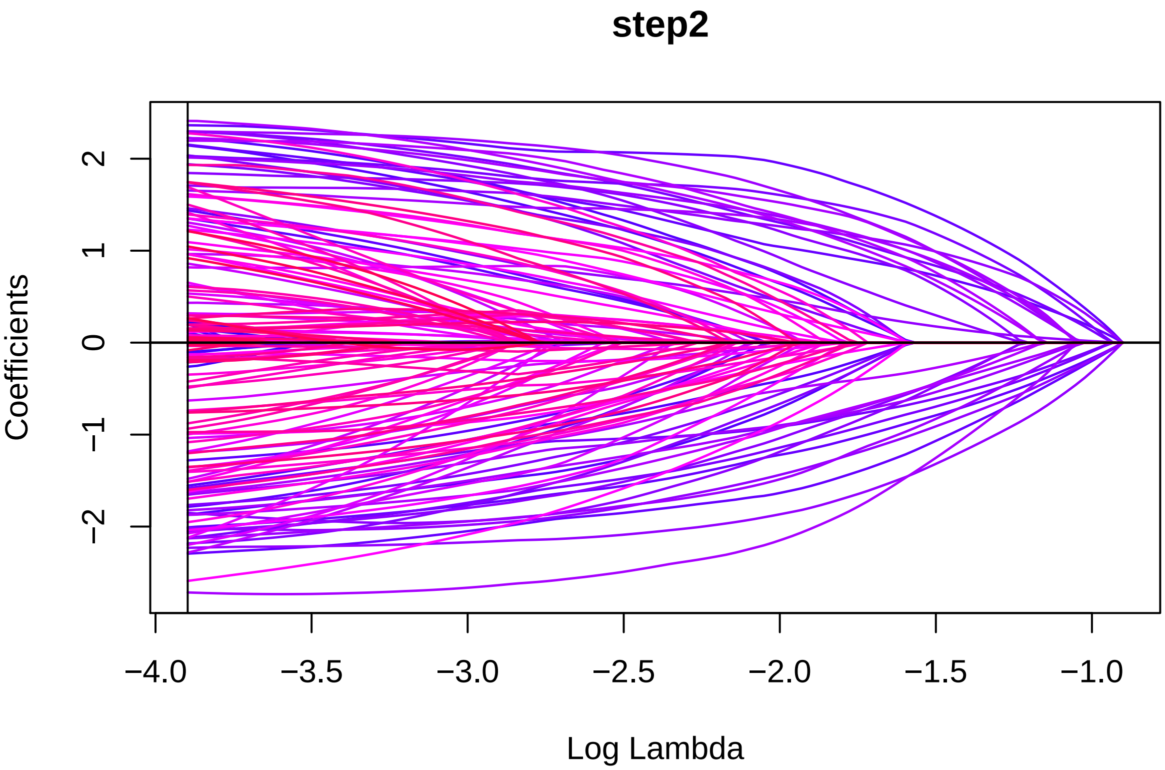
```
(colMeans(x1[,-(1:2)]))[1:5] # checking if the means are 0. showing the first 5 predictors here.
#>        Pred_V1        Pred_V2        Pred_V3        Pred_V4        Pred_V5
#> -3.491401e-17 -7.910627e-18 -1.741921e-17  1.808863e-17  1.569169e-18
(apply(x1[,-(1:2)],2,sd))[1:5] # checking if the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

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





