Powerful genome-wide design and robust statistical inference in two-sample summary-data Mendelian randomization

Qingyuan Zhao, Yang Chen, Jingshu Wang, Dylan S. Small
University of Pennsylvania, University of Michigan, University of Pennsylvania, University of Pennsylvania

Abstract:

Background. Mendelian randomization (MR) uses genetic variants as instrumental variables to estimate the causal effect of risk exposures in epidemiology. Two-sample summary-data MR that uses publicly available genome-wide association studies (GWAS) summary data have become a popular design in practice. With the sample size of GWAS continuing to increase, it is now possible to utilize genetic instruments that are only weakly associated with the exposure.

Methods. To maximize the statistical power of MR, we propose a genome-wide design where more than a thousand genetic instruments are used. For the statistical analysis, we use an empirical partially Bayes approach where instruments are weighted according to their true strength, thus weak instruments bring less variation to the estimator. The final estimator is highly efficient in the presence of many weak genetic instruments and is robust to balanced and/or sparse pleiotropy.

Results. We apply our method to estimate the causal effect of blood lipids on coronary artery disease. In our primary analysis, the estimated odds ratio (95% CI) for LDL cholesterol, HDL cholesterol, and triglycerides are 1.61 (1.45 – 1.80), 0.82 (0.73 – 0.91), and 1.00 (0.84 – 1.21), respectively. Compared to previous MR studies, these numbers are closer to observational epidemiology estimates and much more precise. We also discuss diagnostics of the modeling assumptions and caveats of the results.

Conclusions. By employing a genome-wide design and robust statistical methods, the statistical power of MR studies can be greatly improved. Unlike previous MR studies which all reported null findings for the HDL cholesterol, our results give support to the much debated HDL hypothesis.

Keywords: Conditional score, HDL hypothesis, Partially Bayes, Robust statistics, Spike-and-slab prior.

Key messages:

- We utilize common variants across the whole genome (typically over a thousand) as instrumental variables.

- We extend a previously proposed method—robust adjusted profile score—to account for the measurement error in GWAS summary data and biases due to weak instruments. A new method—

\[ Address for correspondence: \] Department of Statistics, The Wharton School, University of Pennsylvania, Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340 USA. E-mail: qyzhao@wharton.upenn.edu. May 15, 2018.

\[ arXiv:1804.07371v2 [stat.AP] 17 May 2018 \]
empirical partially Bayes—is developed to increase the statistical power when some genetic instruments are strong but many are very weak. The estimator is robust to balanced and/or sparse pleiotropy.

• Our new and more powerful analysis finds the protective effect of HDL cholesterol on coronary artery disease is statistically significant ($p$-value = 0.0003), while all previous MR studies could not reject the null.

• Code to replicate the results (including diagnostics) is available in the R package mr.raps (https://github.com/qingyuanzhao/mr.raps).
1 Introduction

Mendelian randomization (MR) is a method of using genetic variation to infer the causal effect of a modifiable risk exposure on disease outcome. Since MR can give unbiased estimates in presence of unmeasured confounding, it has become a widely used tool for epidemiologists and health scientists [1]. A prominent example is the overwhelming evidence of a causal link between low-density lipoprotein cholesterol (LDL-c) and coronary heart disease found by several MR studies [2, 3, 4, 5], which is consistent with the results of earlier landmark clinical trials [6].

However, when gathering all the evidence from observational studies, Mendelian randomizations, and clinical trials, the role of other blood lipids, especially the high-density lipoprotein cholesterol (HDL-c), remain quite controversial [5, 7, 8]. Observational epidemiology studies have long suggested that HDL-c is inversely associated with the risk of myocardial infarction [9, 10, 11], but a recent clinical trial reported no reduction of risk of recurrent cardiovascular events for patients taking dalcetrapib, a drug shown to increase HDL-c levels while not substantially changing LDL-c [12]. Existing Mendelian randomization studies also provided ambiguous evidence [13, 14, 15]. These inconclusive results have created a lot of confusion about the “HDL is good” dogma in medical practice [8, 16].

From a statistical perspective, MR is a special instance of instrumental variable (IV) methods [17, 18]. Compared to classical applications of IV methods in economics [19] and health research [20], the most distinctive feature of MR is the enormous number of candidate instruments. Potentially, any one of the millions of single nucleotide polymorphisms (SNP) in the human genome can be used as an IV as long as it satisfies the following three validity criteria [20]:

1. Relevance: the SNP must be associated with the risk exposure.

2. Effective random assignment: the SNP must be independent of any unmeasured confounder.

3. Exclusion restriction (ER): the SNP affects outcome only through the risk exposure.

Among the three criteria above, the relevance and ER assumptions are often disputable in MR practice. The ER assumption is especially concerning due to a widespread phenomenon called pleiotropy [21, 22], a.k.a. multiple functions of genes. For example, it is common to find SNPs that are associated with both LDL-c and HDL-c, thus its effect on cardiovascular outcomes are possibly mediated by both lipids. When these SNPs are used as instruments in a MR analysis of HDL-c, the ER assumption is likely violated.
To alleviate these concerns, most existing MR studies select a handful of genome-wide significant SNPs that are associated with the exposure and then seek to justify that the ER assumption is reasonable. This simple design is very transparent, but it has some major limitations. First, a full justification of the ER assumption requires a deep understanding of the causal mechanism of the genes and can be invalidated by new findings. For example, Katan, an early exponent of MR, proposed to use polymorphic forms of the \textit{APOE} gene to estimate the causal effect of blood cholesterol on cancer. However, as Davey Smith and Ebrahim later argued, they may be invalid instruments due to pleiotropic effects on other biomarkers. Second, the statistical power of MR is greatly limited as the vast majority of SNPs are excluded, including a lot of known genetic variation of the exposure and the outcome (Figure 1).

Meanwhile, it is well known to econometricians and statisticians that weak IVs can still provide valuable information, especially if there are a number of them. In a previous paper, we found that using weakly significant SNPs can greatly increase the efficiency of MR studies. In a different but related application, Bulik et al. also found that a genome-wide analysis is much more powerful than using just the significant SNPs to estimate the genetic correlation. Using weak instruments also helps to identify candidate IVs that do not satisfy the ER assumption, and the causal effect can still be consistently estimated when the invalid IVs are rare or the pleiotropic effects are balanced.

In this paper, we will introduce two new strategies that can greatly increase the statistical power of MR studies. The first innovation is a truly genome-wide design: unlike previous MR studies including our earlier work, no threshold will be used in the selection phase. Typically, over a thousand independent genetic instruments are used in a single study. The blueprint of genome-wide MR has been considered in the literature, but it was not feasible until recently because most existing summary-data MR methods are heavily biased by weak IVs.

Our second innovation is an estimator that adaptively assigns weights to the IVs according to their strength. This method is based on a general empirical partially Bayes approach introduced by Lindsay. Our previous method of adjusting the profile score can be viewed as a special case in which a flat prior is used. Both approaches yield consistent and asymptotically normal estimators of the causal effect, but using an empirically estimated prior increases statistical efficiency.
2 Genome-wide design

We will use a working example to illustrate the genome-wide MR design. In this example the goal is to estimate the causal effect of HDL-c on the risk of coronary artery disease (CAD). Our two-sample summary-data MR makes use of three non-overlapping GWAS:

1. **Selection dataset**: A GWAS of blood lipids involving ≥100,000 individuals [35];

2. **Exposure dataset**: A follow-up GWAS of blood lipids using Metabochip with about 100,000 individuals [36];

3. **Outcome dataset**: A GWAS of CAD conducted by the CARDIoGRAMplusC4D consortium involving 63,746 cases and 130,681 controls [37].

For each GWAS, the summary data are publicly available, which report the linear or logistic regression coefficients and standard errors (typically following a meta-analysis) of all the genotyped or imputed SNPs.

We preprocess the data to select genetic instruments for our statistical analysis. We first remove SNPs that do not coappear in all three datasets. Then we use the remaining selection dataset to find independent SNPs (distance ≥10 megabase pairs, linkage disequilibrium $R^2 ≤ 0.001$) that are most associated with the target blood lipid (HDL-c) and are not associated with the other two lipids (LDL-c and TG, $p$-value ≥ 0.01). This is done in a greedy fashion using the linkage-disequilibrium (LD) clumping function in the PLINK software [38]. Using independent SNPs makes the statistical analysis more convenient and is common in MR studies [23]. Suppose $p$ SNPs are selected after LD clumping. In our primary analysis of HDL-c, $p = 1122$.

