# Constrained Maximum Likelihood-Based Mendelian Randomization Robust to both Correlated and Uncorrelated Pleiotropic Effects

Haoran Xue[1,2], Xiaotong Shen[1], and Wei Pan[2,3]

[1]School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455.
[2]Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota 55455.
[3]Corresponding author. Email: panxx014@umn.edu. Phone: 612-624-4655. Fax: 612-626-0660.

January 9, 2021; revised April 3, 2021 and May 1, 2021

**Abstract**

With the increasing availability of large-scale GWAS summary data on various complex traits and diseases, there have been tremendous interests in applications of Mendelian randomization (MR) to investigate causal relationships between pairs of traits using SNPs as instrumental variables (IVs) based on observational data. In spite of the potential significance of such applications, the validity of their causal conclusions critically depend on some strong modeling assumptions required by MR, which may be violated due to the widespread (horizontal) pleiotropy. Although many MR methods have been proposed recently to relax the assumptions by mainly dealing with uncorrelated pleiotropy, only few can handle correlated pleiotropy, in which some SNPs/IVs may be associated with hidden confounders, such as some heritable factors shared by both traits. Here we propose a simple and effective approach based on constrained maximum likelihood and model averaging, called cML-MA, applicable to GWAS summary data. To deal with more challenging situations with many invalid IVs with only weak pleiotropic effects, we modify and improve it with data perturbation. Extensive simulations demonstrated that the proposed methods could control the type I error rate better while achieving higher power than other competitors. Applications to 48 risk factor-disease pairs based on large-scale GWAS summary data of three cardio-metabolic diseases (coronary artery disease, stroke and type 2 diabetes), asthma and 12 risk factors confirmed its superior performance.

## 1 Introduction

Mendelian randomization (MR) has been widely applied to assess causal relationships between pairs of complex traits (called exposures and outcomes respectively) using genetic variants as instrumental variables (IVs) for observational data. The practice is not only motivated by fundamental scientific questions on causal relationships, but also largely facilitated by recent

advances in human genetics with increasing availability of large-scale GWAS summary data on various complex traits and of the simplicity of such analyses. However, the validity of such an analysis critically depends on the IV assumptions, which are often violated due to the widespread genetic (horizontal) pleiotropy, leading to biased inference and false conclusions [1, 2]. The three assumptions on a valid IV, as shown in Figures 1, are listed below:

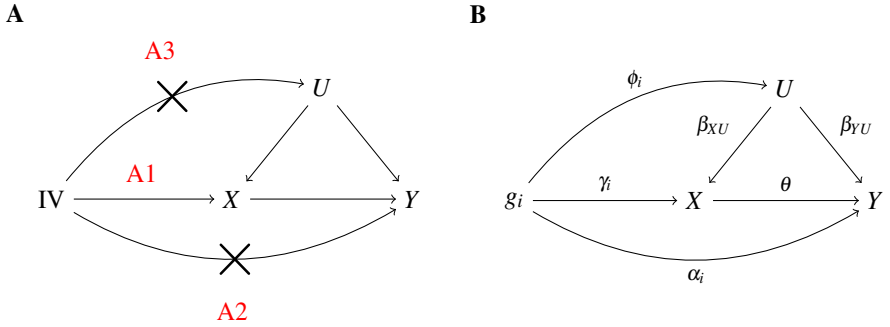
(A1). The IV is associated with the exposure $X$; i.e. $\gamma_i \neq 0$.

(A2). The IV is not associated with the outcome $Y$ conditional on the exposure $X$; i.e. $\alpha_i = 0$.

(A3). The IV is not associated with unmeasured confounder $U$; i.e. $\phi_i = 0$.

Among the three, the first assumption is more straightforward to handle by using highly significant SNPs as IVs. The violation of Assumption A2 introduces so-called uncorrelated pleiotropic effects, for which some MR methods, such as MR Egger regression [3, 4], have been proposed and applied. The most challenging is Assumption A3, introducing so-called correlated pleiotropic effects. Under our general causal model as shown in Figure 1, the total effects of IV $g_i$ on the outcome $Y$ can be decomposed into two parts: that mediated through exposure $X$, $\beta_{Yi,M} = \theta \cdot (\gamma_i + \beta_{XU} \cdot \phi_i)$, and other direct effect (not mediated through $X$), $\beta_{Yi,D} = \alpha_i + \beta_{YU} \cdot \phi_i$ It is clear that, under the violation of Assumption A3, we have $\phi_i \neq 0$, leading to the correlation of the mediated and direct effects $\beta_{Yi,M}$ and $\beta_{Yi,D}$, and thus the violation of the Instrument Strength Independent of Direct Effect (InSIDE) assumption required by MR-Egger regression and other methods (e.g. RAPS) [5, 6] that model the direct effects $\alpha_i$'s as independent random effects; in turn, these methods have to impose that the pleiotropic effect of any SNP must be uncorrelated with its SNP-exposure association.

Figure 1: Causal model with exposure $X$ and outcome $Y$. (A) Three IV assumptions. (B) A general causal model.




Here we propose a simple MR method based on constrained maximum likelihood and model averaging, denoted cML-MA, that is robust to the violation of both Assumptions A2 and A3; i.e., it is robust to invalid IVs with uncorrelated or correlated pleiotropic effects. Table 1 compares our proposed cML-MA with some most popular and new methods that are all applicable to GWAS summary data. Our method only depends on the "plurality valid" assumption: in large samples, while (Wald) ratio estimates of the target causal effect from invalid IVs will take different values, ratio estimates from all valid IVs should approach the true causal effect and thus the valid IVs form the largest group of SNPs among all the groups giving different ratio estimates [7, 8]. This assumption is weaker than the "majority valid" assumption, which states

that more than 50% of the SNPs being used are valid IVs. It is noted that three new methods with this or other weak assumptions, MR-Mix, MR-ContMix and CAUSE, impose a normal mixture model with more unknown parameters to estimate, while our proposed cML-MA does not impose such an assumption and estimates only a minimum number of necessary parameters. It is well known that mixture models are statistically difficult to estimate with small numbers of SNPs/IVs. This brings up an important point: in addition to its modeling assumptions, another key factor determining the performance of a method is how it is implemented. This point might explain why our cML-MA performed better than MR-Mix, MR-ContMix and CAUSE as to be shown. Another example is the MR-Weighted-Mode method: although it imposes minimum modeling assumptions as cML-MA, it is difficult to estimate the mode of a distribution with small numbers of SNPs/IVs, leading to often its poor performance o as shown by others [9, 10] and to be confirmed later too. A similar argument in the Discussion section will be made on the advantage of our method over two other related ones, MR-Lasso and MR-PRESSO: all three share the basic idea of selecting/removing invalid IVs, but due to their different implementations, they perform quite differently. As shown by previous studies [9, 10, 11], existing MR methods may not perform well with inflated type I errors and/or biased estimates, especially with a high proportion of invalid IVs among a small number of SNPs, prompting an urgent need for more robust and efficient MR methods. Here we develop an efficient algorithm, an effective model selection criterion, a model averaging approach and its variant based on data perturbation for such a purpose; our proposed cML-MA and its data perturbation-based variant are simple and easy to implement while imposing less stringent modeling assumptions, and as to be shown, perform consistently better than other methods across a wide range of scenarios.

Table 1: Comparison of different MR methods, including whether valid IV Assumption A2 or A3 can be violated. The notations are defined in Fig 1 and Eq (1), and $q$ is the (unknown) proportion of invalid IVs while $\hat{\theta}_i = \hat{\beta}_{Yi}/\hat{\beta}_{Xi}$ and $SE(\hat{\theta}_i)$ are the Wald ratio estimate of $\theta$ based on SNP $i$ and its standard error respectively. NOME refers to NO Measurement Error assumption: the variance of any IV-exposure association estimate is negligible [12].

| Method | A2 | A3 | Key assumptions | Implementation challenges/performance |
|---|---|---|---|---|
| cML-MA | ✔ | ✔ | Plurality valid | Controlling type-I errors with high power. |
| MR-Mix [13] | ✔ | ✔ | Plurality valid; $\hat{\beta}_{Yi} - \theta\hat{\beta}_{Xi} \sim$ a mixture of Normals | Biased to the null, thus conservative. |
| MR-ContMix [9] | ✔ | ✔ | Plurality valid; $\hat{\theta}_i \sim (1-q)N(\theta, SE(\hat{\theta}_i)^2) + qN(0, \psi^2 + SE(\hat{\theta}_i)^2)$; NOME | Difficult to pre-choose a fixed value for tuning parameter $\psi$. |
| CAUSE [14] | ✔ | ✔ | $< 50\%$ IVs have correlated pleiotropy; $\gamma_i\phi_i = 0$; $\beta_{XU} = 1$; $\beta_{Xi} = \gamma_i$ or $\phi_i$; $\hat{\beta}_{Yi} \sim (1-q)N(\theta\beta_{Xi} + r_i, \sigma^2) + qN((\theta + \beta_{YU})\beta_{Xi} + r_i, \sigma^2)$ | Difficult to estimate some parameters depending on the *hidden* confounder $U$; sensitive to assumption of $\gamma_i\phi_i = 0$. |
| MR-Lasso [15] | ✔ | ✔ | Plurality valid [7]; some condition on the exposure-association strengths of invalid IVs relative to that of valid IVs to ensure consistency [16]; NOME | Depending on the heterogeneity criterion for choosing the tuning parameter for the Lasso penalty. |
| MR-Weighted-Mode [17] | ✔ | ✔ | Plurality valid | Sensitive to the difficult bandwidth selection for mode estimation. |
| MR-Weighted-Median [3] | ✔ | ✔ | Majority valid | Robust to outliers; low powered; sometimes biased. |
| MR-PRESSO [1] | ✔ | ✘ | Majority valid; InSIDE; Good delete-1 causal estimates | Inflated type I errors; unable to completely remove invalid IVs. |
| MR-Egger [18] | ✔ | ✘ | InSIDE: $\phi_i = 0$; $\alpha_i \sim N(\mu, \tau^2)$ for a small $m$ (but no normality needed for a large $m$); NOME | Often biased and low powered. |
| MR-RAPS [6] | ✔ | ✘ | InSIDE: $\phi_i = 0$; $\alpha_i \sim N(0, \tau^2)$ if overdispersion is specified | Maybe sensitive to directional pleiotropy; robust to outliers with Tukey's loss |
| MR-IVW (RE) [19, 20] | ✔ | ✘ | Balanced pleiotropy; NOME | Sensitive to directional pleiotropy; low powered. |
| MR-IVW (FE) [19, 20] | ✘ | ✘ | All IVs are valid; NOME | Efficient when all IVs are valid; sensitive to invalid IVs. |

# 2 Methods

## 2.1 Overview

Suppose that we have $m$ independent SNPs, $g_1, g_2, \cdots, g_m$, as IVs, $X$ is the exposure, $Y$ is the outcome, and $U$ is the hidden confounder. Under the true causal model as shown in Figure 1, we obtain the total effects of $g_i$ on $X$ and on $Y$ as $\beta_{Xi}$ and $\beta_{Yi}$ respectively:

$$\begin{aligned} \beta_{Xi} &= \gamma_i + \beta_{XU} \cdot \phi_i, \\ \beta_{Yi} &= \theta \cdot (\gamma_i + \beta_{XU} \cdot \phi_i) + \alpha_i + \beta_{YU} \cdot \phi_i = \theta \cdot \beta_{Xi} + r_i, \end{aligned} \tag{1}$$

where $r_i = \alpha_i + \beta_{YU} \cdot \phi_i$ represents the direct/pleiotropic effects of IV $g_i$ on outcome $Y$, not mediated by exposure $X$. If $g_i$ is a valid IV, IV Assumptions A2 and A3 would imply $\alpha_i = 0$ and $\phi_i = 0$ respectively, leading to $r_i = 0$. For an invalid IV with $r_i \neq 0$, its (Wald) ratio is biased for $\theta$: $\beta_{Yi}/\beta_{Xi} \neq \theta$. Our goal is for unbiased inference of the causal effect $\theta$ in the possible presence of some (unknown) invalid IVs with the corresponding (unknown) $r_i \neq 0$.

