MAPLE User Manual

Liye Zhang, Lu Liu, Xiang Zhou, Zhongshang Yuan

xzhousph@umich.edu and yuanzhongshang@sdu.edu.cn

Contents

1 Introduction

1.1 What is MAPLE

MAPLE (a novel Mendelian randomization method with self-Adaptive determination of samPle structure and multiple pLEiotropic effects), is an R package for efficient statistical inference of mendelian randomization analysis (https://github.com/yuanzhongshang/MAPLE). MAPLE utilizes a set of correlated SNPs, self-adaptively accounts for the sample structure and the uncertainty that these correlated SNPs may exhibit multiple pleiotropic effects, as well as explicitly models both uncorrelated and correlated horizontal pleiotropy. The term 'self-adaptive' represents MAPLE is able to automatically infer the sample structure and the probability that a SNP has one specific pleiotropy effect from the data at hand. In particular, MAPLE first acquires the accurate estimate of the nuisance error parameter using the genome-wide summary statistics, then places the inference of the causal effect into a likelihood-framework and relies on a scalable sampling-based algorithm to obtain calibrated p-values.

1.2 The MAPLE method

The MAPLE model can be constructed as follows,

$$(\mathbf{z}_1, \mathbf{z}_2) \sim MN\left(\mathbf{\Sigma}\left(\sqrt{n_1-1}\boldsymbol{\beta}, \sqrt{n_2-1}\left(\boldsymbol{\beta}\alpha + (\boldsymbol{\beta}\circ\boldsymbol{v})\omega + \boldsymbol{\eta}_u\right)\right), \mathbf{\Sigma}, \mathbf{\Omega}\right)$$

where $(\mathbf{z}_1, \mathbf{z}_2)$ follows a matrix normal distribution, the p by p row covariance matrix $\mathbf{\Sigma}$ characterizes the covariance among the marginal z-scores across SNPs, and the 2 by 2 column covariance matrix $\mathbf{\Omega}$ characterizes the covariance between the marginal z-scores on the exposure and that on the outcome to account for sample structure. $\mathbf{z}_1$ is a p-vector of marginal z-scores measuring the association between the candidate SNPs and the exposure, with n_1 individuals in the exposure GWAS; $\boldsymbol{\beta}$ is a p-vector of the effects of the correlated SNPs on the exposure; $\mathbf{z}_2$ is a p-vector of marginal z-scores measuring the association between the candidate SNPs and the outcome, with n_2 individuals in the outcome GWAS; α is a scalar that represents the causal effect of the exposure on the outcome; $\boldsymbol{\eta}_u$ is p-vector of uncorrelated horizontal pleiotropic effects on the outcome; $\omega\beta_j$

represents the correlated horizontal pleiotropic effect; $\boldsymbol{v}$ is a p-vector of binary indicators to indicate whether the SNP displays correlated pleiotropy effect ($v_j = 1$) or not ($v_j = 0$), with $p(v_j = 1|\beta_j \neq 0) = \pi_c$, the term $\boldsymbol{\beta} \circ \boldsymbol{v}$ represent the element wise product of the two vectors $\boldsymbol{\beta}$ and $\boldsymbol{v}$.

2 Installation

To install the development version of MAPLE, it's easiest to use the `devtools` package. Appropriate setting of Rtools is required, given that MAPLE relies on the `Rcpp`, `RcppArmadillo`, `RcppDist`, `dplyr`, `magrittr` and `readr` packages.

```
#install.packages("devtools")
library(devtools)
install_github("yuanzhongshang/MAPLE")
```

3 Application analysis

Note: All data used in the manual can be found at https://github.com/yuanzhongshang/MAPLE/tree/main/example.

3.1 Step 1: Estimation of sample structure $\boldsymbol{\Omega}$

The function `est_SS` can estimate the parameter $\boldsymbol{\Omega}$ using to account for sample structure (e.g., population stratification, cryptic relatedness, and sample overlap).

```
library(Rcpp)
library(RcppArmadillo)
library(RcppDist)
library(magrittr)
library(data.table)
library(MAPLE)
#load the summary data for exposure and outcome
exp = fread(paste0("betax.assoc.txt"),head=T)
exp_raw = exp[,c("rs","beta","se","af","allele1","allele0","p_wald","n_obs")]
colnames(exp_raw) = c("SNP","b","se","frq_A1","A1","A2","P","N")
out = fread(paste0("betay.assoc.txt"),head=T)
out_raw = out[,c("rs","beta","se","af","allele1","allele0","p_wald","n_obs")]
colnames(out_raw) = c("SNP","b","se","frq_A1","A1","A2","P","N")

paras = est_SS(dat1 = exp_raw,
               dat2 = out_raw,
               trait1.name = "exp",
               trait2.name = "out",
               ldscore.dir = "./eur_w_ld_chr")
```

