A real example for performing two-sample summary-data MR analysis with `MRAPSS` package

Xianghong Hu

13/03/2021

Introduction

The MR-APSS is a unified approach to Mendelian Randomization accounting for Pleiotropy and Sample Structure using genome-wide summary statistics. Specifically, MR-APSS uses a background-foreground model to characterize the estimated effects of SNPs on exposure and outcome traits, where the background model accounts for confounding from pleiotropy and sample structure, and the foreground model captures the valid signal for causal inference.

The `MRAPSS` package implements the MR-APSS approach to infer the causal relationship between an exposure and an outcome. We illustrate how to perform the MR-APSS model analysis using the MRAPSS package by a real example of BMI (exposure) and T2D (outcome). The MR-APSS model analysis comprises the following steps:

- Step 1: Prepare data and estimate nuisance parameters
- Step 2: Fit MR-APSS for causal inference

Step 0: Installation and loading packages

```
#install.packages("devtools")
#install.packages("readr")
devtools::install_github("YangLabHKUST/MR-APSS")
```

```
library(MRAPSS)
library(readr)
```

Step 1: Prepare data and estimate nuisance parameters

This step comprises three sub-steps:

- 1.1. Format data;
- 1.2. Harmonize the formatted data sets and estimate nuisance parameters;
- 1.3. LD clumping.

1.1. Format data

To begin, we should set a working directory using function *setwd*. Then, we download the GWAS summary statistics for BMI (Watanabe et.al. (2019) [PMID: 31427789]) and T2D (Xue et al. (2015) [PMID: 30054458]). We rename the two downloaded files as "BMI_ukb.txt" and "T2D.txt", and read them into R. For this example, users can also skip this step and use the formatted data sets we have prepared below.

```
BMI_raw <- readr::read_delim("BMI_ukb.txt", "\t", escape_double = FALSE,
                             trim_ws = TRUE, progress = F)
#> Parsed with column specification:
```

```
#> cols(
#>   .default = col_double(),
#>   SNP = col_character(),
#>   A1 = col_character(),
#>   TEST = col_character(),
#>   A2 = col_character(),
#>   SNPID_UKB = col_character(),
#>   A1_UKB = col_character(),
#>   A2_UKB = col_character()
#> )
#> See spec(...) for full column specifications.

T2D_raw <- readr::read_delim("T2D.txt", " ",
                             escape_double = FALSE,
                             trim_ws = TRUE, progress = F)
#> Parsed with column specification:
#> cols(
#>   CHR = col_double(),
#>   BP = col_double(),
#>   SNP = col_character(),
#>   A1 = col_character(),
#>   A2 = col_character(),
#>   frq_A1 = col_double(),
#>   b = col_double(),
#>   se = col_double(),
#>   P = col_double(),
#>   N = col_double()
#> )
```

Similar to LDSC, MR-APSS needs the following information from each GWAS data set:

- rs number,
- effect allele,
- the other allele,
- sample size,
- a signed summary statistic (anything that can be converted to a z-score).

The function *format_data* will try to interpret the raw data sets with user-specified column names. Users can set rs number, effect allele and the other allele using "snp_col","A1_col" and "A2_col" arguments. Users can set one or two of the following columns which can be used to calculate z scores: "b_col" (effect size),"se_col" (standrd error), "or_col" (odds ratio), "z_col" (z-score), "p_col" (p-value). Users can set sample size using "n_col", or "ncase_col" and "ncontrol_col". If the column for sample size is not available, users can use argument "n" to specify the total sample size for each SNP. It could be preferred if minor allele frequency ("freq_col") and INFO ("info_col") are available, which helps to remove low-quality SNPs.

The function *format_data* 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),
- remove SNPs with INFO < 0.9 (if info_col column is available),

- exclude SNPs in the complex Major Histocompatibility Region (Chromosome 6, 26Mb-34Mb),
- remove SNPs with $\chi^2 > \chi 2_{max}$. The default value for $\chi 2_{max}$ is $\max(N/1000, 80)$.

Data sets will be formatted to have the following columns using function *format_data*:

- SNP: rs number,
- A1: effect allele,
- A2: the other allele,
- Z: z score,
- chi2: χ^2 statistics,
- P: p-value,
- N: sample size.