From two independent GWAS summary datasets for traits $X$ and $Y$ respectively, we obtain the estimated marginal effect sizes of $g_i$'s on $X$ and $Y$ (and their standard errors) as $\hat{\beta}_{Xi}(\hat{\sigma}_{Xi})$ and $\hat{\beta}_{Yi}(\hat{\sigma}_{Yi})$, $i = 1, \cdots, m$. Asymptotically (or approximately) we have $\hat{\beta}_{Xi} \sim N(\beta_{Xi}, \hat{\sigma}^2_{Xi})$ and $\hat{\beta}_{Yi} \sim N(\beta_{Yi}, \hat{\sigma}^2_{Yi})$ for $i = 1, \cdots, m$. With equation (1), we have the log-likelihood function (up to a constant) as

$$L(\theta, \{\beta_{Xi}, r_i\}; \{\hat{\beta}_{Xi}, \hat{\beta}_{Yi}, \hat{\sigma}^2_{Xi}, \hat{\sigma}^2_{Yi}\}) = -\frac{1}{2} \sum_{i=1}^{m} \left( \frac{(\hat{\beta}_{Xi} - \beta_{Xi})^2}{\hat{\sigma}^2_{Xi}} + \frac{(\hat{\beta}_{Yi} - \theta \cdot \beta_{Xi} - r_i)^2}{\hat{\sigma}^2_{Yi}} \right), \tag{2}$$

which, for simplicity, may be written as $L(\theta, \{\beta_{Xi}, r_i\})$. Throughout we use $\{\beta_{Xi}, r_i\} = \{(\beta_{Xi}, r_i) : i = 1, 2, ..., m\}$ to represent a set of the parameters, and similarly for $\{\hat{\beta}_{Xi}, \hat{\beta}_{Yi}, \hat{\sigma}^2_{Xi}, \hat{\sigma}^2_{Yi}\}$. Under the constraint that the number of the invalid IVs is $K$, a given integer, we can obtain the constrained maximum likelihood estimate (cMLE) $\hat{\theta}(K)$ and its standard error $\mathrm{SE}(\hat{\theta}(K))$. We prove that with correctly selected valid IVs, the cMLE is consistent and asymptotically Normal. Accordingly we can construct Normal-based confidence intervals (CIs) or conduct significance testing for $\theta$. In practice, since $K$ is unknown, we propose a Bayesian information criterion (BIC) to select $K$ consistently before drawing inference on the true causal effect $\theta$; this is our proposed cML-BIC. For finite (especially small) sample sizes, due to model selection errors, such a procedure might have slightly inflated type I error rates as shown in the Supplementary. Instead, to account for model selection uncertainty, we propose a model averaging approach [21], called **cML-MA-BIC**, or **cML-MA** for short. We obtain multiple estimates $\hat{\theta}(K)$, each based on a selected model with each value of $0 \leq K < m - 1$, then take their weighted average as our final estimate of the causal effect and draw inference accordingly; the weights are determined by the BIC values of the models, in which those more likely models (with lower BIC values) are given higher weights. More details are given below.

## 2.2 Estimation and Selection Consistency with the cMLE

We develop some (asymptotic) theory to support our proposed method (for fixed $m$ as $n$ increases). Denote $B_0 = \{i | r_i \neq 0, i = 1, \cdots, m\}$ the set of truly invalid IVs, and its size $|B_0| = K_0$.

Suppose that $N_1$ and $N_2$ are the sample sizes of the two GWAS summary datasets for $X$ and $Y$ respectively. With (10), we obtain the cMLEs by solving

$$\min_{\theta, b_{Xi}, r_i, 1\le i\le m} -L(\theta, \{b_{Xi}, r_i\}; \{\hat{\beta}_{Xi}, \hat{\beta}_{Yi}, \hat{\sigma}^2_{Xi}, \hat{\sigma}^2_{Yi}\}) \quad \text{subject to} \quad \sum_{i=1}^{m} I(r_i \neq 0) = K. \tag{3}$$

Here $I(\cdot)$ is the indicator function, $K$ is a tuning parameter representing the unknown number of invalid IVs. Denote the cMLEs from (3) as $\hat{\theta}(K)$, $\hat{b}_{Xi}(K)$, and $\hat{r}_i(K)$ for $i = 1, \cdots, m$, and $\hat{B}_K = \{i | \hat{r}_i(K) \neq 0, i = 1, \cdots, m\}$ the estimated set of invalid IVs. We propose a Bayesian information criterion (BIC) based on GWAS summary data to select the best $K$ in a candidate set $\mathscr{K}$:

$$\text{BIC}(K) = -2L\left(\hat{\theta}(K), \{\hat{b}_{Xi}(K), \hat{r}_i(K)\}\right) + \log(N) \cdot K. \tag{4}$$

Here $N$ could be either $N_1$ or $N_2$; we recommend using $N = \min(N_1, N_2)$. We select $\hat{K} = \arg\min_{K\in\mathscr{K}} BIC(K)$, and estimate $\hat{B}_{\hat{K}} = \{i : \hat{r}_i(\hat{K}) \neq 0\}$ as the set of invalid IVs (and its complement $\hat{B}^c_{\hat{K}} = \{1, \cdots, m\} - \hat{B}_{\hat{K}}$ as the set of valid IVs).

Now we state two assumptions used to prove the selection consistency of our proposed BIC.

**Assumption 1.** *(Plurality valid condition.) Suppose that $B_0$ is the index set of the true invalid IVs with $K_0 = |B_0|$. For any $B \subseteq \{1, \cdots, m\}$ and $|B| = K_0$, if $B \neq B_0$, then the $(m - K_0)$ ratios $\{r_i/\beta_{Xi}, i \in B^c\}$ are not all equal.*

**Assumption 2.** *(Orders of the variances and sample sizes.) There exist positive constants $l_X, l_Y, l_N$ and $u_X, u_Y, u_N$ such that we have $l_X/N_1 \le \hat{\sigma}^2_{Xi} \le u_X/N_1$, $l_Y/N_2 \le \hat{\sigma}^2_{Yi} \le u_Y/N_2$, and $l_N \cdot N_2 \le N_1 \le u_N \cdot N_2$ for $i = 1, \cdots, m$.*

Assumption 2 says that the two sample sizes $N_1$ and $N_2$ are comparable, and that either variance $\hat{\sigma}^2_{Xi}$ or $\hat{\sigma}^2_{Yi}$ is of order $1/N_1$ or $1/N_2$, which is satisfied by the usual least squares or maximum likelihood estimates obtained from GWAS summary data. We also note that the sample sizes for the SNPs/IVs being used may vary; as long as they are comparable (in the sense as defined in Assumption 2), we can take their minimum in each GWAS dataset as the corresponding sample size $N_1$ or $N_2$. With Assumptions 1 and 2 we prove (in the Supplementary) that our proposed BIC consistently selects invalid IVs.

**Theorem 1.** *With Assumptions 1 and 2 satisfied, if $K_0 \in \mathscr{K}$, we have $P(\hat{K} = K_0) \to 1$ and $P(\hat{B}_{\hat{K}} = B_0) \to 1$ as $N_1, N_2 \to \infty$.*

As shown in the Supplementary, it was confirmed that in the simulations the proposed BIC selected increasing proportions of the correct models as the sample size increased.

After correctly selecting (and implictly removing) invalid IVs, our proposed cMLE of $\theta$ is the same as the maximum profile likelihood estimate (MPLE) being applied to *all* valid IVs [6]. Applying Theorems 3.1 and 3.2 in [6], coupled with the above selection consistency, we obtain both the estimation consistency and asymptotic Normality of the cMLE $\hat{\theta}(\hat{K})$. It is proven in the Supplementary that the variance of the cMLE (based on the Fisher information matrix as shown in Section 2.4) and that of the MPLE are asymptotically equal. As confirmed numerically in the Supplementary, our cMLE and the MPLE were essentially the same in both the simulations and real data examples.

## 2.3 Model Selection and Model Averaging Approaches to Inference of $\theta$

After selecting $\hat{K}$, we can use the cMLE $\hat{\theta}(\hat{K})$ and its SE (see below for how to obtain it) to infer $\theta$: based on the asymptotic normal distribution $\hat{\theta}(\hat{K}) \sim N(\theta, SE^2\left(\hat{\theta}(\hat{K})\right)$, we either construct a confidence interval (CI) or conduct significance test. We call this method **cML-BIC**.

In spite of the selection consistency of our proposed BIC, to account for model selection uncertainties, especially with small sample sizes, we propose a model averaging (MA) approach. Following [21], we first obtain the estimate of $\theta$ and its standard error for each candidate $K \in \mathscr{K}$, then take their weighted average as the final estimate of $\theta$ with the weights determined by the BIC values of the corresponding candidate models.

Following [21], for a set $\mathscr{K}$ of $K$'s, we define the initial and standard weights

$$w_K^0 = \exp\left(-BIC\left(K\right)/2\right), \qquad w_K = w_K^0 / \sum_{K \in \mathscr{K}} w_K^0.$$

The final weighted estimate $\hat{\theta}_w$ and its standard error are

$$\hat{\theta}_w = \sum_{K \in \mathscr{K}} w_K \cdot \hat{\theta}(K), \qquad \mathrm{SE}(\hat{\theta}_w) = \sum_{K \in \mathscr{K}} w_K \cdot \sqrt{\mathrm{SE}\left(\hat{\theta}(K)\right)^2 + \left(\hat{\theta}(K) - \hat{\theta}_w\right)^2}. \tag{5}$$

With $\hat{\theta}_w$ and $\mathrm{SE}(\hat{\theta}_w)$, based on the asymptotic normal distribution, we draw inference on $\theta$. We call this method **cML-MA-BIC**. In practice, we use the set of candidate $K$'s, $\mathscr{K} = \{0, 1, \cdots, m-2\}$. $K = 0$ means all IVs are valid; according to Assumption 1, there should be at least two valid IVs. Other choices of candidate sets of $K$'s could also be applied, especially when we roughly know the proportion of invalid IVs.

Instead of using BIC, we can also use the corresponding Akaike information criterion (AIC) to select $K$ or weight its corresponding model, leading to cML-AIC and cML-MA-AIC for model selection- and model-averaging-based approaches respectively. As shown in the Supplementary, they did not perform as well as their BIC versions.

## 2.4 Computation

We propose a coordinate descent-like algorithm to iteratively solve (3) to obtain cMLEs, $\hat{\theta}(K)$, $\hat{b}_{Xi}(K)$ and $\hat{r}_i(K)$ for $i = 1, \cdots m$. We start with the initial values $\theta^{(0)}$ and $b_{Xi}^{(0)}$'s, then update them iteratively as below until convergence: at the $k^{th}$ iteration,

- Step 1: Given $\theta^{(k)}, b_{Xi}^{(k)}$, update $r_i$. Order $\frac{(\hat{\beta}_{Yi} - \theta^{(k)} b_{Xi}^{(k)})^2}{\hat{\sigma}_{Yi}^2}, i = 1, \cdots, m$, decreasingly, as $\frac{(\hat{\beta}_{Y(1)} - \theta^{(k)} b_{X(1)}^{(k)})^2}{\hat{\sigma}_{Y(1)}^2}, \frac{(\hat{\beta}_{Y(2)} - \theta^{(k)} b_{X(2)}^{(k)})^2}{\hat{\sigma}_{Y(2)}^2}, \cdots, \frac{(\hat{\beta}_{Y(m)} - \theta^{(k)} b_{X(m)}^{(k)})^2}{\hat{\sigma}_{Y(m)}^2}$. Then for $i = 1, \cdots, K$, let $r_{(i)}^{(k+1)} = \hat{\beta}_{Y(i)} - \theta^{(k)} b_{X(i)}^{(k)}$; for $i = K+1, \cdots, m$, let $r_{(i)}^{(k+1)} = 0$.

- Step 2: Given $\theta^{(k)}$, $r_i^{(k+1)}$'s, update $b_{Xi}$ as

$$b_{Xi}^{(k+1)} = \frac{\frac{\hat{\beta}_{Xi}}{\hat{\sigma}_{Xi}^2} + \frac{\theta^{(k)}(\hat{\beta}_{Yi} - r_i^{(k+1)})}{\hat{\sigma}_{Yi}^2}}{\frac{1}{\hat{\sigma}_{Xi}^2} + \frac{(\theta^{(k)})^2}{\hat{\sigma}_{Yi}^2}}. \tag{6}$$

- Step 3: Given $b_{Xi}^{(k+1)}$'s, $r_i^{(k+1)}$'s, update $\theta$ as

$$\theta^{(k+1)} = \frac{\sum_{i=1}^{m} \frac{(\hat{\beta}_{Yi} - r_i^{(k+1)}) b_{Xi}^{(k+1)}}{\hat{\sigma}_{Yi}^2}}{\sum_{i=1}^{m} \frac{(b_{Xi}^{(k+1)})^2}{\hat{\sigma}_{Yi}^2}}. \tag{7}$$

It is noted that at the convergence, by the expression of $r_i$ in Step 1 and that for $\hat{\theta}$ in Step 3, if $r_i \neq 0$ (i.e. for an invalid IV), SNP $i$ and its data do not contribute to estimating $\theta$.