After preprocessing, we obtain $2p$ marginal genetic effect estimates (linear/logistic regression coefficients) and their standard errors from the exposure and outcome datasets:

- $\hat{\gamma}_j$, $j = 1, \ldots, p$ are the genetic effects on the exposure (HDL-c). The standard errors are denoted by $\sigma_{Xj}$.

- $\hat{\Gamma}_j$, $j = 1, \ldots, p$ are the genetic effects on the outcome (CAD). The standard errors are denoted by $\sigma_{Yj}$.

This type of dataset is typically called a two-sample summary-data MR [39]. Using a separate and non-overlapping dataset for SNP selection is very important for the unbiasedness of the genetic effect estimates, thus avoiding any “winner’s curse” bias [40]. A common misconception is that, when
the same GWAS is used for selection and to obtain \( \hat{\gamma}_j \), the “winner’s curse” could be eliminated by only using genome-wide significant SNPs. This is not true, because although these SNPs are most likely true hits, the estimated genetic effects \( \hat{\gamma}_j \) are still biased. As a consequence, the causal effect estimate is generally biased towards zero [28].

Figure 1 shows the distribution of genetic signal strength measured by \( \|\gamma\|^2 \) (for HDL-c) and \( \|\Gamma\|^2 \) (for CAD) as a function of the selection threshold. These quantities are closely related to the genetically inherited phenotypic variance and can be unbiasedly estimated by \( \sum_j \hat{\gamma}_j^2 - \sigma_{\chi_j}^2 \) and \( \sum_j \hat{\Gamma}_j^2 - \sigma_{\tau_j}^2 \) over the SNPs that pass the selection threshold. Compared to the conventional analysis that only uses 23 genome-wide significant SNPs (\( p \)-value \( \leq 5 \times 10^{-8} \)), the genome-wide MR design using all the 1122 SNPs contains almost twice the signal for HDL-c and more than ten times the signal for CAD. This suggests a great potential of gaining statistical power.
3 Statistical model

Following our previous article [28], our main modeling assumptions are:

Assumption 1 (Measurement error model).

\[
\begin{pmatrix}
\hat{\gamma} \\
\hat{\Gamma}
\end{pmatrix}
\sim
N
\left(
\begin{pmatrix}
\gamma \\
\Gamma
\end{pmatrix},
\begin{pmatrix}
\Sigma_X & 0 \\
0 & \Sigma_Y
\end{pmatrix}
\right),
\Sigma_X = \text{diag}(\sigma^2_{X_1}, \ldots, \sigma^2_{X_p}),
\Sigma_Y = \text{diag}(\sigma^2_{Y_1}, \ldots, \sigma^2_{Y_p}).
\]

Assumption 2 (Pleiotropy model). We assume the causal effect \(\beta\) satisfies \(\Gamma_j \approx \beta \gamma_j\) for most \(j = 1, \ldots, p\). More specifically, let \(\alpha = \Gamma - \beta \gamma\). We assume \(\alpha_j\) is independent of \(\gamma_j\) and most \(\alpha_j \overset{\text{ind}}{\sim} N(0, \tau^2)\) where \(\tau^2\) is a small overdispersion parameter. A few SNPs might deviate from this model and have very large \(|\alpha_j|\).

The normality and independence assumptions in Assumption 1 can be immediately justified by the large sample size of GWAS, non-overlapping samples in the selection, exposure, and outcome datasets, and independence of the SNPs [28]. We have also implicitly assumed the standard errors \(\sigma_{X_j}, \sigma_{Y_j}, j = 1, \ldots, p\) reported in the GWAS are well calibrated.

Assumption 2 presumes that for most SNPs, the genetic associations with the exposure and the outcome approximately satisfy a pair-wise linear relationship, with the common slope parameter being the causal effect \(\beta\). Where all the SNPs are valid IVs, the linear relationship \(\Gamma = \beta \gamma\) can be derived by assuming the SNPs, exposure and outcome satisfy a linear structural model [41] and can be extended to nonlinear structural models by assuming the per-SNP effects are minuscule and the SNPs affect the exposure in a homogenous way [28]. When the outcome variable is the binary disease status, \(\beta\) may be interpreted as a conservative estimate of the causal log-odds-ratio [28].

In reality, many genetic variants may violate the ER assumption and have other pathways to affect CAD. For example, a SNP selected for the MR study of HDL-c might also be associated with LDL-c, so its genetic association with CAD includes both the causal effect of HDL-c and LDL-c. Motivated by our previous analysis of the effect of obesity on blood pressure [28], the robust MR model in Assumption 2 considers two types of deviations from the exact linear relationship \(\Gamma = \beta \gamma\): 1. small and balanced pleiotropy represented by the random effects model \(\alpha_j \sim N(0, \tau^2)\); 2. idiosyncratic and large pleiotropy. The first kind of deviation is a special case of the InSIDE (Instrument Strength Independent of Direct Effect) assumption [41, 42], while the second is similar to the sparse invalid IV assumption [30]. We think it is crucial that the statistical method of MR
is robust to both kinds of pleiotropy.

4 Statistical method

In our previous article [28], we proposed a robust estimator based on adjusting the profile score function of $\beta$ and $\tau^2$, in which the nuisance parameters $\gamma$ are profiled out. Here we propose a new method to eliminate the nuisance parameters $\gamma$ based on an empirical partially Bayes approach introduced by Lindsay [34]. The new method has a simple geometric interpretation and can further increase the statistical power.

4.1 Empirical partially Bayes

We first consider the simplest scenario where $\alpha = 0$. The key insight is gained from deriving the contribution of the $j$-th SNP to the conditional score function (Appendix A):

$$C_j(\beta, \gamma_j) = \frac{\gamma_j (\hat{\Gamma}_j - \beta \hat{\gamma}_j)}{\beta^2 \sigma_{Xj}^2 + \sigma_{Yj}^2} \beta \sigma_{Xj}^2 + \sigma_{Yj}^2.$$

It is straightforward to verify that the maximum likelihood estimator (MLE) of $\gamma_j$ given $\beta$, denoted as $\hat{\gamma}_j, \text{MLE}(\beta)$, is a sufficient statistic of $\gamma_j$ and is independent of $\hat{\Gamma}_j - \beta \hat{\gamma}_j$. The decoupling of “instrument strength” $\gamma_j$ and “regression residual” $\hat{\Gamma}_j - \beta \hat{\gamma}_j$ in $C_j(\beta, \gamma_j)$ motivates us to consider a general class of estimating functions:

$$C(\beta) = \sum_{j=1}^p f_j(\beta, \hat{\gamma}_j, \text{MLE}(\beta)) \cdot \psi(\hat{\Gamma}_j - \beta \hat{\gamma}_j) \frac{\beta^2 \sigma_{Xj}^2 + \sigma_{Yj}^2}{\beta^2 \sigma_{Xj}^2 + \sigma_{Yj}^2},$$

where $f_j$ is an arbitrary function of $\beta$ and $\hat{\gamma}_j, \text{MLE}$, and $\psi$ is an odd function. Because $\hat{\gamma}_j, \text{MLE}(\beta)$ and $\hat{\Gamma}_j - \beta \hat{\gamma}_j$ are independent, it is easy to show that $E[C(\beta)] = 0$ at the true $\beta$. Therefore the root of $C(\beta)$, denoted by $\hat{\beta}$, is a reasonable estimator of $\beta$. Geometrically, this estimating function finds the $\beta$ such that a transformation (by $f$) of the estimated “instrument strength” $\hat{\gamma}_j, \text{MLE}(\beta)$ is uncorrelated with a transformation (by $\psi$) of the “regression residual” $\hat{\Gamma}_j - \beta \hat{\gamma}_j$.

Different choices of the weighting scheme $f_j$ do not change the unbiasedness of $C(\beta)$, but may affect the statistical efficiency. The profile score developed in our previous article [28] amounts to using $\hat{\gamma}_j, \text{MLE}(\beta)$ as the weight. To maximize statistical power, Lindsay [34] suggested to use the
empirical Bayes estimate of $\gamma_j$ as the weight,

$$f_j(\beta, \hat{\gamma}_{j, \text{MLE}}(\beta)) = \hat{\gamma}_{j, \text{EB}}(\beta) = \mathbb{E}_{\pi_\eta} [\gamma_j | \hat{\gamma}_{j, \text{MLE}}(\beta)]$$

where $\pi_\eta$ is a prior distribution of $\gamma$ and $\hat{\eta}$ is an empirical estimate of the prior parameter. Intuitively, $\hat{\gamma}_{j, \text{EB}}$ shrinks $\hat{\gamma}_{j, \text{MLE}}$ towards 0. The function $\psi$ is chosen to limit the influence of large outliers (Section 4.3).