```
#> $Homo
#>  [1]   1   2   3   5 106 107 108 109 110 260
#> 
#> $Heter
#>  [1]   4  57  58  78 211 213 215 216 217 219 221 223 225 227 229 230 231 233 235
#> [20] 237 239 241 243 245 247 249 251 253 255 257 259 261 263 264 265 266 267 269
#> [39] 306 353 382 427 491
#> 
#> $coefficients
#>    predictor               study1              study2               study3
#> 1          1     1.61436825240502    1.61436825240502     1.61436825240502
#> 2          2     1.43644561623458    1.43644561623458     1.43644561623458
#> 3          3     2.52935674610714    2.52935674610714     2.52935674610714
#> 4          5     2.55688994795671    2.55688994795671     2.55688994795671
#> 5        106    -1.87011966168368   -1.87011966168368    -1.87011966168368
#> 6        107    -1.70849104174109   -1.70849104174109    -1.70849104174109
#> 7        108    -1.59893824514899   -1.59893824514899    -1.59893824514899
#> 8        109    -2.52433438876173   -2.52433438876173    -2.52433438876173
#> 9        110     -2.8655396712499    -2.8655396712499     -2.8655396712499
#> 10       260   -0.743509787757318  -0.743509787757318   -0.743509787757318
#> 11         4   0.0735420522499069   0.152397252788904 -0.00120288566506035
#> 12        57     1.39231767300901    1.38851680991485     1.39516325465397
#> 13        58   -0.110214216260805  -0.261261644957651  -0.0975290515916247
#> 14        78   -0.921599000133366   -0.96560204508462    -1.03245979584653
#> 15       211    -1.55523665130732    2.22704180319934    -1.86999798587898
#> 16       213      1.3502603868351    1.43814282690357    -1.28047401391989
#> 17       215     2.36407868183156   -2.00846844074168    -2.29456796591514
#> 18       216  0.00686842197349125 -0.0166506290976548  0.00118601997226445
#> 19       217     2.03120195525943   -2.12585916786851    -1.58079271796314
#> 20       219     1.94005089865703    2.03660653717147    -1.76584167329904
#> 21       221     1.84462179775571    -2.0562669132008    -1.85081711602747
#> 22       223     1.70937990539463   -2.01895830638719     1.46104777900689
#> 23       225     2.01302064334895    2.29870618968682    -1.65399136116226
#> 24       227     -1.8223458201328    1.65436502321912    -2.71625497261507
#> 25       229    -1.63340207000478    2.41075451934731    -1.87307530388672
#> 26       230 0.000909185067000518  0.0102917709397361 -0.00452250721624513
#> 27       231     5.50563146487248    4.97511550328775     2.26006747454398
#> 28       233     4.19906401801511    1.85864180638525     3.39231797561328
#> 29       235     3.55644636427836    4.80798454727043     2.11219858921067
#> 30       237     4.94717287811157    2.21318014012036     3.37509784654581
#> 31       239     3.98760372959715    5.30992293014047     3.04925218841902
#> 32       241     3.14913292849923    4.29665263054271     3.67759948481122
#> 33       243     5.31603240061606    5.19694092675546     3.84115211035442
#> 34       245     5.58300331784522    1.40319182807942     3.00387874482157
#> 35       247     3.57184385900681    4.42023749203901      4.5503880682555
#> 36       249      4.6055838326517    2.07527300648694     3.45239501839928
#> 37       251    -1.35625031939979   -4.66849438560417    -3.00586930430188
#> 38       253    -5.11858065922439   -4.03684760019264    -2.70667652245461
#> 39       255    -5.19157788246549   -3.70727240759961     -4.5245546853628
#> 40       257    -1.85742792572966   -4.64647963162203    -3.45532789800845
#> 41       259    -4.62248208406158   -3.62906983738591     -2.4385608487321
#> 42       261    -2.16590814223526   -3.09840409878443    -4.09041380959113
#> 43       263    -4.58466636438651   -4.52649398195941    -2.43093371271621
#> 44       264   -0.165427180136248 -0.0357282413017941    0.196729147074208
#> 45       265    -5.14389778735169   -3.24707187022838    -4.51552095901155
#> 46       266   -0.840022576325197  -0.780671839038432     -1.1903688787438
#> 47       267    -1.39401533333234   -2.87564698690043    -4.33912302927868
#> 48       269    -2.02840794227785   -3.62397267413261    -4.53540370457212
#> 49       306   -0.211825766443553  0.0542849580778234    0.133245933855651
#> 50       353  -0.0836060000572272  0.0121870607642877   0.0314305562376288