By default, as in all our simulations, we set $\theta^{(0)}$ and $b_{Xi}^{(0)}$'s all at 0. More generally, as in our main real data examples, we can also use multiple random starts; in our real data examples, in addition to the above default starting values, we tried 100 random starts, each randomly generating $\theta^{(0)} \sim \text{Uniform}(\min_{1\leq i\leq m} \hat{\beta}_{Yi}/\hat{\beta}_{Xi}, \max_{1\leq i\leq m} \hat{\beta}_{Yi}/\hat{\beta}_{Xi})$, and $b_{Xi}^{(0)} \sim N(\hat{\beta}_{Xi}, \hat{\sigma}_{Xi}^2)$ for $1 \leq i \leq m$. Then we take the cMLE as the one from the initial values giving the maximum likelihood among those multiple starts. As shown in Supplementary Tables S5 and S6, for our real data examples, among all 48 risk factor-disease pairs, only for 7 pairs the 101 starts gave slightly different results from using the default starting values; the differences in the numbers of detected invalid IVs were only 1 or 2, leading to almost the same results at the end. In our secondary real data analysis, for each of 63 null pairs we tried 10 random starts.

Next we estimate the standard error of $\hat{\theta} = \hat{\theta}(K)$ for any given $K$. Denote the set of the indices of $K$ non-zero $\hat{r}_i$'s as $\hat{B}_K$, the $(m-K+1)$ by $(m-K+1)$ Fisher information matrix is

$$\mathscr{I} = \begin{pmatrix} \frac{\partial^2(-l)}{\partial\theta^2} & \frac{\partial^2(-l)}{\partial\theta\partial\beta_{XB}'} \\ \frac{\partial^2(-l)}{\partial\theta\partial\beta_{XB}} & \frac{\partial^2(-l)}{\partial\beta_{XB}\partial\beta_{XB}'} \end{pmatrix}, \tag{8}$$

where $\beta_{XB}$ is a vector of elements $\beta_{Xi}$ with $i \in \hat{B}_K^c$. Plugging $\hat{\theta}$, $\hat{b}_{Xi}$'s into $\mathscr{I}$, we obtain the standard error of $\hat{\theta}$ as $\text{SE}(\hat{\theta}) = \sqrt{(\mathscr{I}^{-1})_{11}}$. Details are shown in the Supplementary.

## 2.5 Data Perturbation

When the sample sizes of GWAS summary data are relatively small and there are many invalid IVs with weak pleiotropic effects, the (asymptotic) selection consistency of cML-BIC as described in Theorem 1 may not be achieved, leading to missing some invalid IVs and ultimately biased inference, such as inflated type-I errors. To alleviate this problem, we propose using data perturbation (DP) [22]. For $t = 1, \cdots, T$, we generate independent perturbed samples $\hat{\beta}_{Xi}^{(t)} \sim N(\hat{\beta}_{Xi}, \hat{\sigma}_{Xi}^2)$ and $\hat{\beta}_{Yi}^{(t)} \sim N(\hat{\beta}_{Yi}, \hat{\sigma}_{Yi}^2)$ for $i = 1, \cdots m$. Then similar to (3), we obtain the cMLEs with perturbed data by solving

$$\min_{\theta, b_{Xi}, r_i, 1\leq i\leq m} -L(\theta, \{b_{Xi}, r_i\}; \{\hat{\beta}_{Xi}^{(t)}, \hat{\beta}_{Yi}^{(t)}, \hat{\sigma}_{Xi}^2, \hat{\sigma}_{Yi}^2\}) \quad \text{subject to} \quad \sum_{i=1}^{m} I(r_i \neq 0) = K. \tag{9}$$

Denote the cMLEs from (9) as $\hat{\theta}^{(t)}(K)$, $\hat{b}_{Xi}^{(t)}(K)$, and $\hat{r}_i^{(t)}(K)$ for $i = 1, \cdots, m$, we get the maximized log-likelihood as

$$L^{(t)}(K) = L(\hat{\theta}^{(t)}, \{\hat{b}_{Xi}^{(t)}, \hat{r}_i^{(t)}\}; \{\hat{\beta}_{Xi}^{(t)}, \hat{\beta}_{Yi}^{(t)}, \hat{\sigma}_{Xi}^2, \hat{\sigma}_{Yi}^2\}). \tag{10}$$

Averaging over $T$ perturbed estimates, we have

$$\hat{\theta}_{DP}(K) = \frac{\sum_{t=1}^{T} \hat{\theta}^{(t)}(K)}{T}, \quad L_{DP}(K) \quad = \frac{\sum_{t=1}^{T} L^{(t)}(K)}{T}, \tag{11}$$

and estimate the standard error of $\hat{\theta}_{DP}(K)$ as the sample standard deviation of $\hat{\theta}^{(t)}(K)$'s,

$$\text{SE}\left(\hat{\theta}_{DP}(K)\right) = \sqrt{\frac{\sum_{t=1}^{T}\left(\hat{\theta}^{(t)}(K) - \hat{\theta}_{DP}(K)\right)^2}{T-1}}. \tag{12}$$

Then, as for cML-BIC and cML-MA-BIC, with the DP estimates in (11) and (12), we obtain their DP versions called **cML-BIC-DP** and **cML-MA-BIC-DP** respectively.

## 2.6 Goodness-of-fit Tests for the Variance Estimates

In general, as to be shown numerically, cML-MA-BIC-DP is more conservative for inference and thus controls the type I errors better than cML-MA-BIC, but may lose power while being computationally more demanding. To help decide which one to use for a given problem, we develop two goodness-of-fit tests, denoted **GOF1** and **GOF2**, to check whether the (asymptotic) model-based and DP-based variance estimates converge to the same estimate; if so, then we recommend using cML-MA-BIC; otherwise, cML-MA-BIC-DP is preferred.

Suppose that cML-BIC selects a set of $\hat{K}$ invalid IVs, with the estimate $\hat{\theta}(\hat{K})$ and its model-based variance calculated using Fisher Information matrix (8) as $\widehat{\text{Var}}_M\left(\hat{\theta}(\hat{K})\right)$. If the BIC-based model selection is correct with only small model selection uncertainty, we'd expect that $\widehat{\text{Var}}_M\left(\hat{\theta}(\hat{K})\right)$ would be close to the DP-based variance estimate, $\widehat{\text{Var}}_D\left(\hat{\theta}_{DP}(\hat{K})\right)$. Our proposed goodness-of-fit tests aim to test whether the two variance estimates converge to the same estimate (asymptotically).

First, based on each perturbed data set, we obtain $V^{(t)} = \widehat{\text{Var}}_M\left(\hat{\theta}^{(t)}(\hat{K})\right)$, from which we estimate the sample variance of $V^{(t)}$'s as $\widehat{\text{Var}}[\text{Var}_M\left(\hat{\theta}(\hat{K})\right)]$. Second, as shown by (12), $\widehat{\text{Var}}_D\left(\hat{\theta}_{DP}(\hat{K})\right)$ is the sample variance of $T$ i.i.d. random $\hat{\theta}^{(t)}(K), t = 1, \cdots, T$ from some distribution $f(\cdot)$; Theorem 2 in Chapter 6 of [23] shows that its variance is

$$\text{Var}\left[\widehat{\text{Var}}_D\left(\hat{\theta}_{DP}(\hat{K})\right)\right] = \frac{1}{T}\left(\mu_4 - \frac{T-3}{T-1}\sigma^4\right), \tag{13}$$

where $\mu_4$ and $\sigma^2$ are the central fourth moment and variance of $f(\cdot)$. We use the $T$ samples to estimate them as

$$\hat{\mu}_4 = \frac{\sum_{t=1}^{T}\left(\hat{\theta}^{(t)}(\hat{K}) - \hat{\theta}_{DP}(\hat{K})\right)^4}{T}, \quad \hat{\sigma}^2 = \frac{\sum_{t=1}^{T}\left(\hat{\theta}^{(t)}(\hat{K}) - \hat{\theta}_{DP}(\hat{K})\right)^2}{T}.$$

Plug them into (13), we obtain

$$\widehat{\text{Var}}_1\left[\widehat{\text{Var}}_D\left(\hat{\theta}_{DP}(\hat{K})\right)\right] = \frac{1}{T}\left(\hat{\mu}_4 - \frac{T-3}{T-1}\hat{\sigma}^4\right), \tag{14}$$

and the first GOF test statistic is

$$Z_{GOF1} = \frac{\widehat{\text{Var}}_D\left(\hat{\theta}_{DP}(\hat{K})\right) - \widehat{\text{Var}}_M\left(\hat{\theta}(\hat{K})\right)}{\sqrt{\widehat{\text{Var}}_1\left[\widehat{\text{Var}}_D\left(\hat{\theta}_{DP}(\hat{K})\right)\right] + \widehat{\text{Var}}[\widehat{\text{Var}}_M\left(\hat{\theta}(\hat{K})\right)]}}. \tag{15}$$

Comparing $Z_{GOF1}$ with the standard normal random variate $Z$, we can calculate the $p$-value as $P(|Z| > |Z_{GOF1}|)$. This is the first goodness-of-fit test **GOF1**.

When $T$ is small or only moderately large, the estimate $\hat{\mu}_4$ could have a large variance. Furthermore, if $\hat{\theta}^{(t)}(K)$'s are normally distributed, we have $\mu_4 = 3\sigma^4$, and the variance estimate in (16) can be simplified to

$$\widehat{\text{Var}}_2\left[\text{Var}_D\left(\hat{\theta}_{DP}(\hat{K})\right)\right] = \frac{2}{T-1}\hat{\sigma}^4. \tag{16}$$

Replacing $\widehat{\text{Var}}_1$ in $Z_{GOF1}$ with $\widehat{\text{Var}}_2$, we obtain a new GOF test statistic, $Z_{GOF2}$, by which and a standard normal as the null distribution, we can calculate a $p$-value.

## 2.7 Other Methods

We compared cML-MA with other exiting two-sample MR methods, including MR-ContMix, MR-Mix, MR-CAUSE, MR-Lasso, MR-PRESSO, MR-IVW (random-effect (RE) meta-analysis), MR-Egger regression, MR-Weighted-Median, MR-Weighted-Mode, and MR-RAPS (the Robust Adjusted Profile Score) methods. We applied MR-RAPS with four different combinations of its parameters: for MR-RAPS1 and MR-RAPS2, we set the over-dispersion parameter as TRUE, and used the $L_2$ loss and the Tukey loss respectively; for MR-RAPS3 and MR-RAPS3, we set the over-dispersion as FALSE and used the $L_2$ and the Tukey loss respectively; we present RAPS2 to represent RAPS in the main text. We also applied the Oracle MR-IVW to only valid IVs in the simulations, called MR-IVW-Oracle.

Each method takes GWAS summary data of $\hat{\beta}_{Xi}, \hat{\sigma}_{Xi}, \hat{\beta}_{Yi}, \hat{\sigma}_{Yi}, i = 1, \cdots, m$, as input, and gives an estimate of $\theta$, say $\hat{\theta}$, along with its standard error SE($\hat{\theta}$).

## 2.8 GWAS Data

### 2.8.1 Primary Real Data Examples

We applied various methods to some large-scale GWAS summary data. Following [14], we studied possible causal effects of 12 risk factors on 4 complex diseases: coronary artery disease (CAD [MIM: 608320]), stroke (MIM: 601367), type 2 diabetes (T2D [MIM: 125853]) and asthma (MIM: 600807) (mostly as a negative control). These 12 cardio-metabolic risk factors are triglycerides (TG), high-density lipoprotein (HDL), low-density lipoprotein (LDL), drinks per week (Alcohol), ever regular smoker (Smoke), body fat percentage (BF), birth weight (BW), body mass index (BMI), Height (MIM: 606255), fasting glucose (FG), systolic blood pressure (SBP) and diastolic blood pressure (DBP).

For each risk factor/exposure-disease/outcome pair we used the set of LD-independent SNPs as IVs as described in [14] (in their Supplementary Table 4), and applied all methods to the GWAS summary statistics of these SNPs.

### 2.8.2 Secondary Real Data Examples

If two traits are not genetically correlated, it is unlikely that they are causally related. As suggested by a reviewer, from LD Hub [24], we collected 63 trait pairs without significant genetic correlations (i.e. $p$-value $> 0.05$) as negative controls to study the type-I error properties of the methods. These 63 pairs involve 13 traits in total: fasting proinsulin (FP), height, homeostasis model assessment of beta-cell function (HOMA), LDL, rheumatoid arthritis (RA [MIM: 180300]), schizophrenia (SCZ [MIM: 181510]), T2D, age at smoking (ASmk), anorexia nervosa(AN [MIM: 606788]), childhood IQ (CIQ), ever/never smoked (ESmk), former current smoker (FSmk), and infant head circumference (IHC). For each pair, we used R package `TwoSampleMR` to select LD-independent SNPs as IVs and extract their summary statistics following the standard procedures; sample R code is available in the Supplementary. The GWAS summary data in LD Hub and `TwoSampleMR` are the same for 12 traits except for height. For height, LD Hub contains the GWAS data of the GIANT consortium from year 2010 [25], while `TwoSampleMR` uses the GIANT data from year 2014 [26]. Details of the GWAS data used are in the Supplementary.

