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 in two steps. In the first step, HomUHet estimates and removes the overall effects of predictors (only homogeneous and some of the heterogeneous predictors are expected to have overall effect) by using data concatenated from all studies. After step 1, homogeneous predictors will not be associated with the response variable any more, but the heterogeneous ones are still associated due to deviance between study specific effect and overall effect. In step 2, by using a group LASSO method, HomUHet estimates the remaining effects of predictors, keeping predictors with remaining heterogeneous effects in the model while dropping predictors without remaining effects (both homogeneous and unassociated predictors are expected to be dropped in step 2, for they should not have remaining effects). By combining the estimates from the two steps, HomUHet estimate effects of predictors in all studies and separate the homogeneous, heterogeneous and unassociated predictors apart. 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.beta simulate the coefficient matrix for predictors that can be used in HomUHet.sim.
- 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 (data, solution_path_plot = FALSE)

data is the data frame where each row is information from a subject in one of the studies, 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)
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]
```

```
##       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
```

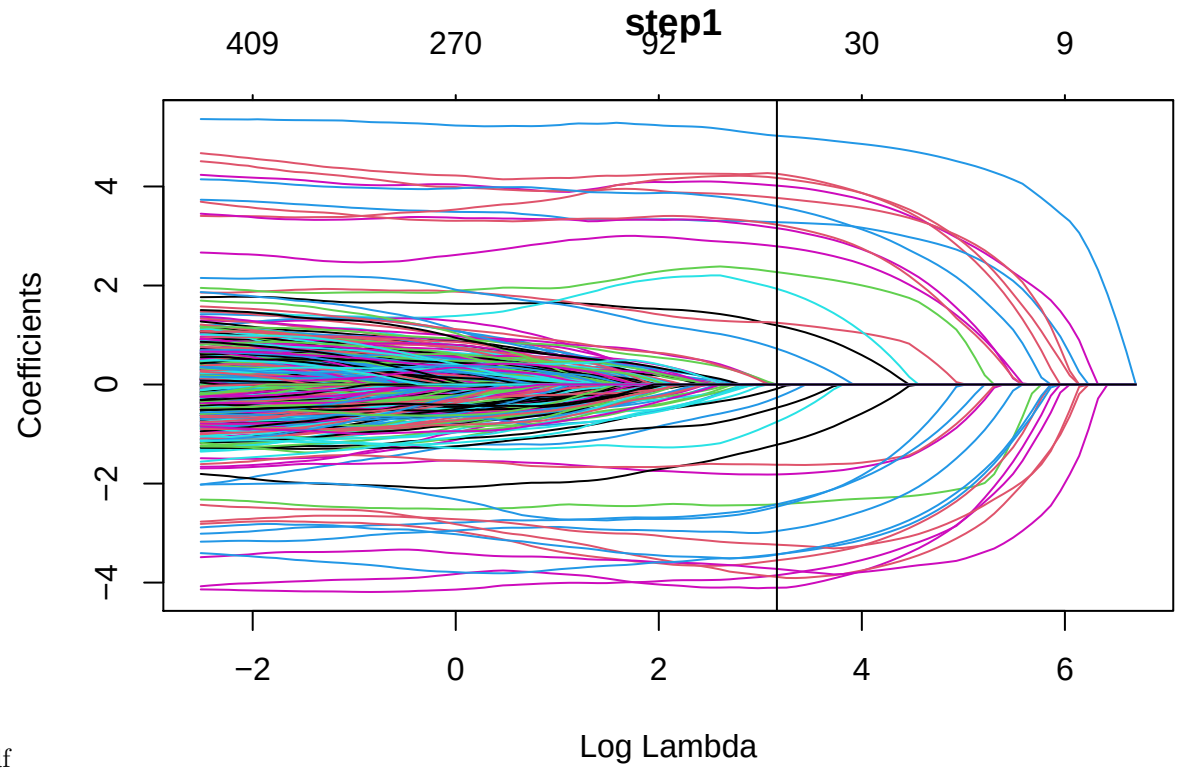
```
# checking if the means are 0 for each predictor in study 1. showing the first 5 predictors here.
(apply(x1[,-(1:2)],2,sd))[1:5] # checking if the variances are 1. showing first 5 here.
```

