METEOR User Manual

Liye Zhang

15122960682@163.com

Contents

1 Introduction

1.1 What is METEOR

METEOR (Multivariate Estimation Tool for Exposure-Outcomes with Robustness), is an R package for efficient statistical inference of multi-outcomes mendelian randomization analysis (https://github.com/Liye222/METEOR). METEOR utilizes a set of correlated SNPs, self-adaptively accounts for the sample structure of both exposure and outcomes, the uncertainty that these correlated SNPs may exhibit multiple pleiotropic effects. The term 'self-adaptive' represents that METEOR is able to automatically infer the sample structure and the probability that a SNP has specific pleiotropic effect from the data at hand. METEOR places the inference of the causal effects into a likelihood-framework and relies on a scalable sampling-based algorithm to obtain calibrated p-values.

1.2 The METEOR method

Our goal is to simultaneously estimate and test causal effects of an exposure on multiple outcomes with self-adaptive determination of pleiotropy and sample structure. We consider a general model in which multiple outcomes may originate from different datasets, and we allow sample overlap between exposure and outcome. We denote $\mathbf{z} = \left(\mathbf{z_x}^T, \mathbf{z_{y_1}}^T, \cdots, \mathbf{z_{y_k}}^T\right)^T$, $\boldsymbol{\alpha} = (1, \alpha_1, \cdots, \alpha_k)^T$,$\boldsymbol{\eta} = (\boldsymbol{\eta}_0^T, \boldsymbol{\eta}_1^T, \cdots, \boldsymbol{\eta}_k^T)^T$.The METEOR model can be constructed as follows,

$$\mathbf{z} = \mathbf{I}_{k+1} \bigotimes (\mathbf{\Sigma}\boldsymbol{\beta}) \cdot (\boldsymbol{\xi} \circ \boldsymbol{\alpha}) + \mathbf{I}_{k+1} \bigotimes \mathbf{\Sigma} \left(\left(\boldsymbol{\xi} \bigotimes \mathbf{1}_p \right) \circ \boldsymbol{\eta} \right) + \boldsymbol{\epsilon}$$

$\mathbf{z}_x$ 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; $\mathbf{z}_{y_i}$ is a p-vector of marginal z-scores measuring the association between the candidate SNPs and the i-th outcome, with n_{2i} individuals in the i-th outcome GWAS; $\boldsymbol{\beta}$ is a p-vector of correlated SNPs effects on the exposure; α_i is a scalar that represents the causal effect of the exposure on the i-th outcome; $\boldsymbol{\eta}_i$ is a p-vector of horizontal pleiotropic effects on the i-th outcome and $\boldsymbol{\eta}_0$ is a p-vector of zeros. $\boldsymbol{\epsilon}$ is residual term. Additionally, $\boldsymbol{\xi} = \left(\sqrt{n_1 - 1}, \sqrt{n_{21} - 1}, \cdots, \sqrt{n_{2k} - 1}\right)^T$, $\mathbf{I}_{k+1}$ is a

k-dimensional identity matrix, $\mathbf{1}_p$ is a p-vector of ones. $\bigotimes$ denotes Kronecker product, and the term $(\boldsymbol{\xi} \circ \boldsymbol{\alpha})$ represents the Hadamard product, also known as the element wise product, of the two vectors $\boldsymbol{\xi}$ and $\boldsymbol{\alpha}$.

$$\mathbf{z} \sim MVN\left(E\left(\mathbf{z}\right), \boldsymbol{\Omega} \bigotimes \boldsymbol{\Sigma}\right)$$

$$E\left(\mathbf{z}\right) = \left(E\left(\mathbf{z_x}\right)^T, E\left(\mathbf{z_{y_1}}\right)^T, \cdots, E\left(\mathbf{z_{y_k}}\right)^T\right)$$

$$E\left(\mathbf{z_x}\right) = \sqrt{n_1 - 1}\boldsymbol{\Sigma}\boldsymbol{\beta}$$

$$E\left(\mathbf{z_{y_i}}\right) = \sqrt{n_{2i} - 1}\boldsymbol{\Sigma}\left(\boldsymbol{\beta}\alpha_i + \boldsymbol{\eta}_i\right), i = 1, \cdots, k$$

where $\mathbf{z}$ follows a multivariate normal distribution, the p by p row covariance matrix $\boldsymbol{\Sigma}$ characterizes the covariance among the marginal z-scores across SNPs, and the $k+1$ by $k+1$ column covariance matrix $\boldsymbol{\Omega}$ characterizes the covariance between the marginal z-scores on the exposure and and k outcomes to account for sample structure, such as sample overlap and any correlations among them.

2 Installation

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

```
#install.packages("devtools")
library(devtools)
install_github("Liye222/METEOR")
```

3 Application analysis

Note: All data used in the manual can be found at https://github.com/Liye222/METEOR/tree/main/example. We consider a simple situation with two outcomes.

3.1 Step 1: Estimation of correlation matrix Ω

The function `Omega_est` or `Omega_est_nopar` can estimate the parameter $\boldsymbol{\Omega}$ using to account for sample structure (e.g., population stratification, sample overlap and any correlations among them).

```
library(Rcpp)
library(RcppArmadillo)
library(RcppDist)
library(magrittr)
library(data.table)
library(METEOR)
#load the summary data for exposure and two outcomes
load(file=paste0("./exposure.rda")) ##file name: exp
dat_x <- exp[,c("SNP","b","se","frq_A1","A1","A2","P","N")]
load(file=paste0("./outcome1.rda")) ##file name: out1
dat_y1 <- out1[,c("SNP","b","se","frq_A1","A1","A2","P","N")]
load(file=paste0("./outcome2.rda")) ##file name: out2
dat_y2 <- out2[,c("SNP","b","se","frq_A1","A1","A2","P","N")]

summarydata <- list(dat_x,dat_y1,dat_y2)
```

Users can use two functions, `Omega_est` or `Omega_est_nopar`, to estimate $\mathbf{\Omega}$. The former utilizes `parallel`, which accelerates the computing time but requires more cores. The latter does not deed multiple cores, but its computing speed is relatively slow.