## 2.9 Simulation Set-ups

### 2.9.1 Main simulations

We compared different methods through extensive simulations. The simulation set-ups were similar to those in [9]. We set the number of the SNPs/IVs $m = 10$, 20 or 100, and sample size $N = 50000$, 100000 or 200000. For each SNP $i = 1, \cdots, m$, we generated $\gamma_i$'s from a uniform distribution on $(-0.2,-0.1) \cup (0.1,0.2)$; its MAF $f_i$ from a uniform distribution Unif(0.1,0.3), then its genotypes $\text{SNP}_{ij}$ from a bonimial Bin(2, $f_i$) for $j = 1, \cdots, N$. For each $m$, we tried different proportions $q = 0\%$, $20\%$, $40\%$, $60\%$ of invalid IVs: for each SNP $i = 1, \cdots, m \cdot q$, we generated its direct effect size $\alpha_i$ from Uniform$(0.2, 0.3)$, and set $\phi_i = 0$ (when the InSIDE assumption was satisfied) or generated $\phi_i$ from Uniform(-0.1,0.1) (when InSIDE was violated). We set $\beta_{XU} = \beta_{YU} = 1$, generated the random errors $\varepsilon_U$, $\varepsilon_X$ and $\varepsilon_Y$ independently from $N(0,1)$. Then we generated $U$, $X$ and $Y$ from the causal model (17):

$$\begin{aligned} U &= \sum_{i=1}^{m} \phi_i \cdot g_i + \varepsilon_U, \\ X &= \sum_{i=1}^{m} \gamma_i \cdot g_i + \beta_{XU} \cdot U + \varepsilon_X, \\ Y &= \theta \cdot X + \sum_{i=1}^{m} \alpha_i \cdot g_i + \beta_{YU} \cdot U + \varepsilon_Y. \end{aligned} \tag{17}$$

We generated two independent samples, each of size $N_1 = N_2 = N$, and used the first sample to fit marginal linear regressions of $X$ on SNP's, and using the second sample to fit marginal linear regressions of $Y$ on SNP's, thus obtaining the GWAS summary statistics. We tried different $\theta = -0.1, -0.05, -0.03, -0.01, 0, 0.01, 0.03, 0.05, 0.1$. When $\theta = 0$ it was the null case; i.e. $X$ had no causal effect on $Y$.

### 2.9.2 Secondary simulations

We did simulations to compare various methods with CAUSE. We generated the simulated data in the framework of CAUSE as described in the original CAUSE paper [14]. We set the sample size $N_1 = N_2 = N = 50000$ or 100000. Denote the direct effects of SNP $i$ on the exposure as $\beta_{Xi}$, and on the outcome $Y$ as $\alpha_i$. For the expected number of SNPs with non-zero associations with the exposure and the outcome (i.e. $\beta_{Xi} \neq 0$ and $\alpha_i \neq 0$), denoted by $m_X$ and $m_\alpha$, we set $m_X = m_\alpha = m = 10$ or 100. Denote the true causal effect size from the exposure to the outcome as $\theta$. We set $\theta = 0$ for the null case and $\theta = 0.05$ for the non-null case. In each simulation, we generated $p = 100000$ independent SNPs with MAF $f_i$ independently drawn from Uniform$(0.1, 0.3)$. In CAUSE, it is assumed that the hidden confounder $U$ is standardized with $\beta_{XU} = 1$; we set the effect size from $U$ to $Y$ at $\beta_{YU} = 1$. We set the proportion of invalid IVs with correlated pleiotropic effects at $q = 0.3$. We set the heritabilities of $X$ and $Y$ at $h_X^2 = h_Y^2 = 0.3$. Then we generated the standardized effect sizes $\tilde{\beta}_{Xi}$ and $\tilde{\alpha}_i$ from a mixture of bivariate normal distribution:

$$\begin{pmatrix} \tilde{\beta}_{Xi} \\ \tilde{\alpha}_i \end{pmatrix} \sim \sum_{k=1}^{4} \pi_k \cdot N\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{Xk}^2 & 0 \\ 0 & \sigma_{\alpha k}^2 \end{pmatrix} \right). \tag{18}$$

Here $\pi = m/p$, and $\pi_1 = 0.2\pi, \pi_2 = \pi_3 = 0.8\pi, \pi_4 = 1 - \pi_1 - \pi_2 - \pi_3$; $\sigma_{X1}^2 = \sigma_{X2}^2 = \sigma_X^2, \sigma_{\alpha 1}^2 = \sigma_{\alpha 3}^2 = \sigma_\alpha^2, \sigma_{X3}^2 = \sigma_{X4}^2 = \sigma_{\alpha 2}^2 = \sigma_{\alpha 4}^2 = 0$, and $\sigma_X^2 = h_X^2/m_X$, $\sigma_\alpha^2 = [h_Y^2 - (\theta^2 + q\eta^2)h_X^2]/m_\alpha$. The non-standardized effect sizes were $\beta_{Xi} = \tilde{\beta}_{Xi}/\sqrt{2f_i(1-f_i)}$, and $\alpha_i = \tilde{\alpha}_i/\sqrt{2f_i(1-f_i)}$. We generated the indicator of an invalid IV, $Z_i$, from Bernoulli$(q)$, and $\beta_{Yi} = \theta\beta_{Xi} + Z_j\beta_{YU}\beta_{Xi} + \alpha_i$. With the standard error $s_i = 1/\sqrt{2Nf_i(1-f_i)}$, the GWAS summary statistics were generated as $\hat{\beta}_{Xi} \sim N(\beta_{Xi}, s_i^2)$, $\hat{\beta}_{Yi} \sim N(\beta_{Yi}, s_i^2)$.

For each combination of $(m, \theta, N)$ we did simulations 200 times (while the simulations in [14] were repeated only 100 times due to longer running time of CAUSE). For CAUSE, we used all $p$ SNPs to estimate the parameters, and used its default p-value threshold 0.001 to select the SNPs associated with the exposure. For other MR methods, we used the usual p-value threshold $5 \times 10^{-8}$ to select the exposure-associated SNPs as IVs. For comparison, we also applied the p-value threshold $5 \times 10^{-8}$ (instead of its default threshold 0.001) to select the SNPs for CAUSE.

### 2.9.3 Simulations with Weak Invalid IVs

As suggested by a reviewer, we did more simulations with many invalid IVs with weak effects (so-called "weak invalid IVs"), representing a scenario more challenging to identify invalid IVs with only weak effects. We set the number of IVs $m = 50$, sample size $N = 20000$, and $\gamma_i$'s from $N(0, h_x/m)$ for $i = 1, \cdots, 50$. We had the first 60% IVs as invalid IVs with uncorrelated pleiotropic effects $\alpha_i$'s from $N(0, h_y/m)$ and correlated pleiotropic effects $\phi_i$'s from $N(0, h_u/m)$ for $i = 1, \cdots, 30$, and set $\alpha_i = \phi_i = 0$ for $i = 31, \cdots 50$. Then we set $\hat{\sigma}_{Xi} = \hat{\sigma}_{Yi} = 1/\sqrt{N}$, and generated $\hat{\beta}_{Xi} \sim N(\gamma_i + \phi_i, \hat{\sigma}_{Xi}^2)$, $\hat{\beta}_{Yi} \sim N(\theta \cdot (\gamma_i + \phi_i) + \alpha_i + \phi_i, \hat{\sigma}_{Yi}^2)$, where $\theta$ was the true causal effect. We set $h_x = 0.5$, and tried different $h_y = 0.1, 0.2, 0.4, 0.6$, different $h_u = 0, 0.1$, and different $\theta = -0.2, -0.1, -0.05, 0, 0.05, 0.1, 0.2$. Here $h_x$, $h_y$ and $h_u$ could be viewed as the heritability of the exposure, outcome and confounder due to direct effects of the IVs. Note that, for $h_u = 0$ there was no correlated pleiotropy, while for $h_u = 0.1$ there was. The smaller $h_y$, the weaker the direct/pleiotropic effects and thus more difficult to identify the invalid IVs.

# 3 Results

## 3.1 Simulations: better type I error control and higher power of the new method than other MR methods

We compared our proposed method with ten most popular and new MR methods as shown in Table 1 through extensive simulations: MR-Mix, MR-ContMix, MR-CAUSE, MR-Lasso, MR-PRESSO, MR-Weighted-Median, MR-Weighted-Mode, MR-Egger, MR-RAPS, and MR-IVW (with a random-effect model throughout this paper). For evaluation, we also added MR-IVW-Oracle, an ideal but impractical method with all valid IVs known and being used, giving best possible performance. Since CAUSE requires full GWAS summary data (with both trait-associated and non-associated SNPs) with much longer running time, we divide the simulations into two parts. For main simulations, following [9, 10] we only generated summary statistics for exposure-associated SNPs, and compared cML-MA with other nine methods except CAUSE. For secondary simulations, we simulated both exposure-associated and non-associated SNPs as required by CAUSE [14], and compared cML-MA to all other 10 methods.

### 3.1.1 Main simulations: comparison with major MR methods

We did extensive simulations with the true causal model shown in Figure 1, including $N = 50,000$ or $100,000$ or $200,000$ subjects in each GWAS dataset, using $m = 10$, 20 or 100 SNPs as IVs, among which 0% to 60% were invalid IVs with IV Assumptions A2 or/and A3 being violated. For each setup we did 1000 simulations to compare the proposed cML methods and other existing MR methods, and for the 10 setups shown in Figure 2 we also increased the number of simulations to 10,000 to better estimate the type-I errors. For the 9 setups with 60% invalid IVs and with both IV Assumptions A2 and A3 being violated, we also applied cML-BIC-DP and cML-MA-BIC-DP. We used the nominal significance level of 0.05. Here we only show some representative results while all others are available in the Supplementary.

Figure 2 shows the empirical type-I errors (at the nominal level 0.05). First, in the cases with all valid IVs, the methods generally performed well, though MR-Weighted-Mode, MR-Weighted-Median and MR-Mix might be too conservative. On the other hand, MR-ContMix could have an inflated type I error rate, perhaps due to its inappropriately pre-selected tuning parameter value. Second, in the presence of 20% or 60% invalid IVs with IV Assumption A2 violated but Assumption A3 (thus the InSIDE assumption) holding, MR-PRESSO, MR-Lasso and MR-IVW could have inflated type I error rates. It is noted that, though the InSIDE assumption held, MR-Egger could have a slightly inflated type I error rate for small $m = 10$, but not for large $m = 100$. MR-Weighted-Mode gave the most highly inflated type I error rate with the large proportion (60%) of invalid IVs and with the small number of SNPs ($m = 10$). Third, in the cases with both IV Assumptions A2 and A3 violated, MR-IVW, MR-Egger and MR-PRESSO all had inflated type I error rates, while MR-Weighted-Mode and MR-Weighted-Median had largely inflated type I error rates with 60% invalid IVs, and so did MR-Lasso with 60% invalid IVs and with only $m = 10$ SNPs. In summary, as expected, MR-IVW was problematic in the presence of invalid IVs, MR-Egger did not perform if IV Assumption A3 is violated, and MR-PRESSO often had inflated type I errors; on the other hand, in agreement with [10], MR-Lasso, MR-Weighted-Median and MR-Weighted-Mode did not perform well with a small number of SNPs and with a high proportion of invalid IVs. We conclude that only

MR-cML-BIC and MR-Mix could control the type I error rates across all the scenarios, though MR-Mix was often too conservative (with too small type I errors), especially for a large number of SNPs/IVs. The results for type-I errors based on 10,000, instead of 1000, simulations, as shown in Supplementary Figure S6, were essentially the same.

Figure 3 shows the empirical type-I error (for $\theta = 0$) and power (for $\theta \neq 0$) curves. It is confirmed that our proposed method cML-MA-BIC always yielded a power curve close to that of MR-IVW-Oracle, the ideal test based on using only valid IVs. In particular, cML-MA-BIC was more powerful than MR-Mix and other methods (when their type I errors were close to the nominal level).

Figures 4 and 5 show the distributions of the causal estimates by each method for the true causal effect sizes $\theta = 0$ and 0.1 respectively. Again it is confirmed that the distribution of the $\theta$ estimates from our proposed cML-MA-BIC was almost the same as that from the ideal MR-IVW-Oracle. In particular, cML-MA-BIC, MR-ContMix and MR-Lasso (and MR-IVW-Oracle) always yielded (almost) unbiased estimates with smaller variances, while other methods sometimes gave biased estimates (and/or with much larger variances). In particular, as shown in Figure 5, MR-Mix was slightly biased (towards 0) for $m = 100$.

