Applying Causal Direction Methods to LDL and CAD

Haoran Xue

5/18/2020

Overview

We proposed three methods to infer causal direction between two traits, they are **CD-Ratio** for the ideal model, **CD-Egger** and **CD-GLS** for pleiotropic model. **CD-Ratio** could be applied to a single SNP, or a set of possibly correlated SNPs. **CD-Egger** and **CD-GLS** could be applied to a set of possibly correlated SNPs. We will use LDL and CAD as example to show how to apply these three methods with our functions.

R Functions

First load functions, and data for LDL and CAD:

```
source("Compute_V_Matrix_May18.R")
source("CalSingle_Locus_or_SNP_May18_2020.R")
load("LDLandCAD_May18.Rdata")
```

Functions in "Compute_V_Matrix_May18.R" are used to compute the asymptotic covariance matrix of sample correlations; functions in "CalSingle_Locus_or_SNP_May18_2020.R" are used to apply CD methods. After loading "LDLandCAD_May18.Rdata" we will get the data for LDL and CAD named as *pruned*, which is a list consists of two elements. The first element "loci_bed" is genotype data of the reference panel, which is a n(sample size) by p(number of SNPs) matrix, rows corresponding to individuals in the reference panel and columns corresponding to SNPs. The second element "sig_part" is for the summary statistics, which is a data frame has p rows and 12 columns, rows corresponding to SNPs, and columns are: (1)chr: chromosome of the SNP; (2)pos: base pair position of the SNP; (3)rsid: rs name of the SNP; (4)A1: allele 1; (5)A2: allele 2; (6)beta_T1: effect size for trait 1; (7)se_T1: standard error of effect size for trait 1; (8)N_T1: sample size for trait 1; (9)beta_T2: effect size for trait 2; (10)se_T2: standard error of effect size for trait 2; (11)N_T2: sample size for trait 2; (12)loci: which loci is this SNP from. Here we have 489 individuals in the reference panel, and using 22 SNPs.

```
dim(pruned$loci_bed)
```

```
## [1] 489  22
```

```
dim(pruned$sig_part)
```

```
## [1] 22 12
```

```
head(pruned$sig_part)
```

```
##   chr       pos       rsid A1 A2   beta_LDL   se_LDL  N_LDL beta_CAD  se_CAD
## 1   1  55496039 rs11206510  C  T -0.0695200 0.003555 294565 -0.06272 0.01138
## 2   1  55505647 rs11591147  T  G -0.4752703 0.011494 265213 -0.22090 0.03500
## 3   1 109817590 rs12740374  T  G -0.1618858 0.003197 294565 -0.10990 0.01022
## 4   2  44074431  rs4245791  C  T  0.0721910 0.003093 270962  0.05444 0.00887
## 5   2 203880992  rs2351524  T  C -0.0239770 0.004204 295826  0.10227 0.01228
```

```
## 6   6 160557643  rs2282143  T  C  0.0582204 0.008825 247909  0.26186 0.03325
##    N_CAD loci
## 1 336860   34
## 2 268736   34
## 3 268733   67
## 4 268741  160
## 5 268745  253
## 6 243575  734
```

Then we can apply functions `CD_Ratio` to *pruned*, and get results:

```
CD_Ratio(pruned)
```

```
## $T1toT2
##           K       se(K)
## 0.150739236 0.007116689
##
## $T2toT1
##         K     se(K)
## 0.9187663 0.0571499
##
## $Q_T1toT2
##          [,1]
## [1,] 337.0535
##
## $Q_T2toT1
##          [,1]
## [1,] 527.2407
```

From the above result, for LDL to CAD, $\hat{K} = 0.151, se(\hat{K}) = 0.007$; for CAD to LDL, $\hat{K} = 0.919, se(\hat{K}) = 0.057$. And the two goodness-of-fit test statistics are 337.1 and 527.2.

Next we can apply functions `CD_Egger` to *pruned*, and get results:

```
CD_Egger(pruned)
```

```
## $T1toT2
##          b0           K      se(b0)       se(K)
## 0.006017342 0.165614296 0.002022583 0.032130605
##
## $T2toT1
##           b0            K       se(b0)        se(K)
## -0.025326887  3.302742188  0.009490421  0.638876900
##
## $Q_T1toT2
##          [,1]
## [1,] 22.21132
##
## $Q_T2toT1
##          [,1]
## [1,] 22.01223
```

From the above result, for LDL to CAD, $\hat{K} = 0.166, se(\hat{K}) = 0.032$, and $\hat{b}_0 = 0.006, se(\hat{b}_0) = 0.002$; for CAD to LDL, $\hat{K} = 3.303, se(\hat{K}) = 0.639$, and $\hat{b}_0 = -0.253, se(\hat{b}_0) = 0.009$. And the two goodness-of-fit test statistics are 22.2 and 22.0. Similarly we can apply functions `CD_GLS` to *pruned*, and get results:

```
CD_GLS(pruned)
```

```
## $T1toT2
##          b0           K      se(b0)       se(K)
## 0.005932237 0.165189643 0.001980317 0.031520027
##
## $T2toT1
##           b0            K       se(b0)        se(K)
## -0.023271919  3.100093240  0.008453923  0.576205502
##
## $Q_T1toT2
##          [,1]
## [1,] 22.56534
##
## $Q_T2toT1
##          [,1]
## [1,] 24.19261
```

Besides run individual functions, we can use function `CD_3_methods` to get results for all 3 CD methods, and only need to calculated asympototic covariance matrices once to save time.

We can also apply **CD-Ratio** to a signle SNP. Let us use the first SNP rs11206510 as an example.

```
ind = 1
SNP = pruned$loci_bed[,ind]
sig_part_all = pruned$sig_part[ind,]
pruned = list(loci_bed = SNP,sig_part = sig_part_all)
CD_Ratio_SingleSNP(pruned)
```

```
## $T1toT2
##          K      se(K)
## 0.26370687 0.04970712
##
## $T2toT1
##         K     se(K)
## 3.7920893 0.7147855
```