4.2 Spike-and-slab prior

In principle, a good choice of the prior distribution $\pi_\eta$ should have the following properties: 1. the parametric family $\pi_\eta$ should fit the distribution of $\gamma$ reasonably well, so we can gain efficiency by using the empirical partially Bayes estimator; 2. The empirical Bayes estimator of $\gamma_j$ should be easy to compute since it will be evaluated many times when iteratively solving the estimating equations. For these reasons, we choose to use a spike-and-slab Gaussian mixture prior [43, 44] to model $\gamma_j/\sigma_{Xj}$:

$$\gamma_j/\sigma_{Xj} \sim \pi_{p_1, \sigma_1, \sigma_2} = p_1 \cdot N(0, \sigma_1^2) + (1-p_1) \cdot N(0, \sigma_2^2).$$

We decide to model the effect sizes $\gamma_j/\sigma_{Xj}$ instead of the effects $\gamma_j$ because this scale is more familiar and the shrinkage rule is easier to interpret. Typically, $p_1$ is close to 1, $\sigma_1^2$ is close to zero (the spike component), and $\sigma_2^2$ is much larger than $\sigma_1^2$ (the slab component). The selective shrinkage offered by the spike-and-slab prior [45] is essential to gain efficiency in empirical partially Bayes (Appendix A.2).

4.3 Robust estimator

To account for invalid IVs in Assumption 2 we need to further estimate the overdispersion parameter $\tau^2 = \text{Var}(\alpha_j)$ while being robust to large outliers of $\alpha_j$. Intuitively, we need two estimating equations after eliminating the nuisance $\gamma$: one for $\beta$ and one for $\tau^2$. For $\beta$, we can follow the empirical partially Bayes approach described above by replacing $\sigma_{Y_j}^2$ with $\sigma_{Y_j}^2 + \tau^2$. For $\tau^2$, we need to adjust the profile score function of $\tau^2$ due to a Neyman-Scott phenomenon [46, 28]. To be robust against outliers, we propose to use a bounded function of the “regression residual” $\hat{\Gamma}_j - \beta \hat{\gamma}_j$.

Next we describe the robust estimating function of $\beta$ and $\tau^2$. Derivation of these functions is very similar to our previous RAPS (Robust Adjusted Profile Score) approach [28] and the details are omitted. Let $\psi_1(\cdot)$ and $\psi_2(\cdot)$ be two differentiable odd functions. The empirical partially Bayes ver-
sion of the RAPS estimator \((\hat{\beta}, \hat{\tau}^2)\) is given by the solution to \(\tilde{C}(\beta, \tau^2) = (\tilde{C}_1(\beta, \tau^2), \tilde{C}_2(\beta, \tau^2))^T = 0\), where

\[
\tilde{C}_1(\beta, \tau^2) = \sum_{j=1}^{p} \frac{\hat{\gamma}_{j,EB}(\beta, \tau^2) \cdot \psi_1(t_j(\beta, \tau^2))}{s_j(\beta, \tau^2)}, \quad \text{and}
\]

\[
\tilde{C}_2(\beta, \tau^2) = \sum_{j=1}^{p} \frac{\psi_2(t_j(\beta, \tau^2)) - \delta}{s_j^2(\beta, \tau^2)}, \quad \text{where} \quad \delta = \mathbb{E}[\psi_2(Z)] \text{ for } Z \sim N(0, 1),
\]

where \(t_j(\beta, \tau^2) = (\hat{\Gamma}_j - \beta \hat{\gamma}_j)/s_j(\beta, \tau^2)\) is the standardized regression residual and \(s_j(\beta, \tau^2) = \sqrt{\beta^2 \sigma^2_{X_j} + \sigma^2_{Y_j} + \tau^2}\).

In the numerical analysis below, we will use \(\psi_2(t) = t \cdot \psi_1(t)\) and consider two choices of \(\psi_1\): the non-robust identity function \(\psi_I\) and the robust Huber’s score function \(\psi_H\). Further implementation details including how to compute the standard error of \(\hat{\beta}\) can be found in Appendix A.2.

### 4.4 Diagnostics

A potential concern of using many weak instruments is that they might have more pleiotropy (i.e. more likely to violate the ER assumption) than strong instruments [22, 47]. In general, the conclusions of a genome-wide MR study are stronger if the weak instruments and strong instruments produce similar estimates of the causal effect. We propose a simple diagnostic plot for this purpose, where the standardized regression residual \(t_j(\beta, \tau^2)\) is plotted against (a standardized version of) the estimated instrument strength

\[
\hat{\gamma}_{j,EB}(\beta, \tau^2) / \sqrt{\text{Var}(\hat{\gamma}_{j,EB}(\beta, \tau^2))},
\]

evaluated at \((\beta, \tau^2) = (\hat{\beta}, \hat{\tau}^2)\). At the true value of \((\beta, \tau^2)\) and when our modeling assumptions are satisfied, \(t_j\) should be independent of \(\hat{\gamma}_{j,EB}\) and follow a standard normal distribution (possibly with some outliers). This proposition can be empirically checked in a diagnostic plot (Figure 2, Appendix A.2.5, Figure 4).

### 5 Validation studies

We perform two validation studies of the proposed statistical method. The first is a simulation study to mimic the real data analysis of HDL-c. The true causal effect \(\beta\) is set to be \(-0.2\) and the GWAS summary data are simulated in five settings: NO (No Outlier), ALL (All SNPs), ST
(Strong SNPs), WK (Weak SNPs), MVW (Many Very Weak SNPs). More specifically, the effect sizes $\gamma_j/\sigma_{Xj}$, $j = 1, \ldots, p$ are generated from the Gaussian mixture distribution with

**Settings NO and ALL** $p = 1122$, $p_1 = 0.91$, $\sigma_1 = 0.73$, $\sigma_2 = 4.57$, resembling the analysis of HDL-c using all the 1122 SNPs.

**Setting STR** $p = 23$, $p_1 = 0.99$, $\sigma_1 = \sigma_2 = 7.85$, resembling the analysis using the 23 SNPs that are genome-wide significant in the *selection* GWAS.

**Setting WK** $p = 1099$, $p_1 = 0.83$, $\sigma_1 = 0.58$, $\sigma_2 = 2.17$, resembling the analysis using the 1099 SNPs that are not significant in the *selection* GWAS.

**Setting MVW** $p = 1122$, $p_1 = 0.95$, $\sigma_1 = 0.2$, $\sigma_2 = 5$, a setting similar to ALL but the weak instruments are even weaker.

Next $\Gamma$ is generated by $\Gamma_j = \beta \gamma_j + \alpha_j$ where $\alpha_j$ is an independent Gaussian variable with mean 0 and variance $3.2 \times 10^{-5}$, the estimated $\tau^2$ in our primary HDL-c analysis. In all settings besides NO, we include an outlier (add $5\tau$ to the $\alpha_j$ corresponding to the strongest SNP) to test the method’s robustness to large idiosyncratic pleiotropy. Finally, the standard deviations ($\sigma_{Xj}, \sigma_{Yj}$) are the same as the standard errors in the primary analysis of HDL-c.

Table 1 shows the results of this simulation study for four estimators: MR-Egger [42], weighted median [31], RAPS with MLE weights, and RAPS with shrinkage weights (both using the Huber score function). The RAPS estimators are unbiased in setting NO and have much smaller bias in other settings with a large outlier. They are also more precise and powerful than MR-Egger and weighted median. Among the two RAPS estimator, the one with shrinkage weights is about 10% more efficient in settings NO, ALL and WK and about 50% more efficient in setting MVW where the “spike” and the “slab” have greater disparity.

Our second validation study uses a real dataset where the golden truth of $\beta$ is essentially known. In this example, all three datasets—*selection*, *exposure*, *outcome* (as explained in Section 2)—are GWAS of coronary heart disease. In particular, the UK Biobank GWAS of self-reported myocardial infarction is used to select SNPs, the C4D GWAS [48] is used as the *exposure dataset*, and the CARDIoGRAM GWAS [49] is used as the *outcome dataset*. Since the exposure and the outcome are the same variable, it is expected that $\gamma$ and $\Gamma$ are the same in our model. Thus $\beta = 1$ and $\tau^2 = 0$. 