For the 9 setups with 60% invalid IVs and with both IV Assumptions A2 and A3 being violated, we applied cML-BIC-DP and cML-MA-BIC-DP, with various numbers of perturbations $T = 100, 200$ and 500. The complete results are shown in Supplementary Section S5.4. cML-BIC-DP and cML-MA-BIC-DP yielded similar results to those of cML-BIC and cML-MA-BIC, respectively, in terms of both point estimation and statistical inference, and both the GOF tests performed similarly in rejecting the null hypothesis (of unequal variance estimates from the model- and DP-based approaches) with low frequencies.

In summary, overall, among all the methods, cML-MA-BIC, MR-Mix and MR-ContMix performed best across all the scenarios; among the latter three, cML-MA-BIC was the clear winner for its higher power while better controlling the type I error rate.

Figure 2: Main simulations: empirical type-I error rates at the nominal level of 0.05 with sample size $N = 50000$ and with $m = 10$ or 100 SNPs, among which 0 to 60% were invalid IVs with the InSIDE assumption either holding or violated.

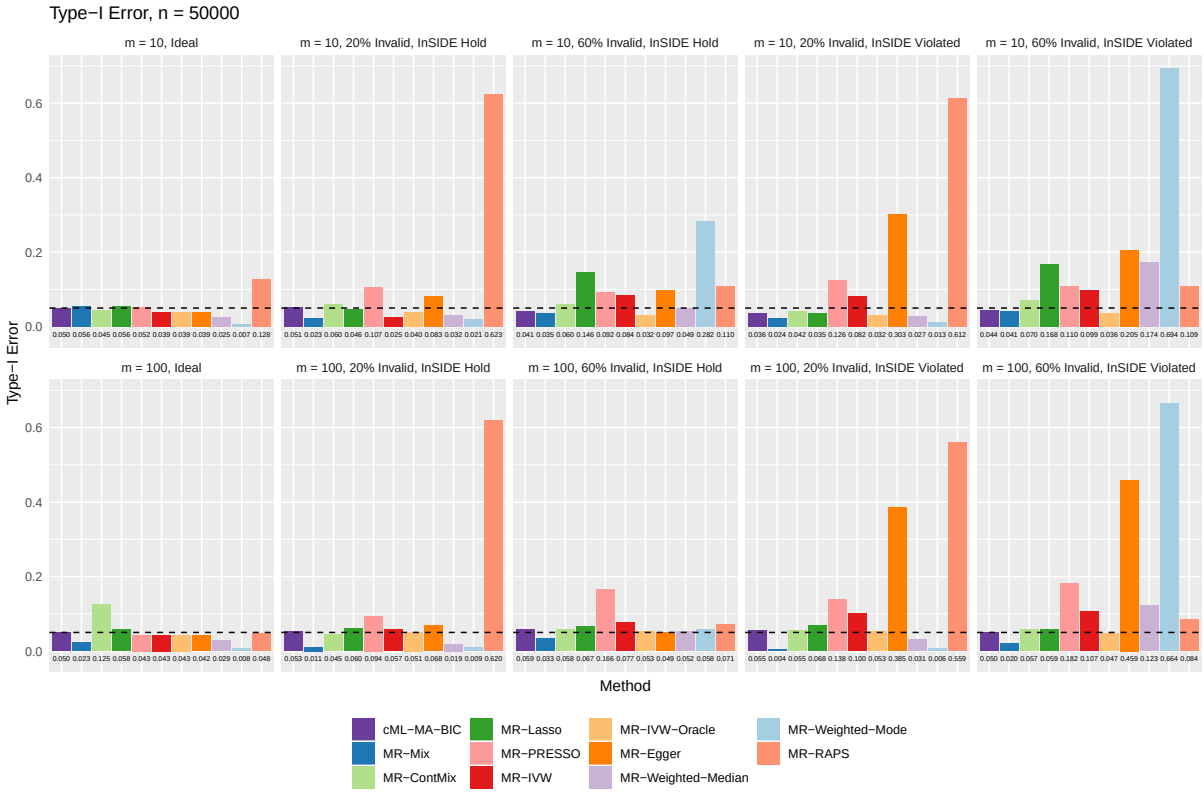


Figure 3: Main simulations: empirical type-I error (for $\theta = 0$) and power (for $\theta \neq 0$) curves with sample size $N = 50000$.

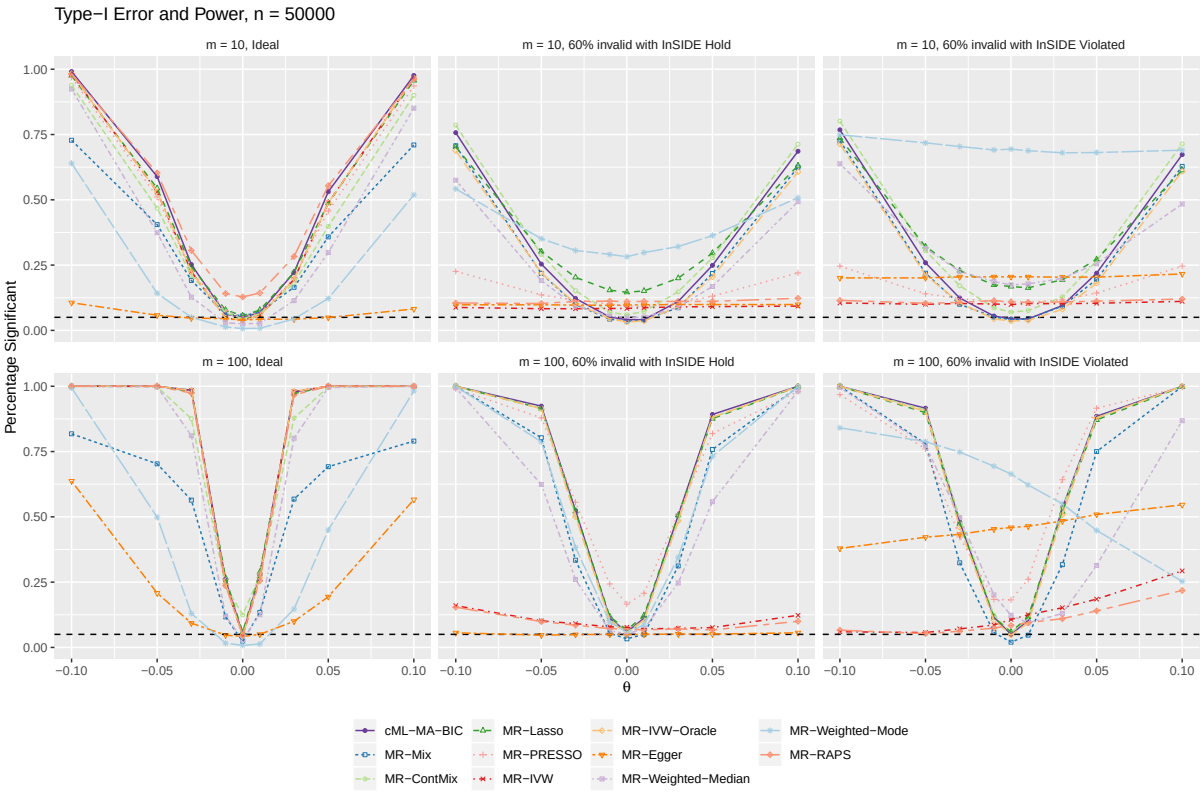


Figure 4: Main simulations: empirical distributions of the estimates of the causal effect $\theta$ by the methods with $N = 50000$ and $\theta = 0$. The numbers below each panel are the mean($\hat{\theta}$), SD($\hat{\theta}$), mean squared error (MSE) of $\hat{\theta}$ from top to bottom.

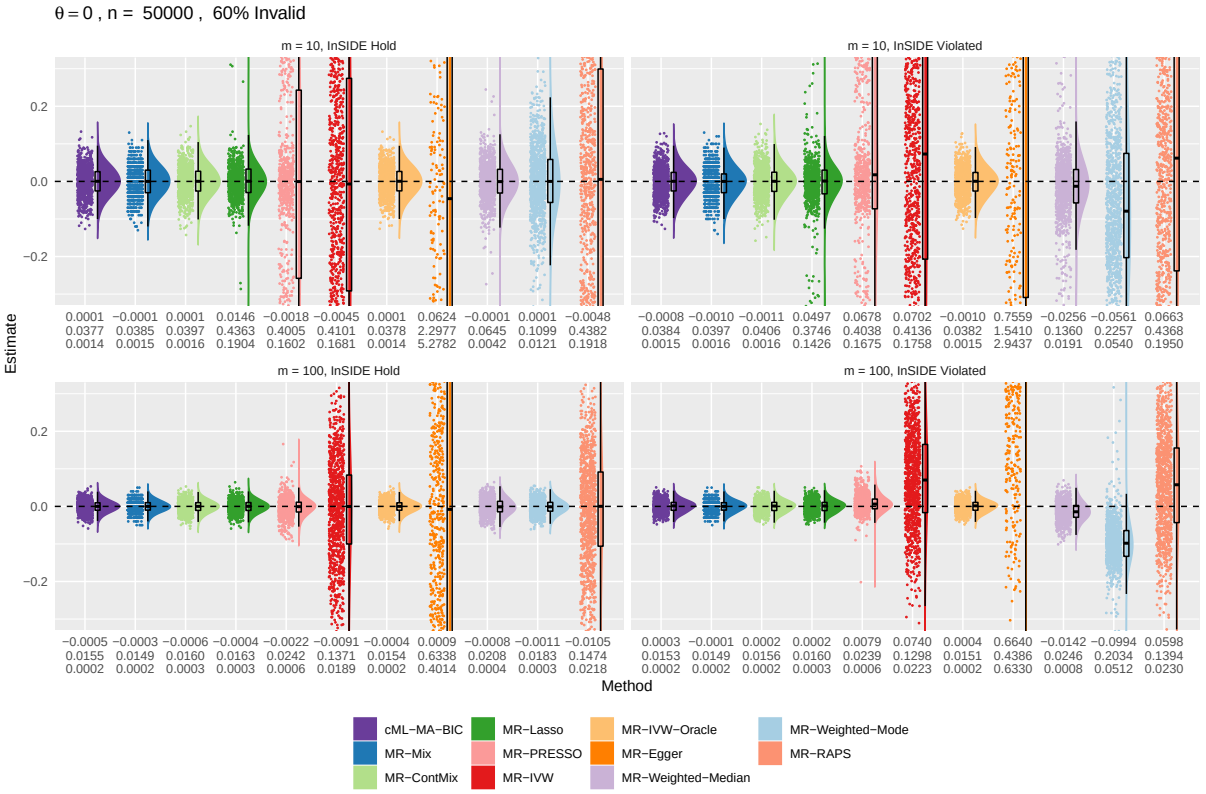
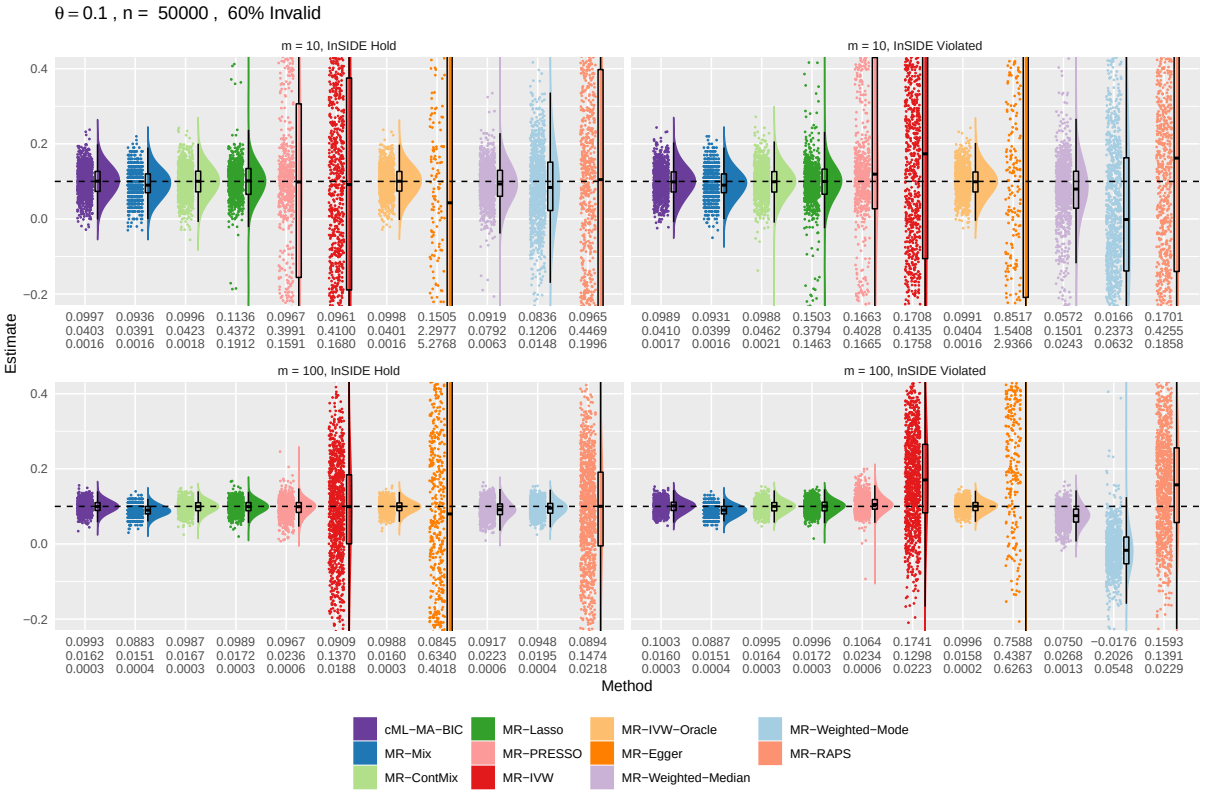