Now, we use function *format_data* to obtain the formatted data sets for BMI and T2D.

```
BMI_ukb = format_data(BMI_raw,
                      snp_col = "SNPID_UKB",
                      b_col = "BETA",
                      se_col = "SE",
                      freq_col = "MAF_UKB",
                      A1_col = "A1",
                      A2_col = "A2",
                      p_col = "P",
                      n_col = "NMISS",
                      info_col = "INFO_UKB")
#> Begin formatting ....
#> The raw data set has 10599054 dat lines
#> Remove SNPs with imputation info less than 0.9 ..., remaining 10599018 SNPs.
#> Remove ambiguous SNPs ..., remaining 8995068 SNPs.
#> Remove SNPs in MHC region ..., remaining 8944044 SNPs.
#> Remove duplicated SNPs ..., remaining 8944044 SNPs.
#> Merge SNPs with the hapmap3 snplist ..., remaining 1161501 SNPs.
#> Remove SNPs with alleles not matched with the hapmap3 snplist, remaining 1161195 SNPs.
#> Remove SNPs with p-value < 0 or p-value > 1, remaining 1037824 SNPs.
#> Inferring z score from p value and b ...
#> Remove missing values, remaining 1037824 SNPs.
#> Remove SNPs with sample size 5 standard deviations away from the mean, remaining 1037824 SNPs.
#> Remove SNPs with chi2 > chi2_max ... , remaining 1037799 SNPs.
#> The formatted data has 1037799 dat lines.

T2D = format_data(T2D_raw,
                  snp_col = "SNP",
                  b_col = "b",
                  se_col = "se",
                  freq_col = "frq_A1",
                  A1_col = "A1",
                  A2_col = "A2",
                  p_col = "P",
                  n_col = "N")
#> Begin formatting ....
#> The raw data set has 5053015 dat lines
#> Remove ambiguous SNPs ..., remaining 4280791 SNPs.
```

```
#> Remove SNPs in MHC region ..., remaining 4274525 SNPs.
#> Remove duplicated SNPs ..., remaining 4274525 SNPs.
#> Merge SNPs with the hapmap3 snplist ..., remaining 1007638 SNPs.
#> Remove SNPs with p-value < 0 or p-value > 1, remaining 936658 SNPs.
#> Inferring z score from p value and b ...
#> Remove missing values, remaining 936658 SNPs.
#> Remove SNPs with sample size 5 standard deviations away from the mean, remaining 936565 SNPs.
#> Remove SNPs with chi2 > chi2_max ... , remaining 936556 SNPs.
#> The formatted data has 936556 dat lines.
```

For this example, we have already prepared the formatted data sets. If users want to skip this step, they can directly download the formatted data sets here and read them into R.

Let's have a look at the formatted data sets:

```
head(BMI_ukb)
#>          SNP A1 A2          Z      N       chi2       P
#> 1  rs1000000  A  G -0.2879304 385270 0.08290393 0.77340
#> 2 rs10000010  C  T -0.7299839 376552 0.53287656 0.46540
#> 3  rs1000002  T  C -1.9961279 385336 3.98452670 0.04592
#> 4 rs10000023  G  T  2.5662659 371893 6.58572050 0.01028
#> 5  rs1000003  G  A -1.5813414 383709 2.50064060 0.11380
#> 6 rs10000033  C  T  1.4286669 382629 2.04108907 0.15310
```

```
head(T2D)
#>          SNP A1 A2          Z      N       chi2       P
#> 1  rs1000000  A  G -0.2881917 573633 0.08305446 0.77320
#> 2 rs10000010  C  T  0.5018117 587226 0.25181496 0.61580
#> 3  rs1000002  T  C -2.5639113 579347 6.57364100 0.01035
#> 4 rs10000023  G  T  0.3003634 557936 0.09021815 0.76390
#> 5  rs1000003  G  A -1.6381040 577450 2.68338473 0.10140
#> 6 rs10000033  C  T  1.9173774 576206 3.67633619 0.05519
```

1.2. Harmonize the formatted data sets and estimate nuisance parameters

At this step, we will first merge and harmonize the formatted data sets from step 1.1 to make sure effect sizes of the same SNP for the exposure and outcome correspond to the same allele. Then, we will use the harmonized data to estimate $\mathbf{\Omega}$ and $\mathbf{C}$ of the background model by LD score regression.