11
Table 1: Simulation results using 1000 replications in 5 settings: NO (No Outlier), ALL (All SNPs), STR (Strong SNPs), WK (Weak SNPs), MVW (Many Very Weak SNPs). For each estimator we report four metrics: mean of $\hat{\beta}$ (the true $\beta = -0.2$), root-mean-squared error (RMSE), coverage of the 95% CI, and statistical power (proportion of 95% CI not covering 0).

In Table 2 we apply our statistical methods to this validation dataset in three ways: using all the SNPs, using SNPs that are genome-wide significant in the selection dataset ($p$-value $\leq 5 \times 10^{-8}$), and using SNPs that are not genome-wide significant in the selection dataset. In all cases the point estimates are close to the truth $\beta = 1$. The shrinkage estimator is about 10% more efficient when using all the SNPs. When only the strong instruments are used so there is virtually no shrinkage, the shrinkage estimator is essentially the same as the non-shrinkage estimator. MR-Egger and weighted median are heavily biased by the weak instruments in this validation study.
Table 2: Validation study: both the exposure and the outcome are coronary artery disease, so $\beta = 1$. Our estimators are roughly unbiased and the shrinkage estimator is about 10% more efficient when all SNPs are used.

6 Application to the effect of blood lipids on cardiovascular diseases

We apply our method to analyze the causal effect of blood lipids on cardiovascular diseases. The main results and comparison with previous studies can be found in Table 3. More detailed results of using various estimators on three set of SNPs (all the 1122 SNPs, the 23 significant SNPs, and the 1099 non-significant SNPs) can be found in Table 4.

6.1 Main results

Our results for LDL-c (OR 1.61, 95% CI 1.45–1.80) are similar to previous MR studies, which all find the harmful effect of LDL-c is highly significant. For HDL-c, all previous MR studies have reported null findings, but we find the protective effect is statistically significant (OR 0.82, 95% CI 0.73–0.91, $p$-value = 0.0003). A closer look at the results reveal that our CI is much shorter than previous studies and the RAPS analysis using the 23 significant SNPs. Therefore, adopting a genome-wide MR design and a more efficient statistical analysis substantially increased the statistical power. For TG, our results (OR 1.00, 95% CI 0.84–1.21) suggest that its causal effect is not statistically significant, however the diagnostic plot shows that some of our modeling assumptions are likely violated in analyzing TG (Section 6.2).

A secondary analysis using an independent outcome dataset—the UK BioBank GWAS of self-reported myocardial infarction is reported in Appendix B. In short, the findings about LDL-c and HDL-c replicate in the secondary analysis, but the results for TG are quite different.
|                          | LDL-c          | HDL-c          | TG             |
|--------------------------|----------------|----------------|----------------|
| **Observational studies**|                |                |                |
| Angelantonio et al. (2009) [11] | 1.50 (1.39–1.61) | 0.78 (0.74–0.82) | 0.99 (0.94–1.05) |
| Voight et al. (2012) [13]    | 1.54 (1.45–1.63) | 0.62 (0.58–0.66) | 1.42 (1.31–1.52) |
| **Previous MR studies**     |                |                |                |
| Voight et al. (2012) [13]    | 2.13 (1.69–2.69) | 0.93 (0.68–1.26) | Not reported   |
| Holmes et al. (2014) [14]    | 1.92 (1.68–2.19) | 0.96 (0.70–1.31) | 1.26 (1.00–1.61) |
| White et al. (2016) [15]     | 1.68 (1.51–1.87) | 0.95 (0.85–1.06) | 1.28 (1.13–1.45) |
| **New MR analysis**         |                |                |                |
| Using all SNPs (main analysis) | 1.61 (1.45–1.80) | 0.82 (0.73–0.91) | 1.00 (0.84–1.21) |
| Using significant SNPs      | 1.76 (1.53–2.03) | 0.88 (0.74–1.04) | Not available† |
| Using non-significant SNPs  | 1.25 (1.02–1.54) | 0.75 (0.62–0.92) | 1.53 (1.01–2.33) |

Table 3: Results (estimated odds ratio and 95% confidence interval) of the primary analysis and comparison with previous observational studies and MR studies. The primary analysis uses RAPS with spike-and-slab shrinkage and Huber’s score function $\psi_H$, which correspond to the bolded numbers in Table 4.

* The original results for the restricted SNPs are reported per 1 mmol/L increase of LDL-c and HDL-c and 1 log unit increase of TG. We transformed the results to per 1 SD increase using the following approximate estimates of the standard deviation: SD(LDL-c) = 1 mmol/L, SD(HDL-c) = 0.4 mmol/L, SD(Log TG) = 0.5.

† Result is not available because there are multiple roots of the robust estimating function (Appendix A.2.3).
### Table 4: Primary analysis for coronary artery disease (CAD).

|                  | $p_{sel} \in (0, 1)$ | $p_{sel} \in (0, 5 \times 10^{-8})$ | $p_{sel} \in (5 \times 10^{-8}, 1)$ |
|------------------|-----------------------|-------------------------------------|-------------------------------------|
| # SNPs           | 1120                  | 27                                  | 1093                                |
| $p_1             | 0.92                  | 0.01                                | 0.91                                |
| $\sigma_1$      | 0.52                  | 7.96                                | 0.46                                |
| $\sigma_2$      | 5.01                  | 7.96                                | 2.61                                |
| MR-Egger         | 0.368 (0.053)         | 0.892 (0.134)                       | 0.057 (0.08)                        |
| Wtd. Med.        | 0.534 (0.071)         | 0.552 (0.075)                       | 0.166 (0.094)                       |

|                  | MLE Shrinkage         | MLE Shrinkage                       | MLE Shrinkage                       |
|------------------|-----------------------|-------------------------------------|-------------------------------------|
| $\tau^2 = 0$, $\psi_I$ | 0.448 (0.06)          | 0.528 (0.055)                       | 0.528 (0.055)                       |
| $\tau^2 = 0$, $\psi_H$ | 0.466 (0.057)         | 0.56 (0.065)                        | 0.56 (0.065)                        |
| $\tau^2 \neq 0$, $\psi_I$ | 0.394 (0.06)         | 0.55 (0.073)                        | 0.55 (0.073)                        |
| $\tau^2 \neq 0$, $\psi_H$ | 0.41 (0.06)          | 0.566 (0.073)                       | 0.566 (0.073)                       |

(a) Low-density lipoprotein cholesterol (LDL-c).

|                  | $p_{sel} \in (0, 1)$ | $p_{sel} \in (0, 5 \times 10^{-8})$ | $p_{sel} \in (5 \times 10^{-8}, 1)$ |
|------------------|-----------------------|-------------------------------------|-------------------------------------|
| # SNPs           | 1122                  | 23                                  | 1099                                |
| $p_1             | 0.91                  | 0.01                                | 0.83                                |
| $\sigma_1$      | 0.73                  | 7.85                                | 0.58                                |
| $\sigma_2$      | 4.57                  | 7.85                                | 2.17                                |
| MR-Egger         | -0.117 (0.055)        | -0.018 (0.196)                      | -0.132 (0.08)                       |
| Wtd. Med.        | -0.118 (0.078)        | -0.108 (0.099)                      | -0.135 (0.086)                      |

|                  | MLE Shrinkage         | MLE Shrinkage                       | MLE Shrinkage                       |
|------------------|-----------------------|-------------------------------------|-------------------------------------|
| $\tau^2 = 0$, $\psi_I$ | -0.164 (0.059)        | -0.109 (0.063)                      | -0.109 (0.063)                      |
| $\tau^2 = 0$, $\psi_H$ | -0.21 (0.059)         | -0.147 (0.064)                      | -0.147 (0.064)                      |
| $\tau^2 \neq 0$, $\psi_I$ | -0.146 (0.06)        | -0.103 (0.085)                      | -0.103 (0.085)                      |
| $\tau^2 \neq 0$, $\psi_H$ | -0.186 (0.061)       | -0.126 (0.088)                      | -0.126 (0.088)                      |

(b) High-density lipoprotein cholesterol (HDL-c).

|                  | $p_{sel} \in (0, 1)$ | $p_{sel} \in (0, 5 \times 10^{-8})$ | $p_{sel} \in (5 \times 10^{-8}, 1)$ |
|------------------|-----------------------|-------------------------------------|-------------------------------------|
| # SNPs           | 1114                  | 8                                   | 1106                                |
| $p_1             | 0.98                  | 0.01                                | 0.97                                |
| $\sigma_1$      | 0.58                  | 10.25                               | 0.52                                |
| $\sigma_2$      | 7.04                  | 10.25                               | 2.48                                |
| MR-Egger         | 0.04 (0.075)          | -0.3 (0.354)                        | 0.198 (0.105)                       |
| Wtd. Med.        | 0.071 (0.109)         | 0.007 (0.108)                       | 0.079 (0.103)                       |

|                  | MLE Shrinkage         | MLE Shrinkage                       | MLE Shrinkage                       |
|------------------|-----------------------|-------------------------------------|-------------------------------------|
| $\tau^2 = 0$, $\psi_I$ | 0.19 (0.134)          | -0.208 (0.09)                       | -0.208 (0.09)                       |
| $\tau^2 = 0$, $\psi_H$ | 0.246 (0.111)         | 0.022 (0.085)                       | 0.022 (0.085)                       |
| $\tau^2 \neq 0$, $\psi_I$ | 0.126 (0.111)        | -0.177 (0.145)                      | -0.177 (0.145)                      |
| $\tau^2 \neq 0$, $\psi_H$ | 0.165 (0.109)       | 0.009 (0.092)                       | 0.009 (0.092)                       |

(c) Triglycerides (TG).