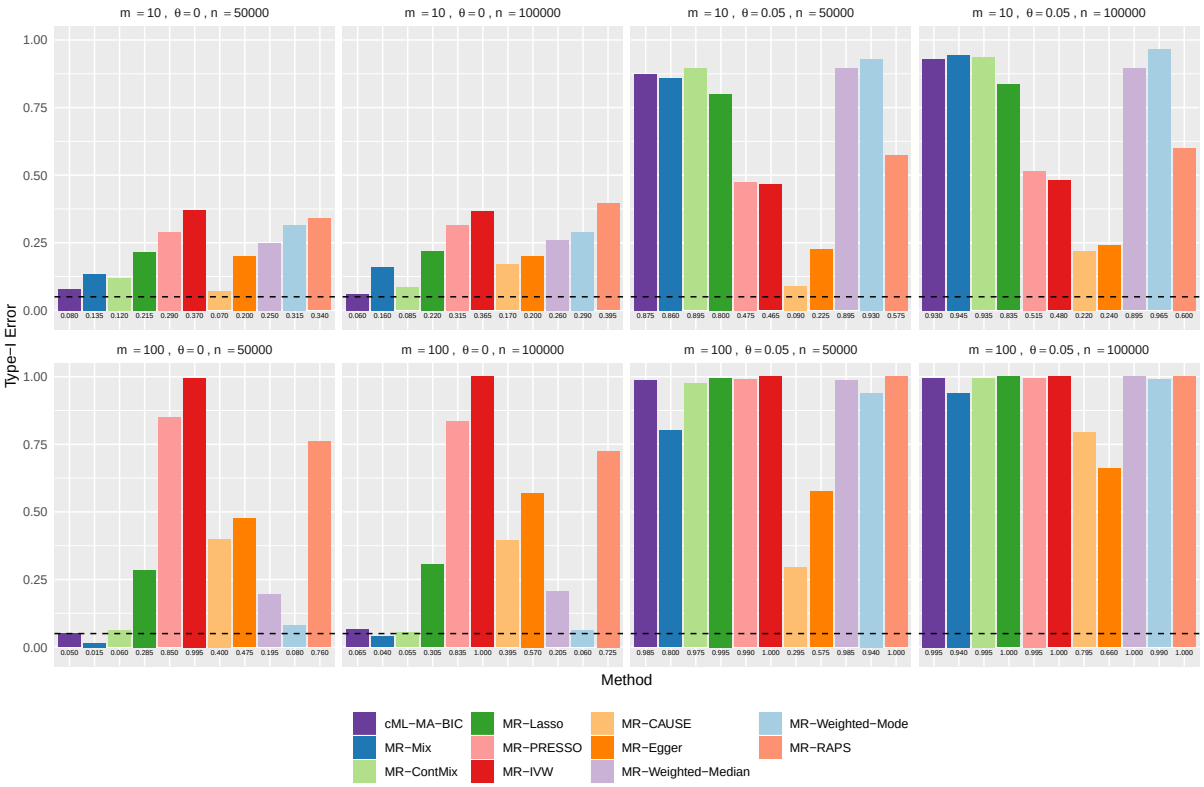
Figure 5: Main simulations: empirical distributions of the estimates of the causal effect $\theta$ by the methods with $N = 50000$ and $\theta = 0.1$. The numbers below each panel are the mean($\hat{\theta}$), SD($\hat{\theta}$), mean squared error (MSE) of $\hat{\theta}$ from top to bottom.




### 3.1.2 Secondary simulations: comparison with CAUSE

We did simulations in the framework of MR-CAUSE as described in [14]. Figure 6 shows the empirical type-I error (for $\theta = 0$) and power (for $\theta = 0.05$) for the methods with $m = 10$ or 100 exposure-associated SNPs and sample size $N = 50000$ or 100000. It is clear that our proposed method cML-MA-BIC could control type-I error well with high power. In contrast, CAUSE could have largely inflated type-I error rates and much lower power than cML-MA-BIC. Here the results for CAUSE were based on using its default p-value threshold of 0.001 to select exposure-associated SNPs; as shown in the Supplementary, using the threshold $5 \times 10^{-8}$ (as for other methods shown here) did not give better results for CAUSE. The poor performance of CAUSE here is in agreement with that shown in the original CAUSE paper ([14], Figure SN1): when it was high powered to detect SNP-exposure and SNP-outcome associations, CAUSE tended to give dramatically inflated false positive rates. In addition, compared to cML-MA-BIC, both MR-Mix and MR-ContMix had higher inflated type I error rates for small $m = 10$; on the other hand, MR-Mix was too conservative with too small type I error rates and lower power for $m = 100$ with the small sample size $N = 50000$.

Figure 6: Secondary simulations: empirical type-I error rates (for $\theta = 0$) and power (for $\theta \neq 0$) with sample size $N = 50000$ or 100000, and with $m = 10$ or 100 exposure-associated SNPs.




### 3.1.3 Simulations with Weak Invalid IVs

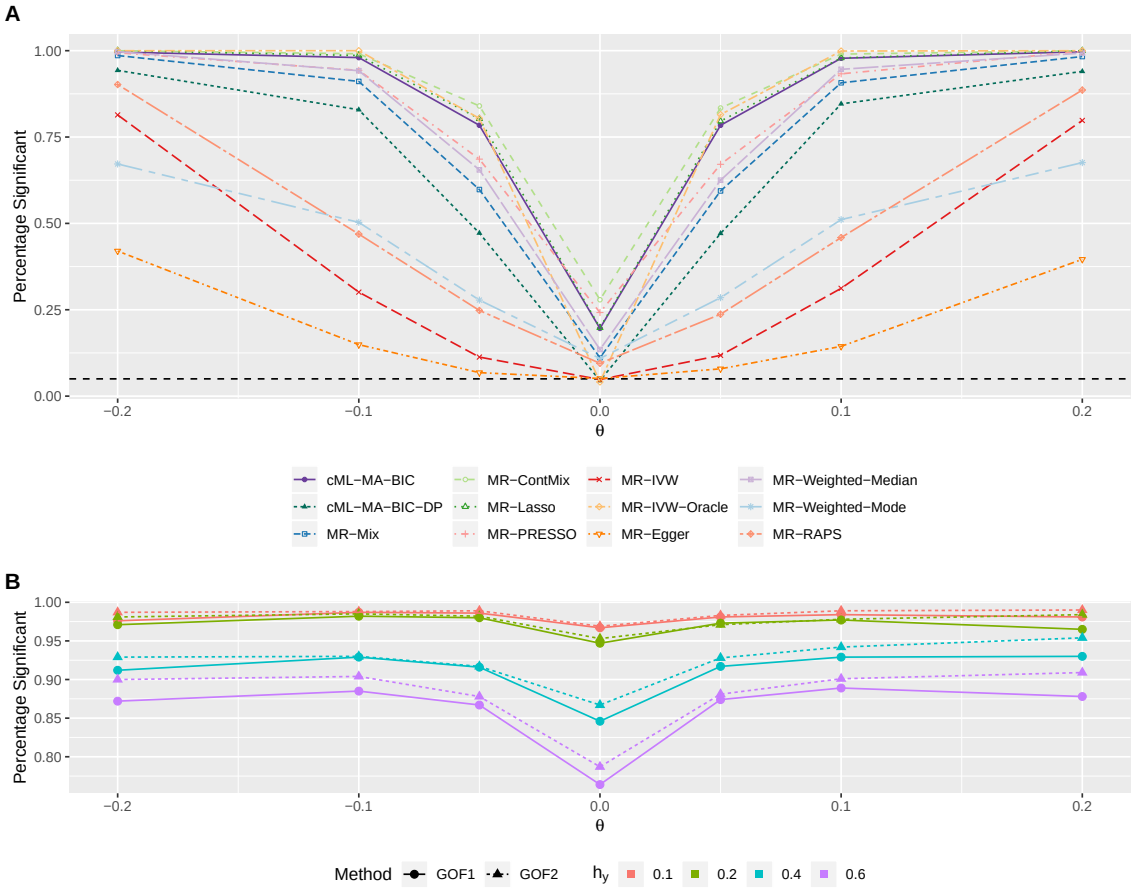
We did 1000 simulations for each setup with a small sample size and many invalid IVs having weak direct/pleiotropic effects. Figure 7(A) shows some representative results for the empirical type-I error (for $\theta = 0$) and power (for $\theta \neq 0$) curves for $h_y = 0.2$ and $h_u = 0$ (i.e. no correlated pleiotropy); the complete results are given in the Supplementary. It is clear that under this challenging situation, in addition to the ideal MR-IVW-Oracle, only three methods could control the type I error satisfactorily (while all others could not): cML-MA-BIC-DP, MR-IVW and MR-Egger, but cML-MA-BIC-DP was much more powerful than the other two. It is noted that here the (weak) direct effects were balanced (i.e. with mean 0) and from a normal distribution, explaining the reasons for the relatively good performance by MR-IVW and MR-Egger. Nevertheless, as shown in Supplementary Tables S107-110, as the pleiotropic effect sizes (i.e. $h_y$) increased, the power of cML-MA-BIC-DP improved (by better identifying invalid IVs), but, perhaps surprisingly, both MR-IVW and MR-Egger became less powerful (because of the increasing error variances in their models by treating the pleiotropic effects as random).

In the presence of correlated pleiotropy with $h_u = 0.1$, as shown in Supplementary Tables S111-114, only our proposed cML-BIC-MA-DP could satisfactorily control type-I error and was well powered, while all other methods, including MR-IVW, MR-Egger and MR-RAPS, yielded inflated type-I errors and possibly low power.

Figure 7(B) shows the relative frequencies of the goodness-of-fit tests' rejecting the null hypothesis that the model-based variance was equal to the DP-based variance by cML. The

proposed goodness-of-fit tests could detect with high power the problem with cML-MA-BIC; the two GOF tests performed similarly, though GOF2 was slightly more powerful (presumably due to its taking advantage of the causal estimates being nearly normally distributed). In addition, as $h_y$ decreased, it became harder to identify invalid IVs, leading to more inflated type I error rates by most methods, including cML-MA-BIC; accordingly the two GOF tests rejected the null hypothesis more frequently, demonstrating their effectiveness.

Figure 7: Simulation results with many invalid IVs having weak pleiotropic effects. (A) Empirical type-I error (for $\theta = 0$) and power (for $\theta \neq 0$) curves with $h_y = 0.2$ and $h_u = 0$. (B) Relative frequencies of the goodness-of-fit tests rejecting the null hypothesis.




### 3.1.4 Computational Time

We did simulations to compare the running times of different methods as detailed in Supplementary Section S8. In summary, cML-MA-BIC runs reasonably fast: its computing time was comparable to that of MR-ContMix and MR-RAPS, while being faster than MR-Mix and MR-Weighted-Mode but slower than MR-IVW, MR-Egger and MR-Weighted -Median. As expected, using more random starting points or data perturbation would take much more time. Nevertheless, on a MacBook Pro laptop, with 10 to 100 SNPs/IVs, it took from a few seconds to less than 10 minutes with cML-MA-BIC-DP with five random starts and $T = 200$ data perturbations; in contrast, cML-MA-BIC with five random starts ran from 0.3 to 4 seconds.