The input from summary statistics:

- **dat1**: GWAS summary-level data for exposure, including
  1. rs number,
  2. effect allele,
  3. the other allele,
  4. sample size,
  5. a signed summary statistic (used to calculate z-score).

  For example, the **dat1** with 3 SNPs can be represented as follows:

```
       SNP           b          se      frq_A1  A1 A2    P      N
1: rs144155419 -0.0001011242 0.04969397  0.010  A  G 0.9983764 19563
2:  rs58276399 -0.0040463990 0.01609338  0.111  C  T 0.8014823 19125
3: rs141242758 -0.0062410460 0.01608838  0.111  C  T 0.6980775 19179
```

- **dat2**: GWAS summary-level data for outcome which is similar as dat1.
- **trait1.name**: specify the name of exposure, default `exposure`.
- **trait2.name**: specify the name of outcome, default `outcome`.
- **ldscore.dir**: specify the path to the LD score files.

  The argument **ldscore.dir** specifies the path to LD score files. Because the two GWASs for this example are based on European samples, we can use the LD score files in *example* file, which are provided by the ldsc software (https://github.com/bulik/ldsc). These LD Scores were computed using 1000 Genomes European data. Users can also calculate the LD scores by themselves.

Users can specify the rs number, effect allele, and the other allele using the arguments "snp_col," "A1_col," and "A2_col," respectively. Users may designate one or both of the following columns for calculating z-scores: "b_col" (effect size), "se_col" (standard error), "z_col" (z-score), and "p_col" (p-value). The sample size can be defined using the "n_col" argument. Alternatively, in the absence of a designated sample size column, users can utilize the "n" argument to indicate the total sample size for each SNP. Incorporating the minor allele frequency ("freq_col") column, if available, is advisable as it aids in filtering out low-quality SNPs.

The function `est_SS` will also conduct the following quality control procedures:

- extract SNPs in HapMap 3 list,
- remove SNPs with minor allele frequency < 0.05 (if freq_col column is available),
- remove SNPs with alleles not in (G, C, T, A),
- remove SNPs with ambiguous alleles (G/C or A/T) or other false alleles (A/A, T/T, G/G or C/C),
- exclude SNPs in the complex Major Histocompatibility Region (Chromosome 6, 26Mb-34Mb),
- remove SNPs with $\chi^2 > \chi^2_{max}$. The default value for χ^2_{max} is $max(N/1000, 80)$.

Now, we can check the estimates with the following commands:

```
paras$Omega
#             [,1]        [,2]
#[1,]  0.99591955 -0.02936449
#[2,] -0.02936449  1.01181862
paras$Omega.se
#           [,1]       [,2]
#[1,] 0.04231683 0.02257965
#[2,] 0.02257965 0.02649007
```

The output contains:

- **Omega**: the estimate of $\mathbf{\Omega}$, the off-diagonal elements of `Omega` are the intercept estimate of cross-trait LD score regression; the diagonal elements of `Omega` are the intercept estimates of single-trait LD score regressions.
- **Omega.se**: the estimated matrix consists of the standard errors of the intercept estimates obtained from LD score regression.

Users have the option to skip this step and set the estimate `Omega` of $\mathbf{\Omega}$ to the identity matrix if there is no confounding arising from sample structure.