```
## Pred_V1 Pred_V2 Pred_V3 Pred_V4 Pred_V5
##       1       1       1       1       1
```

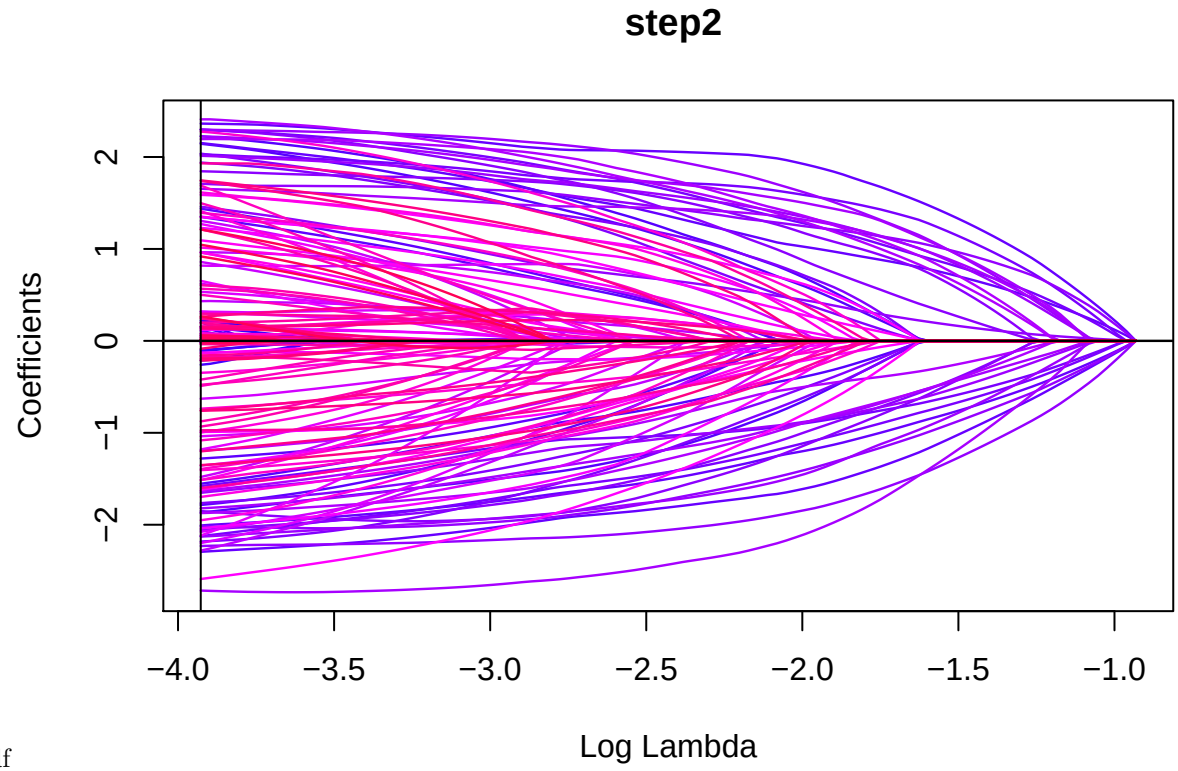
In this sample data, the first column contains the study labels. For example, study_label=1 indicates that row 1 is an observation from study 1. The second column is the response variable, y. Starting from the third column are the predictors, which are standardized.

solution_path_plot TRUE if outputting solution path plots for both step 1 and step 2 is desired. Figures below shows examples of the solution path plots.

```
set.seed(100)
fit=HomUHet(data=HomUHet_data, solution_path_plot = TRUE)
```



plots-1.pdf



plots-2.pdf

Figure 1: solution path plots

2.2 Output values from HomUHet

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) of identified homogeneous predictors. If names are not provided, **Homo** will display the column numbers of them, shown as "V_n", where n is the column number (excluding the study label and response variable). The estimates are constant across studies.

Heter is a vector containing the names (if provided in the input data) of identified heterogenous predictors. If names are not provided, **Heter** will display the column numbers of them, shown as "V_n", where n is the column number (excluding the study label and response variable).

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

In the following, we are going to show an example by applying HomUHet() to HomUHet_data. HomUHet_data contains 500 predictors that are correlated, out of which 10 have homogeneous effects and 30 have heterogeneous effects. In the result below, HomUHet correctly identified 9 out of 10 homogeneous effects while including only one falsely discovered unassociated predictor. HomUHet also correctly identified all 30 heterogeneous predictors while falsely identified 13 out of 460 unassociated predictors as heterogeneous.