## 3.2 Identifying Causal Risk Factors of Complex Diseases

We compare our proposed cMLwith other methods to identify possible causal effects of 12 risk factors on three cardio-metabolic diseases, coronary artery disease (CAD) [27], stroke [28] and type 2 diabetes (T2D) [29], plus asthma largely used as a negative control [30]. These 12 risk factors (and their corresponding GWAS) are LDL cholesterol, HDL cholesterol, triglycerides (TG) [31], drinks per week (Alcohol), ever regular smoker (Smoke) [32], body fat percentage (BF) [33], birth weight (BW) [34], body mass index (BMI) [35], Height [26], fasting glucose (FG) [36], systolic blood pressure (SBP) and diastolic blood pressure (DBP) [37]. As used and shown in [14], the sample sizes of these GWAS ranged from 46186 for FG and 69033 for T2D, to 100716 for BF, 142486 for asthma, 188577 for TG, HDL and LDL, 253288 for Height, 322154 for BMI, then 446696 for Stroke, 547261 for CAD, 757601 for DBP and SBP, finally to near and above a million for Alcohol and Smoke respectively. For each risk factor/exposure-disease/outcome pair we used the set of LD-independent SNPs as IVs as described in [14] (in their Supplementary Table 4), and applied all methods except CAUSE to the GWAS summary statistics of these SNPs; for CAUSE, we extracted the results from the original paper [14].

In [14], the 48 exposure-outcome pairs were classified into 5 categories: considered causal (9 pairs), likely causal as supported by the literature (10 pairs), correlated but unknown to be causal or with conflicting evidence (17 pairs), unrelated (10 pairs), and considered non-causal (2 pairs); here we combined the first two categories into one to represent (likely) causal pairs. In Figure 8 we compare cML-MA-BIC with three representative methods, CAUSE, a new one specifically proposed to deal with correlated pleiotropy, and two MR methods, one robust and competitive (MR-Mix) and the other perhaps most popular (MR-IVW), for all of these 48 risk factor-disease pairs; Figure 9 compares the numbers of the detected pairs by these and other methods, and other detailed results (of the causal parameter estimates, SEs and p-values) for all methods are available in the Supplementary. Here we discuss the results based on the Bonferroni adjusted significance level of $0.05/48 \approx 0.001$. For the 19 known or likely causal risk factor-disease pairs, cML-MA-BIC, CAUSE, MR-Mix and MR-IVW identified 15, 7, 12 and 12 significant pairs respectively; among the 17 correlated pairs, the four methods detected 6, 0, 4 and 1 pairs respectively; among the 10 unrelated pairs, none of the methods identified any, while for the two pairs of non-causal pairs, all four methods indicated one and the same one (i.e. HDL-CAD, which is still under debate as whether it is truly causal). In addition, although none of the methods detected causal Smoke-asthma, Smoke-T2D and BMI-stroke, our method was the only one among the four methods giving marginally significant p-values. It is clear that our proposed cML-MA-BIC identified the largest numbers of the known or likely causal pairs, showcasing its highest power. On the other hand, it also detected more pairs from the category of "correlated" pairs. It is possible that these pairs, such as BW-CAD, BW-T2D and DBP/SBP-T2D, may be false positives, but at the same time, they may be truly causal as to be confirmed (or refuted) by further studies. As shown in Figure 10, only based on the data, there seems to be evidence to support these causal relationships as detected by our cML-MA-BIC and many other tests (shown in the Supplementary). It is noted that CAUSE also gave marginally significant p-values $< 0.05$ for three of these six pairs.

Among the tens to about 1000 SNPs used as IVs for the 48 risk factor-disease pairs, 0 to 96 SNPs, mostly $< 10$, ranging from 0% to 30%, mostly $< 3\%$, were identified as invalid IVs by our method (Supplementary Table S5). However, as for BW-CAD and BW-T2D in Figure 10, although only five (out of 65 and 54) SNPs were identified as invalid IVs based on the BIC-

Figure 8: Results of cML-MA-BIC, MR-CAUSE, MR-Mix and MR-IVW to detect causal relationships among 48 risk factor-disease pairs.

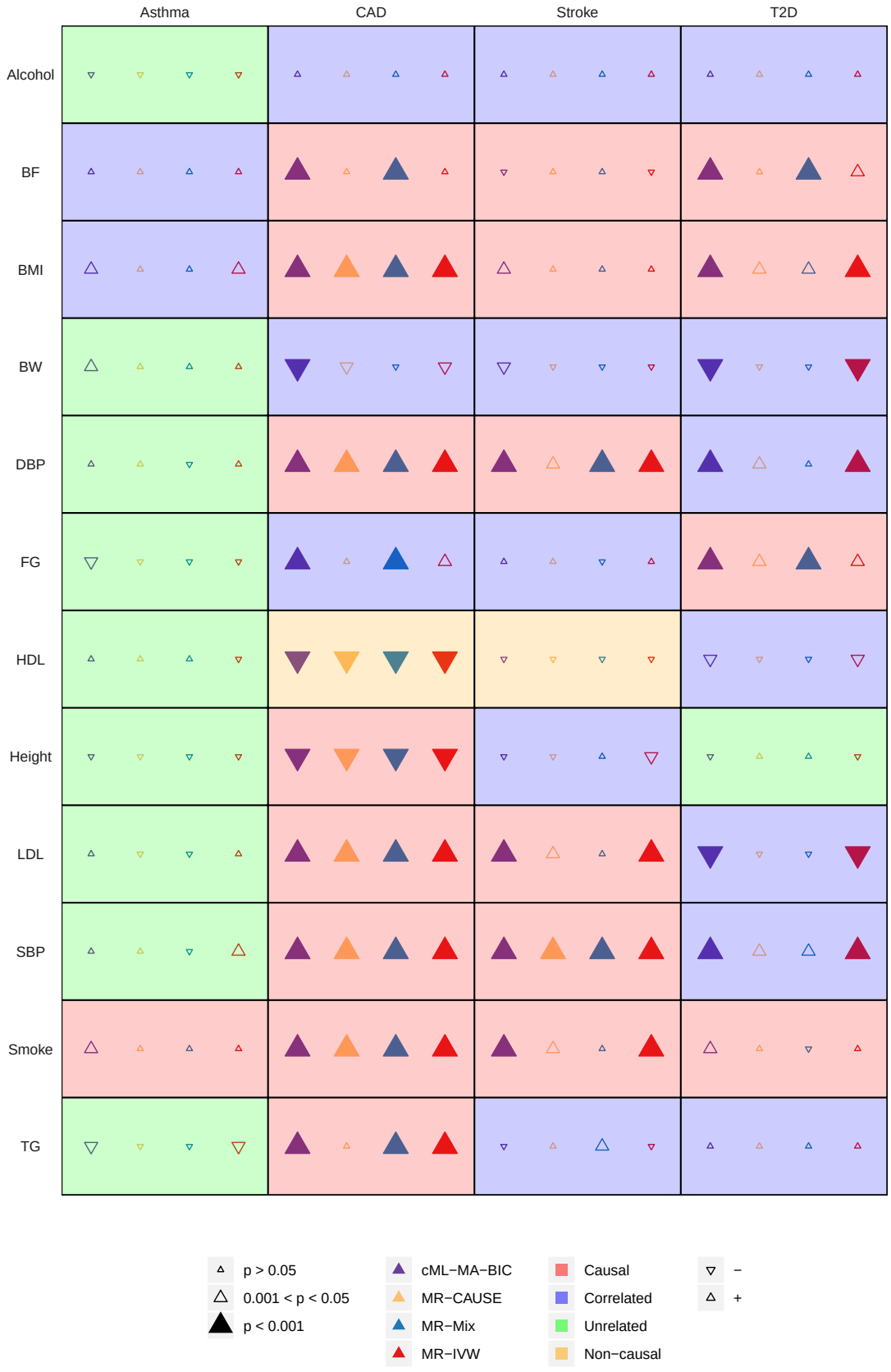
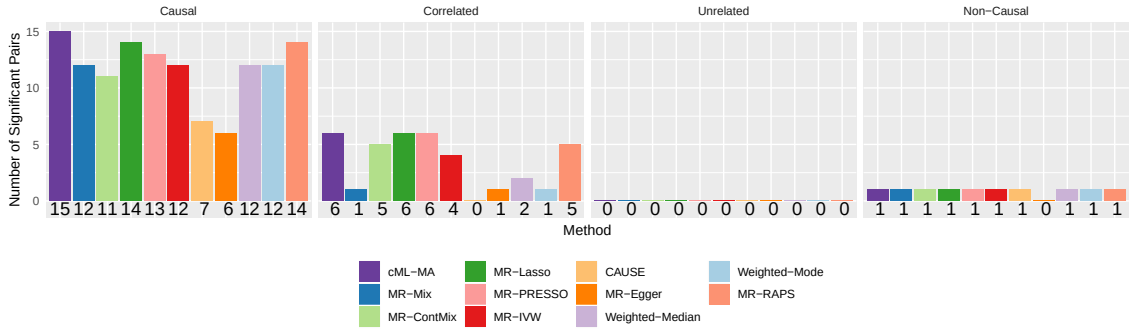


Figure 9: The numbers of the significant risk factor-disease pairs detected by various methods at the significance cutoff of p-value $< 0.001$. MR-RAPS refers to RAPS2 (with the Tukey loss and overdispersion).




selected best models, the models containing up to $K = 20$ SNPs as invalid IVs were estimated to be better (with lower BIC values) than the model treating all SNPs as valid IVs (i.e. $K = 0$). In general there were fewer invalid IVs for asthma, stroke and T2D, but more for CAD. It can be seen that the influence of invalid IVs is somewhat complex; the difference between the two causal estimates with and without invalid IVs may or may not simply depend on the presence or the number of invalid IVs, but more on the configuration of invalid IVs relative to that of valid ones. For example, with the same number of invalid IVs detected for the BW-CAD and BW-T2D pairs, the causal estimates with and without the detected invalid IVs were almost the same for the former pair, but more different for the latter (Figure 10).

We then applied cML-MA-BIC-DP with $T = 200$ perturbations and 10 additional random starts in each perturbation. Compared to those of cML-MA-BIC, the results remained the same for 46 pairs in terms of statistical significance, but changed from being significant to marginally significant for only two pairs: one known or likely causal pair, BF-CAD; and one correlated pair, BW-T2D.

In summary, it is encouraging that most methods detected similar numbers of significant pairs, though MR-Egger and CAUSE detected much fewer as shown in Figure 9. In addition, there were some notable differences in specific pairs detected across the methods. For example, while our methods detected the causal pairs BF-T2D and FG-T2D, MR-IVW missed both, and MR-RAPS missed BF-T2D. We conclude that our proposed methods performed competitively.

## 3.3 Secondary Real Data Analysis

We compared the type-I errors of the cML methods and other existing MR methods with 63 pairs of traits that were not genetically correlated (with their $p$-values greater than 0.05). For 10 of 63 pairs with HOMA as the exposure, `TwoSampleMR` gave only 2 LD-independent SNPs as IVs, which was too small to apply MR-ContMix, MR-Lasso, MR-Egger, MR-Weighted-Median, MR-Weighted-Mode, and MR-PRESSO; although cML methods are applicable to only 2 IVs, they would require $K = 0$, i.e. no invalid IVs. Hence we applied all methods to the other 53 pairs (without HOMA as the exposure). Figure 11 shows the Q-Q plots of cML-MA-BIC, cML-MA-BIC-DP, MR-Mix, MR-ContMix, MR-IVW and MR-RAPS for these 53 pairs; Supplementary Figure S5 shows the results for all methods. While the methods based on selection of invalid IVs, i.e. cML-MA-BIC, MR-Mix and MR-ContMix, all seemed to

Figure 10: For 4 risk factor-disease pairs, the left panels show the numbers of invalid IVs versus BIC values, while the right panels shows $\hat{\beta}_{Xi}$ versus $\hat{\beta}_{Yi}$ (with their errors bars indicating $\hat{\sigma}_{Xi}$ and $\hat{\sigma}_{Yi}$). In the right panels, those for invalid IVs detected by BIC are blue colored; the red solid lines give the causal estimates (after removing the detected invalid IVs), and the black dashed lines are for the estimates based on all IVs.