Function 1: `Omega_est`

```
library(parallel)
Omega <- Omega_est(data_list=summarydata,
                   ldscore.dir = "./eur_w_ld_chr",
                   nCores=NA,
                   system_used="linux")
```

Function 2: `Omega_est_nopar`

```
Omega <- Omega_est_nopar(data_list=summarydata,
                         ldscore.dir = "./eur_w_ld_chr")
```

The input from summary statistics:

- **summarydata**: a list of GWAS-summary-level data for exposure (*dat_x*) and two outcomes (*dat_y1*,*dat_y2*)

  dat_x: 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 *dat_x* 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
```

  dat_yi: GWAS summary-level data for the i-th outcome which is similar as *dat_x*.

- **ldscore.dir**: specify the path to the LD score files.
- **nCores**: The number of required cores or *NA*.
- **system_used**: The system used.

The argument **ldscore.dir** specifies the path to LD score files. Because the GWASs used for this example are based on European samples, we can use the LD score files from https://github.com/yuanzhongshang/MAPLE/tree/main/example/eur_w_ld_chr, 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 functions `Omega_est` and `Omega_est_nopar` 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:

```
Omega
#$Omega
#             [,1]        [,2]        [,3]
#[1,]  1.06320820 -0.04611552 -0.01709125
#[2,] -0.04611552  1.02055474  0.49859715
#[3,] -0.01709125  0.49859715  1.03129944

#$Omega_se
#            [,1]       [,2]       [,3]
#[1,] 0.03158760 0.02169158 0.01934258
#[2,] 0.02169158 0.03412790 0.02355318
#[3,] 0.01934258 0.02355318 0.04228503
```

The output contains:

- **Omega**: the estimate of $\mathbf{\Omega}$, the off-diagonal elements of `Omega` are the intercept estimates 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 METEOR

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

```
#load the z-score
zscorex =fread(paste0("./zscorex.txt"),head=F)
zx<-as.matrix(zscorex,ncol=1)
zscorey_1 =fread(paste0("./zscorey1.txt"),head=F)
zy1<-as.vector(zscorey_1[[1]])
zscorey_2 =fread(paste0("./zscorey2.txt"),head=F)
zy2<-as.vector(zscorey_2[[1]])
Zscore <- cbind(zx,zy1,zy2)
Zscore <- t(Zscore)
#load the LD matrix
sigma<-fread(paste0("./Sigma.txt"),head=F)
Sigma <- as.matrix(sigma)
#load the sample size
N <- matrix(c(50000,50000,50000),ncol=1)
#load the correlation matrix and corresponding standard errors
load("./Omega.rda") #file name: Omega
Omega_est <- Omega$Omega
Omega_se <- Omega$Omega_se

result<-METEOR(Zscore,Sigma,N,Omega_est,Omega_se,Gibbsnumber=1000,burninproportion=0.2,
               pi_beta_shape=0.5,pi_beta_scale=4.5,pi_1_shape=0.5,pi_1_scale=1.5,
```

```
                pi_0_shape=0.05,pi_0_scale=9.95)
```

The input from summary statistics:

- **Zscore**: the Zscore with k+1 rows and p columns of the SNP effect size matrix for the exposure and k outcomes.
- **Sigma**: the LD matrix for the SNPs selected from the exposure can be obtained by using the weighted average LD matrix from k outcomes. If individual data is unavailable, the LD matrix can also be derived from a reference panel.
- **N**: the sample size of exposure and k outcomes GWASs.
- **Omega_est**: the correlation matrix derived by LDSC.
- **Omega_se**: the standard errors of elements of *Omega* derived by LDSC.
- **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%.
- **lambda**: the tuning parameter used to ensure that the correlation matrix is invertible.
- **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_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.

Note that, we use $p = 5 \times 10^{-8}$ for METEOR to select candidate IVs without LD clumping. However, if the number of SNPs is too much, such as (greater than 10000), suggesting using LD clumping with $r^2 = 0.5$ to select candidate IVs. Additionally, users can employ the LD matrix derived from the weighted average LD matrix from k outcomes, or an LD reference panel as **Sigma**, provided that no additional LD matrices are available for the SNPs in the k outcome data.

Now, we can check the estimates from METEOR:

```
result$causal_effect
#           [,1]       [,2]
#[1,] 0.0785606 0.01103452

result$causal_pvalue_single
#              [,1]     [,2]
#[1,] 1.577082e-05 0.595196

result$causal_pvalue_overall
#              [,1]
#[1,] 6.608792e-05
```

The output from `METEOR` is a list containing:

- **causal_effect**: the estimate of k causal effects.
- **causal_pvalue_single**: the p values for the causal effects in single tests.
- **causal_pvalue_overall**: the p value for the causal effect in overall tests.
- **cause_sd**: the standard deviation for the causal effects.
- **cause_cov**: the covariance for the causal effects.
- **sigmabeta**: the variance estimate for the SNP effect sizes on the exposure.
- **sigmaeta**: the variance estimates for the horizontal pleiotropy effects.
- **Omega**: the correlation matrix for one exposure and k outcomes.
- **pi_beta**: the proportion of selected SNPs, which show non-zero effects on exposure.
- **pi_1**: the proportion of selected SNPs showing horizontal pleiotropy.
- **pi_0**: the proportion of non-selected SNPs showing horizontal pleiotropy.