The numbers reported in this table can be interpreted as causal log odds ratio. When there were multiple roots to the estimating equation, the results were reported as NA (not available).
6.2 Diagnostics

A further separate analysis of the strong instruments (p-value $\leq 5 \times 10^{-8}$ in the selection GWAS) and the weak instruments (p-value $> 5 \times 10^{-8}$ in the selection GWAS) shows substantial heterogeneity of the results for LDL-c and TG (Table 3). This can be also visualized using the proposed diagnostic plot (Figure 2). For HDL-c the heterogeneity is not significant, i.e. the weak and strong instruments produced similar estimates of the causal effect of HDL-c. These diagnostics suggest that the results for HDL-c are more trustworthy and further analyses are required to reliably estimate the causal effect of TG.

7 Discussion

Our simulation and real data analyses in Sections 5 and 6 suggest that a genome-wide MR analysis is much more powerful than an analysis just using a small number of strong genetic instruments. The empirical partially Bayes technique further increases the statistical efficiency. Our primary analysis reaffirms the harmful effect and LDL-c and finds the protective effect of HDL-c on cardiovascular diseases is statistically significant (two-sided p-value $= 3.4 \times 10^{-4}$). These findings replicate in the secondary analysis (Appendix B).

Our empirical results should be interpreted cautiously as the real data analysis is still subject to several limitations:

1. Pleiotropy (violation of the exclusion restriction) seems to be wide-spread in the MR study of blood lipids. This phenomenon is particularly pertinent for TG, as it is rare to find a SNP exclusively associated with TG. The estimates for TG are thus quite unstable using different sets of instruments in the primary and secondary analyses. Although our statistical method is robust against systematic pleiotropy that is balanced (as modeled by $\alpha_j \sim \text{N}(0, \tau^2)$) and idiosyncratic pleiotropy that can be regarded as outliers, it is unclear if this holds for blood lipids and other MR studies.

2. Relatedly, the way we selected the genetic instruments does not rule out the possibility of horizontal pleiotropy through the other lipids. In our study design, a SNP is eligible for the HDL-c analysis if it is seemingly not associated with LDL-c or TG (p-values $> 0.01$ in the selection GWAS). However, this does not prove there is no association as the selection dataset has limited sample size. To alleviate this concern, a future work is to extend the RAPS
Figure 2: Diagnostic plots in the primary analysis. If our modeling assumptions are satisfied, most of the standardized residuals (y-axis) should be approximately independent of the instrument weight (x-axis), checked by the scatter-plot in the left panel, and roughly follow the standard normal distribution, checked by the quantile-quantile plot in the right panel.
approach to the multivariate MR setting where the exposure variables simultaneously include all three blood lipids.

3. In some cases (LDL-c and TG in the primary analysis, HDL-c and TG in the secondary analysis), the diagnostic plots show evidence of heterogeneity among the genetic instruments. Besides pleiotropy, another possible explanation is that the underlying biological mechanisms can be different for different genetic instruments. As a consequence, the “causal effect” $\beta$ might not be identical for all instruments, and the magnitude and direction of pleiotropy can be diverse.

4. The outcome dataset in the primary analysis (CARDIoGRAMplusC4D GWAS) has a small fraction of overlapping samples with the selection and exposure datasets. This means that the genetic associations $\hat{\gamma}_j$ and $\hat{\Gamma}_j$ for the same $j$ might have a small correlation. Although we do not expect this minor violation of our modeling assumption can drastically change the results, it would be valuable to take this account in future analyses.

Despite the above caveats, our analysis of HDL-c indicates that it may be too soon to dismiss the HDL hypothesis. Previous MR studies of HDL-c generally lack the statistically power to detect this relation and should not be regarded as a disproof of the HDL hypothesis. This example highlights how a genome-wide design can substantially increase the statistical efficiency of MR studies.

References

[1] Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Human Molecular Genetics. 2014;23(R1):R89–R98.

[2] Linsel-Nitschke P, Götz A, Erdmann J, Braenne I, Braund P, Hengstenberg C, et al. Life-long reduction of LDL-cholesterol related to a common variant in the LDL-receptor gene decreases the risk of coronary artery disease—a Mendelian Randomisation study. PloS ONE. 2008;3(8):e2986.

[3] Ference BA, Yoo W, Alesh I, Mahajan N, Mirowska KK, Mewada A, et al. Effect of long-term exposure to lower low-density lipoprotein cholesterol beginning early in life on the risk of coronary heart disease: a Mendelian randomization analysis. Journal of the American College of Cardiology. 2012;60(25):2631–2639.
[4] Myocardial Infarction Genetics Consortium Investigators. Inactivating mutations in NPC1L1 and protection from coronary heart disease. New England Journal of Medicine. 2014;371(22):2072–2082.

[5] Burgess S, Harshfield E. Mendelian randomization to assess causal effects of blood lipids on coronary heart disease: lessons from the past and applications to the future. Current Opinion in Endocrinology, Diabetes, and Obesity. 2016;23(2):124.

[6] Scandinavian Simvastatin Survival Study Group. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). Lancet. 1994;344(8934):1383–1389.

[7] Rader DJ, Hovingh GK. HDL and cardiovascular disease. Lancet. 2014;384(9943):618–625.

[8] Kingwell BA, Chapman MJ, Kontush A, Miller NE. HDL-targeted therapies: progress, failures and future. Nature Reviews Drug Discovery. 2014;13(6):445.

[9] Miller GJ, Miller NE. Plasma-high-density-lipoprotein concentration and development of ischaemic heart-disease. Lancet. 1975;305(7897):16–19.

[10] Prospective Studies Collaboration. Blood cholesterol and vascular mortality by age, sex, and blood pressure: a meta-analysis of individual data from 61 prospective studies with 55,000 vascular deaths. Lancet. 2007;370(9602):1829–1839.

[11] Emerging Risk Factors Collaboration. Major lipids, apolipoproteins, and risk of vascular disease. JAMA. 2009;302(18):1993–2000.

[12] Schwartz GG, Olsson AG, Abt M, Ballantyne CM, Barter PJ, Brumm J, et al. Effects of dalcetrapib in patients with a recent acute coronary syndrome. New England Journal of Medicine. 2012;367(22):2089–2099.

[13] Voight BF, Peloso GM, Orho-Melander M, Frikke-Schmidt R, Barbalic M, Jensen MK, et al. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. Lancet. 2012;380(9841):572–580.

[14] Holmes MV, Asselbergs FW, Palmer TM, Drenos F, Lanktree MB, Nelson CP, et al. Mendelian randomization of blood lipids for coronary heart disease. European Heart Journal. 2014;36(9):539–550.
[15] White J, Swerdlow DI, Preiss D, Fairhurst-Hunter Z, Keating BJ, Asselbergs FW, et al. Association of lipid fractions with risks for coronary artery disease and diabetes. JAMA Cardiology. 2016;1(6):692–699.

[16] Holmes MV, Ala-Korpela M, Smith GD. Mendelian randomization in cardiometabolic disease: challenges in evaluating causality. Nature Reviews Cardiology. 2017;14(10):577.

[17] Hernán MA, Robins JM. Instruments for causal inference: an epidemiologist’s dream? Epidemiology. 2006;17(4):360–372.

[18] Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. Statistical Methods in Medical Research. 2007;16(4):309–330.

[19] Angrist JD, Keueger AB. Does compulsory school attendance affect schooling and earnings? Quarterly Journal of Economics. 1991;106(4):979–1014.

[20] Baiocchi M, Cheng J, Small DS. Instrumental variable methods for causal inference. Statistics in Medicine. 2014;33(13):2297–2340.

