Accurate genetic and environmental covariance estimation with composite likelihood in genome-wide association studies

Boran Gao

2020-06-15

Introduction

This vignette provides an introduction to the `GECKO` package. R package `GECKO` implements GECKO, an accurate genetic and environmental covariance estimation with composite likelihood in genome-wide association studies. The package can be installed with the command:

```
library(devtools)
```

```
install_github("borangao/GECKO")
```

The package can be loaded with the command:

```
library("GECKO")
```

Fit GECKO using example data within the package

We first load summary statistics and LD score.sumstat_1 and sumstat_2 are the files of the summary statistics with column name chr, bp, SNP, A1, A2, N, Z, P representing chromosome, base pair position, SNP iD, major allele, minor allele, number of individuals in the study, Z score, P value. ldscore is the LD score file generated by LDSC software using 1000 genome reference panel. The example data and LD score could be accessed using the code below.

```
data(sumstat_1);
data(sumstat_2);
data(ldscore);
```

There are 6 arguments needed to be specified by GECKO including number of observation of the both studies, number of overlapped samples, whether to use weighted composite-lieklihood (usually is specified TRUE to increase estimation accuracy), whether fix environmental covariance to be zero if it's known that there is no overlapped sample, whether testing for the genetic and envrionmental covariance (It takes more time to test for the significant genetic and envrionmental covariance).GECKO_R is the function to calculate genetic and environmental covariance. The example code of analysis is listed below.Then we fit the GECKO using code below:

```
n1in<-round(mean(sumstat_1$N))
n2in<-round(mean(sumstat_2$N))
nsin<-0
Weightin = T #is always set for GECKO to improve the efficiency
Fix_Vein = T #(if two studies have non-overlapped samples, otherwise false)
Test = T #Test for significance or not
###Need to specify the number of individuals within each study: n1,n2, and number of the overlapping in
###if the two samples are from the separate studies, nsin = 0, and Fix_Vein = 1
###if the two samples are from the same study, nsin need to be specified
```

```
Result<-GECKO_R(sumstat_1,sumstat_2,n1in,n2in,nsin,ldscore,Weightin,Fix_Vein,Test)
#> 1234567891011121314151617181920212223242526272829303132333435363738394041424344454647484950515253545
Result
#>                Env_Var_Comp1 Env_Covar Env_Var_Comp2 Env_Correlation
#> Estimates        0.993080057         0   0.962412504               0
#> Standard_Error   0.002909489         0   0.006899501               0
#> P_value          0.000000000       NaN   0.000000000             NaN
#>                Gen_Var_Comp1    Gen_Covar Gen_Var_Comp2   Gen_Corr
#> Estimates        0.006919943 0.0002090599  3.758750e-02 0.01296277
#> Standard_Error   0.002909489 0.0002909236  6.899501e-03 0.18381437
#> P_value          0.017387722 0.4723828727  5.098020e-08 0.94377901
```

The result consists 8 columns including the environmental variance estiamte of study 1, environmental covariance estimate, environmental variance estiamte of study 2, environmental correlation estimate, genetic variance estimate(heritability) of study 1, genentic covariance estimate, genetic variance estimate of study 2, and genetic correlation estimate. The environmnetal covariance and correlation are zero in non-overlapping individuals.