#> 51       382  0.00182691681419556 0.00556843974592781  0.00232088947604207
#> 52       427   0.0290436115199617  0.0123155440326989  -0.0250741967166921
#> 53       491   -0.418402777167701  -0.584316169103777    -1.36121953166438
#>                  study4          type
#> 1      1.61436825240502   Homogeneous
#> 2      1.43644561623458   Homogeneous
#> 3      2.52935674610714   Homogeneous
#> 4      2.55688994795671   Homogeneous
#> 5     -1.87011966168368   Homogeneous
#> 6     -1.70849104174109   Homogeneous
#> 7     -1.59893824514899   Homogeneous
#> 8     -2.52433438876173   Homogeneous
#> 9      -2.8655396712499   Homogeneous
#> 10   -0.743509787757318   Homogeneous
#> 11    0.221938633976112 Heterogeneous
#> 12     1.38531711753629 Heterogeneous
#> 13   -0.215070493602625 Heterogeneous
#> 14   -0.872411384107029 Heterogeneous
#> 15     2.14464260824252 Heterogeneous
#> 16     -1.7856493040989 Heterogeneous
#> 17     2.15301991909347 Heterogeneous
#> 18  0.00955412972692142 Heterogeneous
#> 19     2.29632132651933 Heterogeneous
#> 20   -2.18288102492565 Heterogeneous
#> 21    2.29777286753672 Heterogeneous
#> 22   -2.23134684945473 Heterogeneous
#> 23   -2.12868883248477 Heterogeneous
#> 24    2.22400056216871 Heterogeneous
#> 25    2.19496103310858 Heterogeneous
#> 26 0.00753882434390508 Heterogeneous
#> 27    3.34795856073113 Heterogeneous
#> 28    4.64581420863831 Heterogeneous
#> 29    5.00493593603226 Heterogeneous
#> 30    4.73177961273494 Heterogeneous
#> 31    5.48757329510465 Heterogeneous
#> 32    1.66645654136771 Heterogeneous
#> 33    1.92769632083793 Heterogeneous
#> 34    5.34561627457596 Heterogeneous
#> 35    1.83558702646676 Heterogeneous
#> 36    4.55073885547635 Heterogeneous
#> 37   -4.92328461383828 Heterogeneous
#> 38   -4.46937417924448 Heterogeneous
#> 39   -1.50920919896081 Heterogeneous
#> 40   -4.76371308071718 Heterogeneous
#> 41   -5.21388392788347 Heterogeneous
#> 42   -4.14576937333905 Heterogeneous
#> 43   -3.04162007623321 Heterogeneous
#> 44   0.156290807013418 Heterogeneous
#> 45   -1.60985036379544 Heterogeneous
#> 46  -0.790607044669801 Heterogeneous
#> 47   -3.89347999901089 Heterogeneous
#> 48   -5.12635318983443 Heterogeneous
#> 49  -0.187586707529106 Heterogeneous
#> 50  0.0107996040219368 Heterogeneous
#> 51  0.0147801615385905 Heterogeneous
#> 52   0.056943276721692 Heterogeneous
#> 53  -0.711136582473626 Heterogeneous
```

2.2 Output values from HomUHet

Values the output is 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)
fit$Homo # the identified homogeneous predictors
#>  [1]   1   2   3   5 106 107 108 109 110 260
fit$Heter # the identified heterogeneous predictors
#>  [1]   4  57  58  78 211 213 215 216 217 219 221 223 225 227 229 230 231 233 235
#> [20] 237 239 241 243 245 247 249 251 253 255 257 259 261 263 264 265 266 267 269
#> [39] 306 353 382 427 491
```

```
fit$coefficients[1:2,] # the estimated coefficients
#>   predictor           study1           study2           study3           study4
#> 1         1 1.61436825240502 1.61436825240502 1.61436825240502 1.61436825240502
#> 2         2 1.43644561623458 1.43644561623458 1.43644561623458 1.43644561623458
#>          type
#> 1 Homogeneous
#> 2 Homogeneous
fit$coefficients[15:17,] # the estimated coefficients
#>    predictor            study1            study2            study3
#> 15       211 -1.55523665130732  2.22704180319934 -1.86999798587898
#> 16       213   1.3502603868351  1.43814282690357 -1.28047401391989
#> 17       215  2.36407868183156 -2.00846844074168 -2.29456796591514
#>              study4          type
#> 15 2.14464260824252 Heterogeneous
#> 16 -1.7856493040989 Heterogeneous
#> 17 2.15301991909347 Heterogeneous
```

3.1 Simulating multiple data sets

HomUHet.sim<-function(Pred_type=c("Gaussian","SNP"), J, K, beta=NULL, rho=0.5,sigma=2, level=c("l","m","e"), nlower=50,nupper=300, allele_freq)

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.

beta

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

Values

x is a matrix of predictors

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

study_label is a vector of study labels taking on integers.

3.2 simulating coefficients

HomUHet.sim.beta<-function(J, K, homo_coef, heter_distr=c("Gaussian","Uniform"), heter_coef_param)