The analysis for this step is accomplished by function *est_paras*. The argument "ldscore.dir" in function *est_paras* 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 at this link, 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.

```
paras = est_paras(dat1 = BMI_ukb,
                  dat2 = T2D,
                  trait1.name = "BMI",
                  trait2.name = "T2D",
                  ldscore.dir = "./eur_w_ld_chr")
#> Merge dat1 and dat2 by SNP ...
#> Harmonize the direction of SNP effects of exposure and outcome
#> Read in LD scores ...
#> Add LD scores to the harmonized data set...
#> The Harmonized data set will also be used for MR analysis
#> Begin estimation of C and Omega using LDSC ...
```

```
#> Estimate heritability for trait 1 ...
#> Using two-step estimator with cutoff at 30.
#> Mean Chi2:3.0256.
#> Intercept: 1.1521(0.025).
#> Total Observed scale h2:0.2174(0.0071).
#> Estimate heritability for trait 2 ...
#> Using two-step estimator with cutoff at 30.
#> Mean Chi2:1.7045.
#> Intercept: 1.0911 (0.0167).
#> Total Observed scale h2:0.0485 (0.0028).
#> Estimate genetic covariance ...
#> Using two-step estimator with cutoff at 30.
#> Intercept: 0.1618 (0.0145).
#> Total Observed scale gencov: 0.054 (0.003).
```

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

```
paras$Omega
#>               [,1]         [,2]
#> [1,] 1.852220e-07 4.599093e-08
#> [2,] 4.599093e-08 4.132106e-08
```

```
paras$C
#>           [,1]      [,2]
#> [1,] 1.1520504 0.1617687
#> [2,] 0.1617687 1.0911428
```

Note that the off-diagonal elements of $\hat{\mathbf{C}}$ are the intercept estimate of cross-trait LD score regression; the diagonal elements of $\hat{\mathbf{C}}$ are the intercept estimates of single-trait LD score regressions. The diagonal elements can be fixed at 1 by specifying "h2.intercept = T" in function *est_params* if necessary.

The harmonized data set will be used for LD clumping at the next sub-step.

```
head(paras$dat)
#>          SNP A1 A2         b.exp         b.out      se.exp      se.out
#> 1  rs1000000  A  G -0.0004638793 -0.0003805085 0.001611081 0.001320331
#> 2 rs10000010  C  T -0.0011895996  0.0006548444 0.001629624 0.001304960
#> 3  rs1000002  T  C -0.0032156486 -0.0033684772 0.001610943 0.001313804
#> 4 rs10000023  G  T  0.0042081639  0.0004021192 0.001639800 0.001338776
#> 5  rs1000003  G  A -0.0025528462 -0.0021556799 0.001614355 0.001315960
#> 6 rs10000033  C  T  0.0023096280  0.0025259149 0.001616632 0.001317380
#>   pval.exp pval.out       L2
#> 1  0.77340  0.77320 20.90684
#> 2  0.46540  0.61580 23.52035
#> 3  0.04592  0.01035 46.00095
#> 4  0.01028  0.76390 25.62940
#> 5  0.11380  0.10140 37.94950
#> 6  0.15310  0.05519 10.62294
```

1.3. LD clumping

We use LD clumping to obtain a subset of nearly independent SNPs as IVs from the harmonized data set. The analysis for this step is accomplished by function *clump*. The p-value threshold for IV selection can be specified by argument "IV.Threshold". The PLINK clumping procedure will be used for LD clumping. At this step, we will also calculate an adjusted IV threshold to account for bias due to LD clumping where SNPs with smaller p-values will be selected as IVs. The adjusted IV threshold is obtained by multiplying the IV

threshold by the ratio of the median after the LD clumping to the median before LD clumping. It is typically less or equal to the specified IV threshold and will be used as a parameter for correction of the winner's curse in MR-APSS.

```
MRdat = clump(paras$dat,
              IV.Threshold = 5e-05,
              SNP_col = "SNP",
              pval_col = "pval.exp",
              clump_kb = 1000,
              clump_r2 = 0.001)
#> API: public: http://gwas-api.mrcieu.ac.uk/
#> Please look at vignettes for options on running this locally if you need to run many instances of th
#> Clumping xrx2Ke, 26527 variants, using EUR population reference
#> Removing 25300 of 26527 variants due to LD with other variants or absence from LD reference panel
```