3.2 Step 2: Running MAPLE

The `MAPLE` function utilizes a scalable sampling-based algorithm to acquire calibrated p-values.

```
#load the z-score
zscorex = fread(paste0("zscorex.txt"),head=F)
Zscore_1 = as.vector(zscorex[[1]])
zscorey = fread(paste0("zscorey.txt"),head=F)
Zscore_2 = as.vector(zscorey[[1]])

#load the LD matrix
Sigmaxin = fread(paste0("Sigmax.txt"),head=F)
Sigma1in = as.matrix(Sigmaxin)
Sigmayin = fread(paste0("Sigmay.txt"),head=F)
Sigma2in = as.matrix(Sigmayin)

#load the sample size
samplen1 = 20000
samplen2 = 20000

#load the nuisance error parameter
t1 = paras$Omega[1,1]
t2 = paras$Omega[2,2]
t12 = paras$Omega[1,2]

result = MAPLE(Zscore_1,Zscore_2,Sigma1in,Sigma2in,samplen1,samplen2,Gibbsnumber=1000,
               burninproportion=0.2,pi_beta_shape=0.5,pi_beta_scale=4.5,
               pi_c_shape=0.5,pi_c_scale=9.5,pi_1_shape=0.5,pi_1_scale=1.5,
               pi_0_shape=0.05,pi_0_scale=9.95,t1,t2,t12)
```

The input from summary statistics:

- **Zscore_1**: the Zscore vector of the SNP effect size vector for the exposure.
- **Zscore_2**: the Zscore vector of the SNP effect size vector for the outcome.
- **Sigma1in**: the LD matrix for the SNPs in the exposure GWAS data.
- **Sigma2in**: the LD matrix for the SNPs in the outcome GWAS data.
- **samplen1**: the sample size of exposure GWAS.
- **samplen2**: the sample size of outcome GWAS.
- **Gibbsnumber**: the number of Gibbs sampling iterations with the default to be 1000.
- **burninproportion**: the proportion to burn in from Gibbs sampling iterations, with default to be 20%.
- **pi_beta_shape**: the prior shape paramter for π_β with the default to be 0.5.
- **pi_beta_scale**: the prior scale paramter for π_β with the default to be 4.5.
- **pi_c_shape**: the prior shape paramter for π_c with the default to be 0.5.
- **pi_c_scale**: the prior shape paramter for π_c with the default to be 9.5.
- **pi_1_shape**: the prior shape paramter for π_1 with the default to be 0.5.
- **pi_1_scale**: the prior scale paramter for π_1 with the default to be 1.5.
- **pi_0_shape**: the prior shape paramter for π_0 with the default to be 0.05.
- **pi_0_scale**: the prior scale paramter for π_0 with the default to be 9.95.
- **t1**: the intercept estimate of sigle-trait LD score regression from exposure data.
- **t2**: the intercept estimate of sigle-trait LD score regression from outcome data.
- **t12**: the intercept estimate of cross-trait LD score regression from exposure and outcome data.

Note that, we used $p = 5 \times 10^{-8}$ for MAPLE to select candidate IVs without LD clumping. Additionally, users can employ the same LD matrix derived from either exposure data, outcome data, or an LD reference panel as **Sigma1in** and **Sigma2in**, provided that no additional LD matrices are available for the SNPs in the exposure and outcome data, respectively.

Now, we can check the estimates from MAPLE:

```
result$causal_effect
#[1] -0.0404195
result$causal_pvalue
#[1] 0.23265
result$cause.se
#[1] 0.03386465
```

The output from `MAPLE` is a list containing:

- **causal_effect**: the estimate of causal effect.
- **causal_pvalue**: the p value for the causal effect.
- **cause.se**: the standard error of causal effect.
- **correlated_pleiotropy_effect**: The confounder effect on the outcome (ω).
- **sigmaeta**: the variance estimate for the uncorrelated pleiotropy effect.
- **sigmabeta**: the variance estimate for the SNP effect sizes on the exposure.

3.3 Parameter selection

The R package of MAPLE allows users to tune several parameters, including the number of Gibbs sampling iterations (**Gibbsnumber**, default=1000), the burn-in proportion (**burninproportion**, default=20%), and the prior shape and scale parameters for four key proportion parameters:

- π_β (defaults: pi_beta_shape=0.5 and scale pi_beta_scale=4.5);
- π_c (defaults: pi_c_shape=0.5 and pi_c_scale=9.5);
- π_1 (defaults: pi_1_shape=0.5 and pi_1_scale=1.5);
- π_0 (defaults: pi_0_shape=0.05 and pi_1_scale=9.95).

Different choices of priors for π_β, π_1 and π_0 do not substantially affect the performance of MAPLE, while the choice of prior for π_c have a certain impact on causal effect testing, with $Beta(0.5, 9.5)$ as a potentially promising choice. In practice, we suggest users to assess robustness of causal effect estimation with different priors.