[21] Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in complex traits: challenges and strategies. Nature Reviews Genetics. 2013;14(7):483–495.

[22] Verbanck M, Chen CY, Neale B, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. Nature Genetics. 2018;50:693–698.

[23] Hemani G, Zheng J, Wade KH, Laurin C, Elsworth B, Burgess S, et al. MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations. bioRxiv:078972. 2016;.

[24] Katan M. Apoupoprotein E isoforms, serum cholesterol, and cancer. Lancet. 1986;327(8479):507–508.

[25] Davey Smith G, Ebrahim S. "Mendelian randomization": can genetic epidemiology contribute to understanding environmental determinants of disease? International Journal of Epidemiology. 2003;32(1):1–22.
[26] Stock JH, Wright JH, Yogo M. A survey of weak instruments and weak identification in generalized method of moments. Journal of Business & Economic Statistics. 2002;20(4):518–529.

[27] Hansen C, Hausman J, Newey W. Estimation with many instrumental variables. Journal of Business and Economic Statistics. 2008;26(4):398–422.

[28] Zhao Q, Wang J, Hemani G, Bowden J, Small DS. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. arXiv preprint arXiv:180109652. 2018;.

[29] Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, et al. An atlas of genetic correlations across human diseases and traits. Nature Genetics. 2015;47(11):1236.

[30] Kang H, Zhang A, Cai TT, Small DS. Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. Journal of American Statistical Association. 2016;111(513):132–144.

[31] Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. Genetic Epidemiology. 2016;40(4):304–314.

[32] Guo Z, Kang H, Cai TT, Small DS. Confidence Intervals for Causal Effects with Invalid Instruments using Two-Stage Hard Thresholding with Voting. Journal of the Royal Statistical Society Series B (Methodological). 2018;to appear.

[33] Brion MJA, Benyamin B, Visscher PM, Smith GD. Beyond the single SNP: emerging developments in Mendelian randomization in the "Omics" era. Current Epidemiology Reports. 2014;1(4):228–236.

[34] Lindsay BG. Using empirical partially Bayes inference for increased efficiency. Annals of Statistics. 1985;13(3):914–931.

[35] Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010;466(7307):707.

[36] Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, et al. Discovery and refinement of loci associated with lipid levels. Nature Genetics. 2013;45(11):1274.
[37] The CARDIoGRAMplusC4D Consortium P Deloukas, Kanoni S, Willenborg C, Farrall M, Assimes TL, Thompson JR, et al. Large-scale association analysis identifies new risk loci for coronary artery disease. Nature Genetics. 2013;45(1):25.

[38] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics. 2007;81(3):559–575.

[39] Burgess S, Scott RA, Timpson NJ, Smith GD, Thompson SG, Consortium E. Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. European Journal of Epidemiology. 2015;30(7):543–552.

[40] Burgess S, Thompson SG, Collaboration CCG. Avoiding bias from weak instruments in Mendelian randomization studies. International Journal of Epidemiology. 2011;40(3):755–764.

[41] Bowden J, Del Greco M, Minelli C, Davey Smith G, Sheehan N, Thompson J. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. Statistics in Medicine. 2017;36(11):1783–1802.

[42] Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. International Journal of Epidemiology. 2015;44(2):512–525.

[43] Mitchell TJ, Beauchamp JJ. Bayesian variable selection in linear regression. Journal of American Statistical Association. 1988;83(404):1023–1032.

[44] George EI, McCulloch RE. Variable selection via Gibbs sampling. Journal of American Statistical Association. 1993;88(423):881–889.

[45] Ishwaran H, Rao JS. Spike and slab variable selection: frequentist and Bayesian strategies. Annals of Statistics. 2005;33(2):730–773.

[46] Neyman J, Scott EL. Consistent estimates based on partially consistent observations. Econometrica. 1948;16(1):1–32.

[47] Jordan DM, Verbanck M, Do R. The landscape of pervasive horizontal pleiotropy in human genetic variation is driven by extreme polygenicity of human traits and diseases. bioRxiv. 2018;p. 311332.
[48] Coronary Artery Disease (C4D) Genetics Consortium. A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. Nature Genetics. 2011;43(4):339.

[49] Schunkert H, König IR, Kathiresan S, Reilly MP, Assimes TL, Holm H, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nature Genetics. 2011;43(4):333.

[50] Burgess S, Thompson SG. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. American Journal of Epidemiology. 2015;181(4):251–260.

[51] Kiefer J, Wolfowitz J. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Annals of Mathematical Statistics. 1956;27(4):887–906.

[52] Feng L, Dicker LH. Approximate nonparametric maximum likelihood for mixture models: A convex optimization approach to fitting arbitrary multivariate mixing distributions. Computational Statistics & Data Analysis. 2018;122:80–91.

[53] Cox D. A note on partially Bayes inference and the linear model. Biometrika. 1975;62(3):651–654.

[54] Meng XL. Automated bias-variance trade-off: intuitive inadmissibility or inadmissible intuition. In: Chen MH, Müller P, Sun D, Ye K, Dey DK, editors. Frontiers of Statistical Decision Making and Bayesian Analysis. New York: Springer; 2010. .

[55] Lindsay B. Conditional score functions: some optimality results. Biometrika. 1982;69(3):503–512.

[56] Stefanski LA, Carroll RJ. Conditional scores and optimal scores for generalized linear measurement-error models. Biometrika. 1987;74(4):703–716.

[57] James W, Stein C. Estimation with quadratic loss. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability. vol. 1; 1961. p. 361–379.

[58] Efron B, Morris C. Stein’s estimation rule and its competitors—an empirical Bayes approach. Journal of the American Statistical Association. 1973;68(341):117–130.
[59] Huber PJ. Robust estimation of a location parameter. Annals of Mathematical Statistics. 1964;35(1):73–101.

[60] Maronna R, Martin RD, Yohai V. Robust Statistics: Theory and Methods. John Wiley & Sons; 2006.

[61] Churchhouse C, Neale B. Rapid GWAS of thousands of phenotype for 337,000 samples in the UK BioBank; 2017. Available from: http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank/
A Technical details

A.1 The empirical partially Bayes approach

We propose a new way of eliminating the nuisance parameters $\gamma$ that can increase the power of genome-wide MR studies when most IVs are very weak (i.e. $\gamma$ are very close to 0). The key idea is to view the errors-in-variables regression as a semiparametric problem: instead of treating the vector $\gamma$ as the nuisance parameter whose dimension grows as more SNPs are used, we treat the (empirical) distribution of $\gamma$ as the nuisance. This idea originates from the general solution given by Kiefer and Wolfowitz [51] to the Neyman-Scott problem in which the observed data are modeled by a mixture distribution. In principle, the statistical inference can be carried out by solving a nonparametric maximum likelihood problem, but the numerical problem is often extremely challenging [52].

We will take an empirical partially Bayes approach introduced by Lindsay [34] which is numerically feasible and still has several good theoretical properties. The approach is partially Bayes because only the nuisance parameters are modeled by a prior distribution [53, 54]. It is empirical Bayes because the prior distribution is estimated empirically using the observed data.

Consider the simplest scenario where $\alpha = 0$ and derive the conditional score function [55]. When $\alpha = 0$, the log-likelihood function of the data $(\hat{\gamma}, \hat{\Gamma})$ is given by

$$l(\beta, \gamma) = \prod_{j=1}^{p} l_{j}(\beta, \gamma_{j})$$

where

$$l_{j}(\beta, \gamma_{j}) = -\frac{(\hat{\gamma}_{j} - \gamma_{j})^{2}}{2\sigma_{X_{j}}^{2}} - \frac{(\hat{\Gamma}_{j} - \beta \gamma_{j})^{2}}{2\sigma_{Y_{j}}^{2}}.$$ 

Thus the score function of $\beta$ in the $j$-th SNP is given by

$$S_{j}(\beta, \gamma_{j}) = \frac{\partial}{\partial \beta} l_{j}(\beta, \gamma_{j}) = \frac{\gamma_{j}(\hat{\Gamma}_{j} - \beta \gamma_{j})}{\sigma_{Y_{j}}^{2}}.$$

and a sufficient statistic of the nuisance parameter $\gamma_{j}$ is

$$W_{j}(\beta) = \frac{\hat{\gamma}_{j}}{\sigma_{X_{j}}^{2}} + \frac{\beta \hat{\Gamma}_{j}}{\sigma_{Y_{j}}^{2}}.$$ (1)