Now, we can have a look at the data for selected IVs:

```
# The "Threshold" column presents the adjusted IV threshold.
head(MRdat)
#>              SNP A1 A2        b.exp        b.out      se.exp      se.out
#> 956  rs10009336  T  C -0.006792164 -0.002548263 0.001611882 0.001286171
#> 2262  rs1002226  C  T -0.008782110  0.010755686 0.001616300 0.001336168
#> 4794  rs1004807  T  C -0.010160801 -0.004629696 0.001626313 0.001330145
#> 6192  rs1006195  T  G  0.011168351  0.003045159 0.001627663 0.001324508
#> 6325  rs1006316  C  T  0.008082369  0.002534664 0.001612065 0.001320965
#> 7438  rs1007392  G  A -0.007985049  0.001987821 0.001611163 0.001320388
#>       pval.exp  pval.out       L2    Threshold
#> 956  2.511e-05 4.756e-02 35.09356 2.193202e-05
#> 2262 5.527e-08 8.303e-16 34.95761 2.193202e-05
#> 4794 4.164e-10 5.003e-04 37.67589 2.193202e-05
#> 6192 6.810e-12 2.150e-02 22.44078 2.193202e-05
#> 6325 5.340e-07 5.501e-02 19.69577 2.193202e-05
#> 7438 7.193e-07 1.322e-01 60.18694 2.193202e-05
```

By default, *clump* performs LD clumping through API, which means that installation of PLINK is not required (see https://github.com/MRCIEU/ieugwasr). Alternatively, users can perform LD clumping locally using PLINK with *clump* by specifying the LD reference panel with "bfiles" and the path to local PLINK binary with "plink_bin" (see the details of *clump*).

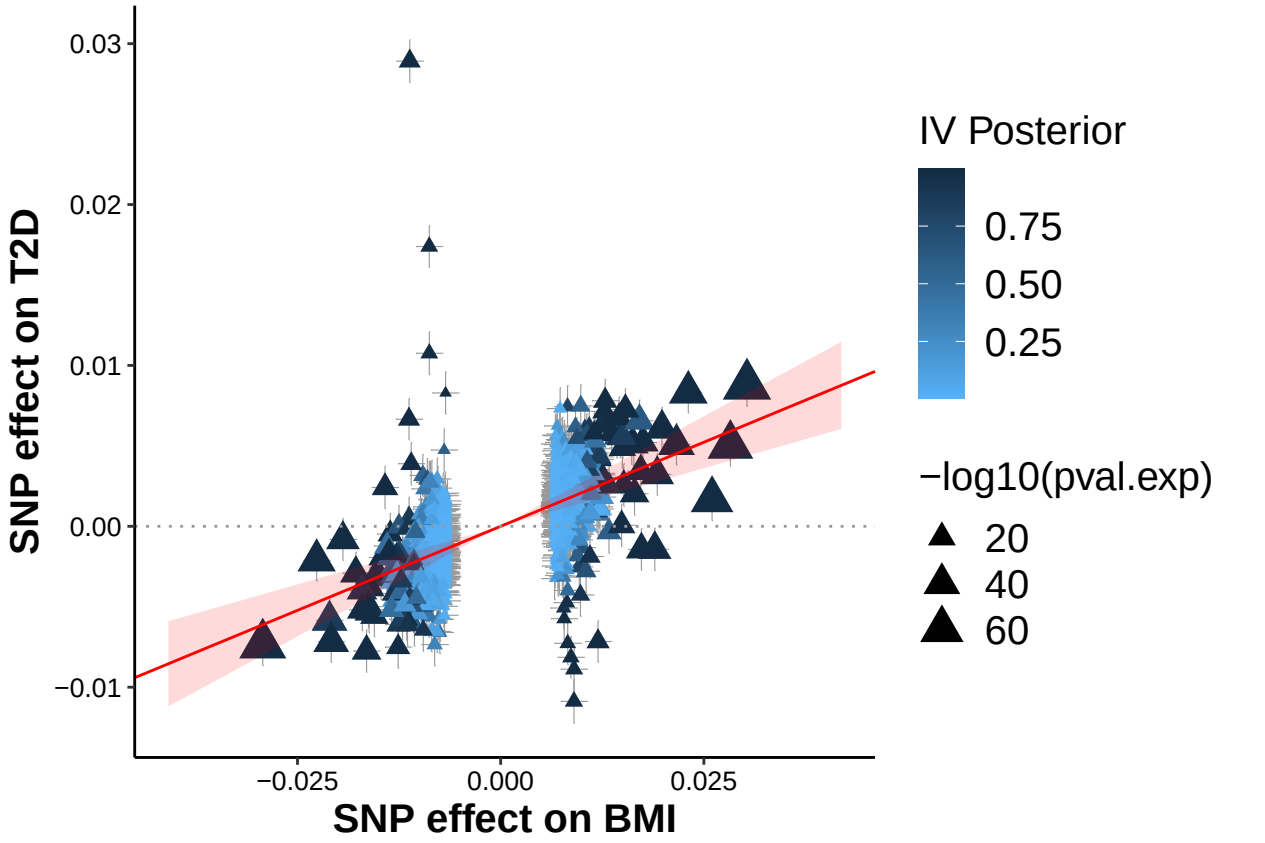
Step 2: Fit MR-APSS for causal inference