```
set.seed(100)
fit=HomUHet(data=HomUHet_data, solution_path_plot = FALSE)
fit$Homo
```

```
##  [1] "Pred_V1"   "Pred_V2"   "Pred_V3"   "Pred_V5"   "Pred_V106" "Pred_V107"
##  [7] "Pred_V108" "Pred_V109" "Pred_V110" "Pred_V260"
```

```
# the identified homogeneous predictors. if names are not provided, the predictor names will be display
(fit$Heter)[1:10]
```

```
##  [1] "Pred_V4"   "Pred_V57"  "Pred_V58"  "Pred_V78"  "Pred_V211" "Pred_V213"
##  [7] "Pred_V215" "Pred_V216" "Pred_V217" "Pred_V219"
```

```
# partial list of the identified heterogeneous predictors
(fit$coefficients[(fit$coefficients)$type=="Homogeneous",])[1:5,]
```

```
##   predictor            study1            study2            study3
## 1   Pred_V1  1.61436825240502  1.61436825240502  1.61436825240502
## 2   Pred_V2  1.43644561623458  1.43644561623458  1.43644561623458
## 3   Pred_V3  2.52935674610714  2.52935674610714  2.52935674610714
## 4   Pred_V5  2.55688994795671  2.55688994795671  2.55688994795671
## 5 Pred_V106 -1.87011966168368 -1.87011966168368 -1.87011966168368
##              study4        type
## 1  1.61436825240502 Homogeneous
## 2  1.43644561623458 Homogeneous
## 3  2.52935674610714 Homogeneous
## 4  2.55688994795671 Homogeneous
## 5 -1.87011966168368 Homogeneous
```

```
# partial display of the estimated homogeneous coefficients
(fit$coefficients[(fit$coefficients)$type=="Heterogeneous",])[1:5,]
```

```
##    predictor             study1             study2               study3
## 11   Pred_V4  0.073434350756619  0.152313955595617 -0.00123870161452215
## 12  Pred_V57   1.39388977839516   1.39062728445416     1.39632806523841
## 13  Pred_V58 -0.109787320305859 -0.260183649683502  -0.0969039904448178
## 14  Pred_V78 -0.921477382441846 -0.965513446770938    -1.03249584755419
## 15 Pred_V211  -1.55424906346995   2.22669879901544    -1.86937860414054
##                study4          type
## 11  0.221842229034671 Heterogeneous
## 12   1.38789187502868 Heterogeneous
## 13 -0.214184858406882 Heterogeneous
## 14 -0.872379242669427 Heterogeneous
## 15   2.14422501267974 Heterogeneous
```

```
# partial display of the estimated coefficients
```

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

3 Simulating data

We provide two functions for simulating data sets. One of them simulates the coefficients of predictors and the other simulates the data sets when the coefficients of predictors are supplied. By having both functions, the users have the option of using their own set of coefficients as well as having a convenient way to simulate the coefficients from scratch.

3.1 Simulating the data sets

HomUHet.sim simulates continuous response variable generated as a linear combination of predictors including common covariates such as age and sex, as well as continuous genetic variants and SNPs. The users can

choose to supply their own coefficient matrix used to generate the response variable, otherwise the coefficients will be generated by calling HomUHet.sim.beta. In this section, we demonstrate the usage of HomUHet.sim without directly supplying the coefficient matrix.

HomUHet.sim (age=TRUE, sex=TRUE, K, n_cont, n_SNP, rho=0.5,sigma=2, nlower=50, nupper=300, beta=NULL)

age TRUE if covariate age should be included. The ages will be simulated from a normal distribution with mean 40 and standard deviation 5. The coefficient value for age is set to be 1.

sex TRUE if covariate sex should be included. sex will be simulated from Bernoulli distribution with the proportion randomly chosen between 0 and 1. The coefficient value for sex is set to be 1.

K is the number of studies.

n_cont is the number of continuous predictors to be simulated. Enter 0 if there will not be continuous predictors.

n_SNP is the number of SNPs to be simulated. Enter 0 if there will not be SNP predictors. The allele frequencies will be simulated from Uniform (0.05, 0.5).

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

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

HomUHet.sim gives the option of using unequal sample sizes across studies, so the sample size of each study will be drawn from Uniform (nlower, nupper), where

nlower sets the lower bound of the sample sizes.

nupper sets the upper bound of the sample sizes.

beta if the users wish to supply the coefficients on their own, enter a coefficient matrix where the columns containing the K coefficients of each predictor, for all genetic and non-genetic predictors. For example, to simulate a data set with age, sex, and 6 genetic variants for K studies, beta should be a $K \times 8$ matrix.

The following shows an example where 4 studies and 20 predictors including age, sex, 5 continuous predictors and 13 SNPs are to be generated. Among the genetic variants, 3 of them have homogeneous coefficients and 5 have heterogeneous coefficients. The sample sizes of k studies range between 50 and 120