When $\beta$ is given, the maximum likelihood estimator (MLE) of $\gamma_{j}$ is

$$\hat{\gamma}_{j, \text{MLE}}(\beta) = \frac{W_{j}(\beta)}{1/\sigma_{X_{j}}^{2} + \beta^{2}/\sigma_{Y_{j}}^{2}}.$$ 

The conditional score function is the residual of the score function $S_{j}$ conditioning on $W_{j}$. After
some algebra, we obtain

\[ C_j(\beta, \gamma_j) = S_j(\beta, \gamma_j) - \mathbb{E} \left[ S_j(\beta, \gamma_j) | W_j(\beta) \right] = \frac{\gamma_j (\hat{\Gamma}_j - \beta \hat{\gamma}_j)}{\beta^2 \sigma_{X_j}^2 + \sigma_{Y_j}^2}. \]  

(2)

We would like to make three remarks on the conditional score (2). First, it is proportional to the “regression residual” \( \hat{\Gamma}_j - \beta \hat{\gamma}_j \) which has mean 0 at the true \( \beta \). Second, the nuisance parameter appears in (2) only as a weight to the “regression residual”, as noticed by Lindsay [34]. Third, the sufficient statistic \( W_j(\beta) \) of \( \gamma_j \) in (1) is independent of \( \hat{\Gamma}_j - \beta \hat{\gamma}_j \) because they are jointly normal and their covariance is 0, regardless of what \( \beta \) is. See [56] for a related application of the conditional score in measurement error models. These observations motivate the following estimating function of \( \beta \):

\[ C(\beta) = \sum_{j=1}^{p} C_j(\beta, \hat{\gamma}_j(\beta, W_j(\beta))) = \sum_{j=1}^{p} \frac{\hat{\gamma}_j(\beta, W_j(\beta)) \cdot (\hat{\Gamma}_j - \beta \hat{\gamma}_j)}{\beta^2 \sigma_{X_j}^2 + \sigma_{Y_j}^2}, \]  

(3)

where \( \hat{\gamma}_j(\beta, W_j(\beta)) \) is any estimator of \( \hat{\gamma}_j \) that only depends on the sufficient statistic \( W_j(\beta) \), not necessarily the MLE. It is obvious that the estimating function is always unbiased, i.e. \( \mathbb{E}[C(\beta)] = 0 \) at the true value of \( \beta \), regardless of what form of \( \hat{\gamma}_j \) is used. The estimator \( \hat{\beta} \) is obtained by solving \( C(\beta) = 0 \).

The profile score approach [28] can be viewed as a special case of the conditional score, where the “weights” are the MLE of \( \gamma_j \): \( \hat{\gamma}_j(\beta, W_j(\beta)) = \hat{\gamma}_{j, \text{MLE}}(\beta) \). This is also equivalent to using a flat prior in the partially Bayes approach that will be explained shortly. Under regularity conditions, we prove in our previous article [28] that the profile score provides a consistent and asymptotically normal estimator of \( \beta \). However, this estimator is not efficient in general. To see this, let’s assume most \( \gamma_j \) are equal to 0. Intuitively, the \( j \)-th IV provides no information on \( \beta \) because the distribution of \( (\hat{\gamma}_j, \hat{\Gamma}_j) \) does not depend on \( \beta \). However, some information is still used to estimate \( \gamma_j \) in the MLE, resulting in a loss of statistical efficiency. This phenomenon is particularly relevant in genome-wide MR as most of the genetic instruments are very weak.

When \( \sigma_{X_j}^2 \) and \( \sigma_{Y_j}^2 \) are equal across \( j \), Lindsay [34] points out that the efficient estimator of \( \beta \) is given by the weight \( \hat{\gamma}_j^* = \mathbb{E}_{\pi^*}[\gamma_j | W_j(\beta)] \), where \( \pi^* \) is the empirical distribution of \( \gamma \). However, since \( \gamma \) and hence \( \pi^* \) is unknown, it is impossible to compute \( \hat{\gamma}_j^* \) directly. Lindsay proposes to use the empirical Bayes (EB) estimator of \( \gamma \). Suppose the distribution of \( \gamma \) is modeled by a parametric family \( \pi_\eta \) and \( \hat{\eta} \) is an estimate of \( \eta \) using the observed data. We can use

\[ \hat{\gamma}_{j, \text{EB}}(\beta, W_j(\beta)) = \mathbb{E}_{\pi_\hat{\eta}}[\gamma_j | W_j(\beta)] \]  

(4)
in the estimating function (3). Since this is usually a better estimator of the whole vector \( \gamma \) than the MLE, a phenomenon known as the James-Stein paradox [57, 58], it is natural to expect that the resulting function of \( \beta \) is also more efficient than the profile score. In fact, Lindsay [34] shows that the estimator has a local efficiency property: when the parametric distribution \( \pi_\eta \) is specified correctly, the estimator \( \hat{\beta} \) is asymptotically efficient; when \( \pi_\eta \) is specified incorrectly, the estimator is not efficient but still consistent.

**A.2 Implementation details**

**A.2.1 Spike-and-slab prior**

Model (4.2) implies that \( \hat{\gamma}_j/\sigma_{Xj} \) also follows a Gaussian mixture distribution marginally:

\[
\hat{\gamma}_j/\sigma_{Xj} \overset{i.i.d.}{\sim} p_1 \cdot N(0, \sigma_1^2 + 1) + (1 - p_1) \cdot N(0, \sigma_2^2 + 1), \; j = 1, \ldots, p.
\] (5)

In practice we use maximum likelihood to estimate the prior parameters \((p_1, \sigma_1^2, \sigma_2^2)\) by fitting the marginal mixture model (5) to the exposure \(z\)-statistics \(\hat{\gamma}_j/\sigma_{Xj}, \; j = 1, \ldots, p\).

The posterior mean of \(\gamma_i/\sigma_{Xi} \) can be computed using the formulas in Proposition 1.

**Proposition 1.** Suppose \(Z \sim N(\gamma, \sigma^2)\), \( \gamma \sim p_1 N(\mu_1, \sigma_1^2) + (1 - p_1) N(\mu_2, \sigma_2^2)\), then \(\gamma|Z \sim \tilde{p} \cdot N(\tilde{\mu}_1, \tilde{\sigma}_1^2) + (1 - \tilde{p}) \cdot N(\tilde{\mu}_2, \tilde{\sigma}_2^2)\), where

\[
\tilde{\mu}_k = \frac{Z/\sigma^2 + \mu_k/\sigma_k^2}{1/\sigma^2 + 1/\sigma_k^2}, \quad \tilde{\sigma}_k^2 = \frac{1}{1/\sigma^2 + 1/\sigma_k^2}, \quad \text{and} \quad \tilde{p} = \frac{p_1 \cdot \varphi(Z; \mu_1, \sigma^2 + \sigma_1^2)}{p_1 \cdot \varphi(Z; \mu_1, \sigma^2 + \sigma_1^2) + (1 - p_1) \cdot \varphi(Z; \mu_2, \sigma^2 + \sigma_2^2)}.
\]

In the above equation, \(\varphi(z; \mu, \sigma^2)\) is the probability density function of the normal distribution \(N(\mu, \sigma^2)\): \(\varphi(z; \mu, \sigma^2) = (\sqrt{2\pi\sigma^2})^{-1} \exp\{- (z - \mu)^2 / (2\sigma^2)\}\). The posterior mean of \(\gamma\) is given by \(\hat{\gamma} = \mathbb{E}[\gamma|Z] = \tilde{p}\tilde{\mu}_1 + (1 - \tilde{p})\tilde{\mu}_2\).

We want to make two remarks about the choice of prior distribution. First, there is an attractive property of setting the means to be 0 in (4.2). Using Proposition 1, it is easy to verify that, when \(\mu_1 = \mu_2 = 0\), \(\mathbb{E}[\gamma|Z] = -\mathbb{E}[\gamma| - Z]\). As a consequence, the estimating functions in (4.3) are invariant to allele-recoding, meaning if a pair of observations \((\hat{\gamma}_j, \hat{\Gamma}_j)\) is replaced by \((-\hat{\gamma}_j, -\hat{\Gamma}_j)\), the point estimate \(\hat{\beta}\) is unchanged. This is desirable because the allele coding used in a GWAS is often arbitrary. The second remark is that the spike-and-slab implementation is actually quite
important in order to gain efficiency. To see this, suppose a single Gaussian prior is used (as in the empirical Bayes interpretation of the James-Stein estimator). It is easy to show that every SNP then receives the same amount of multiplicative shrinkage, so the first estimating function in (4.3) is just scaled by a constant. As a consequence, the estimator $\hat{\beta}$ is the same no matter how large the shrinkage is. By using a spike-and-slab prior, every genetic instrument is shrunk selectively according to its strength and thus efficiency might be gained. It is then natural to expect that the efficiency gain is more substantial when the two components are more distant ($\sigma_1$ and $\sigma_2$ are more different). See Section 5 for an example.