Now we can fit the MR-APSS model with the selected IVs (MRdat) and the pre-estimated background parameters $\hat{\mathbf{\Omega}}$ and $\hat{\mathbf{C}}$.

```
MRres = MRAPSS(MRdat,
               exposure = "BMI",
               outcome = "T2D",
               C = paras$C,
               Omega = paras$Omega ,
               Cor.SelectionBias = T)
#> ***********************************************************
#> MR test results of  BMI  on  T2D :
#> MR-APSS: beta =  0.209 , beta.se =  0.033 , p-value =  2.3976e-10 .
#> Total NO. of IVs=  1227 , NO. of valid IVs with foreground signals:  178.019 .
#> ***********************************************************
```

We can visualize MR-APSS analysis results by function *MRplot*:

```
MRplot(MRres, exposure = "BMI", outcome = "T2D")
```



The figure shows the results of MR-APSS using the default IV threshold 5×10^{-5}. The estimated causal effect is indicated by a red line with its 95% confidence interval indicated by the shaded area in transparent red color. Triangles indicate the observed SNP effect sizes ($\hat{\gamma}_j$ and $\hat{\Gamma}_j$). Error bars indicate their standard errors. The color of triangles indicates the posterior of a valid IV, i.e., the posterior of an IV carrying the foreground signal ($Z_j = 1$, dark blue) or not ($Z_j = 0$, light blue).

Reference

Watanabe K, Stringer S, Frei O, Umićević Mirkov M, de Leeuw C, Polderman TJC, van der Sluis S, Andreassen OA, Neale BM, Posthuma D. A global overview of pleiotropy and genetic architecture in complex traits. Nat Genet. 2019 Sep;51(9):1339-1348.

Xue A, Wu Y, Zhu Z, Zhang F, Kemper KE, Zheng Z, Yengo L, Lloyd-Jones LR, Sidorenko J, Wu Y; eQTLGen Consortium, McRae AF, Visscher PM, Zeng J, Yang J. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. Nat Commun. 2018 Jul 27;9(1):2941.

Bulik-Sullivan, B., Loh, PR., Finucane, H. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet 47, 291–295 (2015). https://doi.org/10.1038/ng.3211

Xianghong Hu, Jia Zhao, Zhixiang Lin, Yang Wang, Heng Peng, Hongyu Zhao, Xiang Wan, Can Yang. MR-APSS: a unified approach to Mendelian Randomization accounting for pleiotropy and sample structure using genome-wide summary statistics. bioRxiv 2021.03.11.434915; doi: https://doi.org/10.1101/2021.03.11.434915.