```
mydata=HomUHet.sim (age=TRUE, sex=TRUE, K=4, n_cont=5, n_SNP=13, n_homo=3, n_heter=5, rho=0.5,sigma=2,
                    nlower=50,nupper=120)

(mydata$data)[1:2,1:5] # a partial display of the simulated data
```

```
##     study_label         y    Pred_V1    Pred_V2    Pred_V3
## X             1  1.838625 -0.3567578 -0.3610555 -0.4401874
## X.1           1 -8.817805  0.4080296  0.7864974  1.2570905
```

The output is a list of the following items:

data is a data frame containing, in that order, the simulated study label, response variable and predictors. **study_label** is the variable indicating to which study each row of observation belongs, **y** is the continuous response variable and the rest of the variables are the predictors arranged in the order of age, sex, continuous predictors, and SNPs.

beta is the simulated (or supplied) coefficient matrix.

homo_index is the column numbers of simulated homogeneous coefficients (or supplied coefficients) in **data**.

heter_index is the column numbers of simulated heterogeneous coefficients (or supplied coefficients) in **data**.

3.2 Simulating the coefficients

This function simulates homogeneous and heterogeneous coefficients of prospective predictors, and outputs them in a matrix including the coefficients of predictors without an effect (in which case, the coefficients are zero)

HomUHet.sim.beta (J, K, n_homo, n_heter, level=c("l","m","h"))

The function allows the user to simulate the coefficients of **J** prospective predictors for **K** studies. The users can choose to simulate **n_homo** number of homogeneous coefficients and **n_heter** number of heterogeneous coefficients. The input values for HomUHet.sim.beta() are the following:

J is the total number of predictors including the predictors with homogeneous and heterogeneous effects and the predictors without effects.

K is the number of studies.

n_homo indicates the number of homogeneous coefficients

n_heter indicates the number of heterogeneous coefficients

level "l", "m", and "h" represent increasing level of heterogeneity in the coefficients, where "l" stands for low, "m" stands for medium and "h" stands for high, each corresponding to a different sets of parameters used to simulate the coefficients.

In the following, we show an example of simulating coefficients for J=20 prospective predictors. Out of which, 3 are homogeneous, 4 are heterogeneous and the rest of them are zero, for k=4 studies.

```
# simulating coefficients
coef_mat = HomUHet.sim.beta(J=20, K=4, n_homo=3, n_heter=4, level="l")

# below is a partial display of the matrix of coefficients
# each column contains the 4 coefficients of one predictor
# across 4 studies.

coef_mat$beta[,1:6]
```

```
##          [,1] [,2]      [,3]      [,4] [,5]      [,6]
## [1,] 2.366374    0 -1.889958 -2.080075    0 -1.221267
## [2,] 2.366374    0 -1.889958 -2.080075    0  1.157911
## [3,] 2.366374    0 -1.889958 -2.080075    0 -1.390863
## [4,] 2.366374    0 -1.889958 -2.080075    0  1.238417
```

```
# we can also extract the heterogeneous coefficients by entering the following:

(coef_mat$beta)[,coef_mat$heter_index]
```

```
##           [,1]     [,2]      [,3]      [,4]
## [1,] -1.221267 2.084252 -3.414029 -1.364768
## [2,]  1.157911 3.453783 -3.028280 -3.173882
## [3,] -1.390863 1.224005 -1.291520 -2.172252
## [4,]  1.238417 3.194185 -2.005434 -3.054532
```

```
# where heter_index tells us the column numbers of the heterogeneous coefficients

coef_mat$heter_index
```

```
## [1]  6  8 10 12
```

```
# and homo_index tells us the column numbers of the heterogeneous coefficients

coef_mat$homo_index
```

```
## [1] 1 3 4
```

The function outputs the simulated coefficient matrix and miscellenance information about it, including:

beta, the K x J coefficient matrix containing homogeneous, heterogeneous coefficients, as well as coefficients of prospective unassociated predictors (which are zeros).

J, the number of predictors including both predictors which have effects and which do not.

K, the number of studies.

homo_index, a vector containing the column numbers of homogeneous coefficients in the coefficient matrix.

heter_index, a vector containing the column numbers of homogeneous coefficients in the coefficient matrix.