In Figure 3 we examine the fit of the Gaussian mixture model (5) for our primary analysis of HDL-c in Section 6. In this example we selected 1122 SNPs and the estimated prior parameters are $p_1 = 0.91$, $\sigma_1 = 0.73$, $\sigma_2 = 4.57$. In the left panel of Figure 3 we compare the empirical distribution of $\hat{\gamma}_j/\sigma_{Xj}$ (black histogram) with the fitted Gaussian mixture distribution in (5) (red curve). We find the empirical fit is quite good. In the right panel of Figure 3 we plot the empirical Bayes shrinkage estimator as a function of the $z$-score. When the $z$-score is close to 0, it is shrunk aggressively towards 0; when the $z$-score is large (e.g. greater than 5), there is essentially no shrinkage.

Figure 3: Examine the fitted prior distribution for HDL-c using 1122 restricted SNPs used in Table 5b. The left panel compares the empirical distribution of $z_j = \hat{\gamma}_j/\sigma_{Xj}$ (black histogram) with the fitted Gaussian mixture distribution in (5) (red curve). The right panel shows the posterior mean as a function of the $z$-score.
A.2.2 Choice of robust score function

In our empirical analysis we will consider two choices of the function $\psi(\cdot)$. The first is the identity function $\psi_I(t) = t$ which is non-robust, and the second is the Huber score function \[59\] that is robust:

$$
\psi_H(t) = \begin{cases} 
  t, & \text{if } |t| \leq k, \\
  k \cdot \text{sign}(t), & \text{otherwise.}
\end{cases}
$$

The tuning constant $k$ is chosen to be 1.345, which corresponds to 95% asymptotic efficiency for normal samples in the standard location problem. Another common choice of the robust score function is Tukey’s biweight \[60\], where large outliers essentially have no influence on the estimator. In practice we find that using Tukey’s biweight usually gives more local roots than Huber’s score. Thus we only report results of the more stable Huber’s score in the application.

A.2.3 Multiple roots of the estimating equations

In practice, the estimating functions in \[4.3\] may have multiple roots. Some roots are trivial: it is straightforward to show that $\tilde{C}(\beta, \tau^2) \to 0$ if $\beta \to \pm \infty$ or $\tau^2 \pm \infty$. These unbounded roots can be easily ruled out. However, often there are multiple finite roots. When this happens, we report the root whose $\hat{\beta}$ is closest to the profile-likelihood estimator of $\beta$ assuming all the SNPs are valid IVs \[28\]. The latter is always unique because it solves an optimization problem. When there is another root that is also close to the profile-likelihood estimator (the criterion we use in the application is the absolute difference is no more than 5 times the closest difference), we report the empirical partially Bayes estimator is not available.

A.2.4 Standard error of the estimator

Our final problem is to compute the standard error of the estimator $\hat{\theta} = (\hat{\beta}, \hat{\tau}^2)$. For the calculation below we assume all SNPs satisfy $\alpha_j \sim N(0, \tau^2)$, i.e. there is no outlier. After taking a first-order Taylor expansion at the true value of $\theta = (\beta, \tau^2)$,

$$
0 = \tilde{C}(\hat{\theta}) \approx \tilde{C}(\theta) + \nabla \tilde{C}(\theta) \cdot (\hat{\theta} - \theta),
$$

29
the variance of $\hat{\theta}$ can be approximated using the Delta method by

$$\text{Var}(\hat{\theta}) \approx \left[ \nabla \tilde{C}(\theta) \right]^{-1} \text{Var} \left( \tilde{C}(\theta) \right) \left[ \nabla \tilde{C}(\theta) \right]^{-T}. $$

By repeatedly using the fact that $\hat{\gamma}_{j,EB}$ is independent of $t_j(\beta, \tau^2)$ (so several terms have mean 0 and are dismissed), we obtain, after some algebra, that

$$\text{Var} \left( \tilde{C}(\theta) \right) \approx \sum_{j=1}^{p} \begin{pmatrix} c_1 \hat{x}_{j}^2 / s_j^2 & 0 \\ 0 & c_2 \hat{x}_{j}^2 / s_j^4 \end{pmatrix}, \quad \text{and}$$

$$\nabla \tilde{C}(\theta) \approx \sum_{j=1}^{p} \begin{pmatrix} \psi_1(t_j) \cdot (\partial / \partial \beta) \hat{\gamma}_{j} + \hat{\gamma}_{j} \psi'_1(t_j) \cdot (\partial / \partial \beta) t_j / s_j & \psi_1(t_j) \cdot (\partial / \partial \tau^2) \hat{\gamma}_{j} / s_j \\ 0 & (\delta + c_3) / (2s_j^4) \end{pmatrix}. \quad (6)$$

The subscript EB and the dependence of $s_j$ and $t_j$ on $\theta$ are suppressed to simplify the expressions.

The constants appeared in (6) are $\delta = \mathbb{E}[\psi_2(Z)]$, $c_1 = \mathbb{E}[\psi_1^2(Z)]$, $c_2 = \text{Var}(\psi_2(Z))$, $c_3 = \mathbb{E}[Z \psi'(Z) - \psi(Z)]$ for $Z \sim \mathcal{N}(0,1)$. Assuming $\hat{\theta}$ is a good estimator of $\theta$, the matrices in (6) can be estimated by replacing $\theta$ with $\hat{\theta}$.

### A.2.5 Diagnostics

To check the modeling assumptions we propose to use a scatterplot of standardized residuals $t_j(\hat{\beta}, \hat{\tau}^2)$ versus the empirical Bayes estimates $\hat{\gamma}_{j,EB}(\hat{\beta}, \hat{\tau}^2)$, $j = 1, \ldots, p$. Under our modeling assumptions, if $(\hat{\beta}, \hat{\tau}^2)$ is close to the true value, most of $t_j(\hat{\beta}, \hat{\tau}^2)$ should be independent of $\hat{\gamma}_{j,EB}(\hat{\beta}, \hat{\tau}^2)$ and distributed like a standard normal. We can verify this implication by computing a smoothing spline of the scatter-plot and check it is different from the x-axis (constant 0). More specifically, we run a linear regression with B-splines of $t_j(\hat{\beta}, \hat{\tau}^2)$ with degrees of freedom $[p/50]$ and report the $F$-test result as “heterogeneity p-value”. We also use the Q-Q plot of $t_j(\hat{\beta}, \hat{\tau}^2)$ against standard normal to check if there is excessive pleiotropy that could not be explained by the normal random effects model.

### B Secondary analysis

In our secondary analysis we use an interim release of the UK BioBank self-reported myocardial infarction GWAS [61]. Notice that in this dataset the genetic associations were reported using a
linear regression model instead of the usual logistic regression model. Thus a directly MR analysis effectively estimates the causal risk difference. In order to compare with the results of the primary analysis, we transformed the estimates to the log odds ratio scale using a crude approximation formula (the overall prevalence of self-reported myocardial infarction is \(7735/337159 = 2.3\%\)).

The harmful effect of LDL-c and protective effect of HDL-c replicate in the secondary analysis (Tables 5a and 5b). However, the analysis for TG (Table 5c) suggest its effect on CAD is significantly positive, and the magnitude of the effect is similar to that of LDL-c. This is drastically different from the null finding regarding TG in our primary analysis.

Different from the primary analysis, the diagnostic plots of the secondary analysis show that the heterogeneity between weak and strong instruments is quite significant for HDL-c (Figure 4b). This can be mostly attributed to the four strongest instruments which all have positive standardized residuals between 2 and 4. Further investigation is required to understand the biological mechanism of this heterogeneity.
Table 5: Secondary analysis for self-reported myocardial infarction. Bolded cells correspond to the recommended analysis: \( \tau^2 \neq 0 \), robust Huber score function \( \psi_H \), and empirical partially Bayes shrinkage. The numbers reported in this table can be interpreted as causal log odds ratios. When there were multiple roots to the estimating equation, the results were reported as NA (not available).
Figure 4: Diagnostic plots in the secondary analysis. If our modeling assumptions are satisfied, most of the standardized residuals (y-axis) should be approximately independent of the instrument weight (x-axis), checked by the scatter-plot in the left panel, and roughly follow the standard normal distribution, checked by the quantile-quantile plot in the right panel.