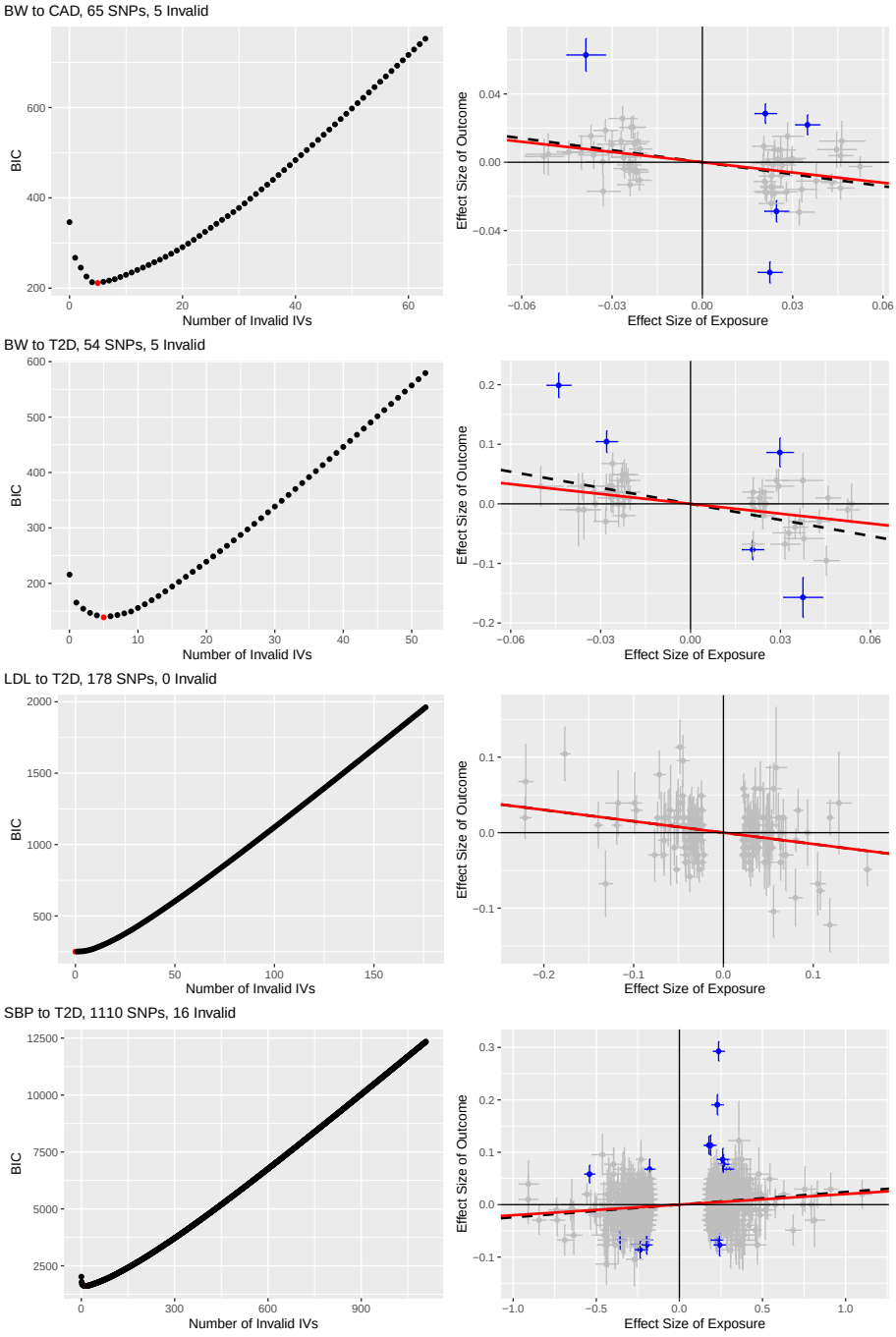
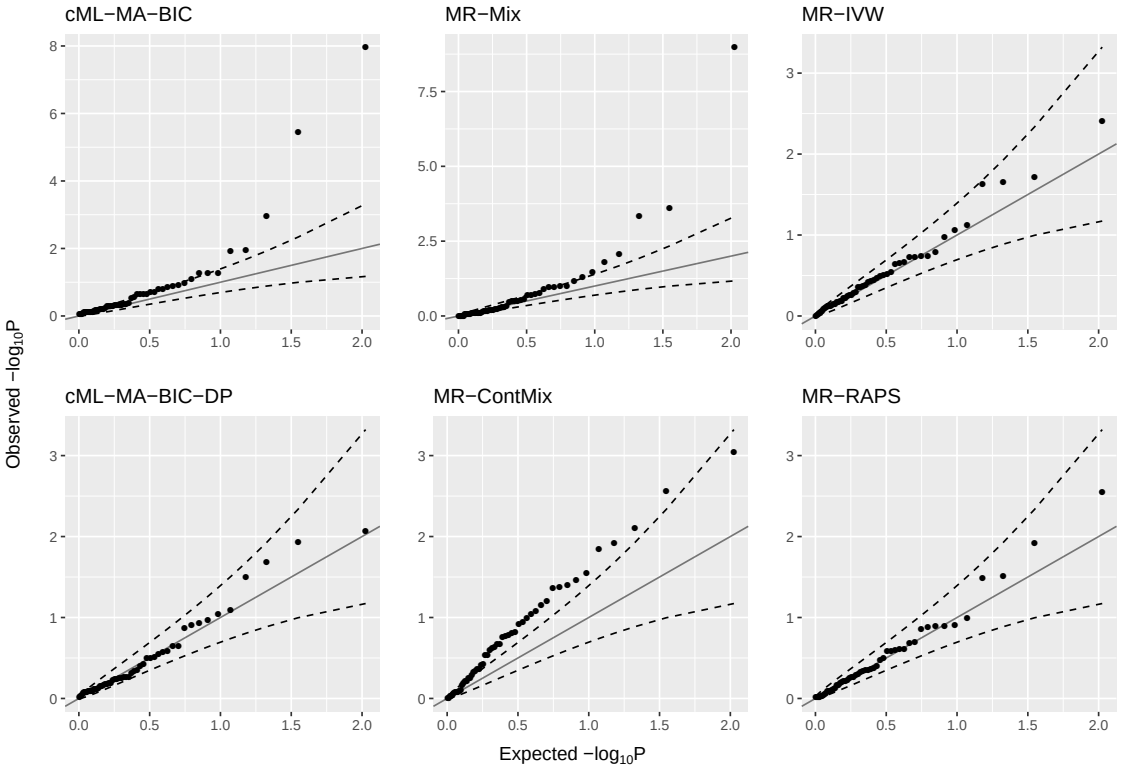


Figure 11: Q-Q plots for 53 (likely) null trait-pairs in the secondary real data examples.




have inflated type I errors, the proposed cML-MA-BIC-DP with $T = 200$ perturbations and 10 additional random starts in each perturbation, along with MR-IVW and MR-RAPS, appeared to perform well in satisfactorily controlling the type I errors. The complete results are in the Supplementary.

# 4 Discussion

We have proposed several methods based on constrained maximum likelihood (cML) to consistently identify invalid IVs with either or both of correlated and uncorrelated pleiotropic effects, thus leading to consistent estimation and inference of the causal effect between an exposure and an outcome. For finite samples, the (asymptotic) selection consistency may not be achieved. To account for model selection uncertainty, we first propose a model-averaging approach, cML-MA-BIC, which performs better, especially in better controlling the type I error rate, than the selection-based version, cML-BIC. In addition, in more challenging situations with many invalid IVs with only weak pleiotropic effects, both model selection and model averaging may not perform well by failing to fully account for model selection uncertainty; accordingly, we propose a version based on data perturbation, cML-MA-BIC-DP, which could control the type I error rate satisfactorily across all simulated and real data examples. However, cML-MA-BIC-DP is computationally more demanding and may be conservative with some loss of power as compared to cML-MA-BIC. To help a user determine which one is preferred, we propose two GOF tests for the null hypothesis that the two approaches give the equivalent variance estimates (and thus inferential results); if the null hypothesis is not rejected, one can simply apply cML-

MA-BIC; otherwise, cML-MA-BIC-DP is preferred. All the proposed methods are applicable to GWAS summary data.

Three new competitors to our methods include CAUSE [14], MR-Mix [13] and MR-ContMix [9]. As shown in our simulations and in agreement with the original study [14], CAUSE may have dramatically inflated type I error rates, presumably due to its complex modeling and estimation of the parameters related to *hidden/unobserved* confounding. Furthermore, CAUSE imposes an assumption of $\gamma_i\phi_i = 0$ (i.e. that an IV can only have a direct effect on either the exposure or the hidden confounder, but not on both simultaneously), though it is yet unknown and debatable whether this assumption is reasonable for real data. It seems advantageous that our proposed methods simply estimate only a small number of necessary parameters (without such an assumption). Both MR-Mix and MR-ContMix are based on multivariate normal mixture models on various effect sizes across the genome, which not only impose stronger modeling assumptions, but also are computationally more demanding. In our experiments, overall, MR-Mix performed only second to our proposed methods with mostly controlled type I error rates and high power, though it might still have either largely inflated or too conservative type I errors while giving biased estimates (e.g. Figures 6 and 5). On the other hand, it is challenging to pre-select a fixed tuning parameter in MR-ContMix, which may negatively influence its performance in some situations as shown in our simulations. At the same time, it is confirmed that two most popular methods, MR-IVW and MR-Egger, do not perform well with dramatically inflated type I error rates and low power in the presence of correlated pleiotropy. Most importantly, we conclude that our proposed cML-MA-BIC (or cML-MA-BIC-DP) was the overall winner based on our extensive numerical studies.

MR-PRESSO and MR-Lasso are two existing methods looking most similar to our proposed methods. For our cML estimates, if an SNP has an estimated total direct effect $\hat{r}_i \neq 0$, then it is an invalid IV and does not contribute to estimating the causal effect $\theta$ (as shown in the Methods section); otherwise it is a valid IV and contributes to estimating $\theta$. Hence, our methods work by selecting and (implicitly) removing invalid IVs, in which sense they are related to MR-PRESSO and MR-Lasso (and an improved variant of MR-PRESSO [38]). However, there are some importantly differences. First, we propose a BIC for consistent model selection (and weighting) while MR-PRESSO uses resampling-based significance testing and MR-Lasso is based on a heuristic heterogeneity criterion. We have a rigorous theory to support our proposed method. Second, both MR-PRESSO and MR-Lasso draw inference on the causal effect $\theta$ based on a single selected model; due to selection bias, they often have inflated type I error rates and in general biased estimates of $\theta$. In contrast, by accounting for model selection uncertainties through model averaging, our method cML-MA-BIC performs much better as shown in simulations. Third, since MR-PRESSO selects invalid IVs one by one while ours and MR-Lasso select multiple ones simultaneously, MR-PRESSO may miss some invalid IVs (e.g. as well known in statistics that two invalid IVs/outliers may not appear so if checked one by one). In addition, MR-PRESSO fails to properly account for the variability of the delete-1 (IV) estimates of the causal effect while assuming that the delete-1 estimates are all accurate, which may be false, leading to both false positives and false negatives in selecting invalid IVs. On the other hand, MR-Lasso depends on a specified candidate set of the tuning parameter values for the Lasso penalty, which may be difficult to specify *a priori*. There is also lack of theoretical justification for its heterogeneity-based model selection criterion. Furthermore, in general, as most penalized methods, MR-Lasso yields biased estimates due to the shrinkage effects of the Lasso penalty. Finally, MR-LASSO, as MR-IVW and MR-Egger, imposes the NOME ("NO

Measurement Errors") assumption by ignoring the variability in estimating each IV-exposure association. Presumably due to these reasons, our proposed methods performed much better than the other two methods in our numerical examples.

Our proposed cML-MA-BIC not only performs extremely well in our numerical examples, but also is quite simple and intuitive with strong theoretical support. In fact, it may be surprising that such a method has not appeared in the MR literature. We note that our methods are based on classic statistical theory (for "large $n$, small $m$", i.e. asymptotics for a large sample size and a fixed/small number of parameters), which is suitable for typical GWAS (with $n$ in tens to hundreds of thousands, while $m$ is no more than a few hundreds). It might be of interest to extend our methods to a high-dimensional "large $n$, large $m$" scenario with even a much larger number of IVs: instead of using the full likelihood, we can use the profile likelihood [6] as shown in the Supplementary; we will also need to adopt or develop some new model selection criteria for high-dimensional data [39, 40, 41]. There are other limitations of our proposed methods. First, we propose a fast algorithm to select valid (or equivalently, invalid) IVs to obtain cML estimates. Since it is a combinatorial and non-convex variable selection problem, the proposed algorithm cannot guarantee finding a global solution. Nevertheless, in our simulations it yielded good results with only one starting value (by setting all parameters at 0). A simple strategy is to use multiple random starting values as used in the real data examples, for which little difference was found. In the future other more sophisticated algorithms [42, 43] may be adapted and applied. Second, we assume that the two GWAS (summary) data for the exposure and outcome are independent. As in CAUSE, we may estimate and model possible correlations between the two GWAS datasets due to overlapping subjects, and then modify the log-likelihood accordingly. More generally, it is critical to adjust for possible sample structure, including population stratification and subject relatedness, present in some GWAS data [44, 45]. Third, as in typical MR applications, we used the same GWAS sample to select significant SNPs as IVs (to meet IV Assumption A1) before using the same data for inference. It is known that this double-use of the data could lead to biased inference due to selection bias. Alternatively, we may use an independent GWAS sample to select SNPs to avoid selection bias as in the three-sample MR design [46], or account for selection explicitly [44, 47]. Fourth, we considered only independent SNPs as IVs; extending to using correlated SNPs as IVs may be useful in other applications, e.g. transcriptome-wide association studies [48, 49, 50, 51, 52, 53, 54], which equally face the analysis challenges with pleiotropic SNPs and thus invalid IVs [55]. These are interesting topics for future investigation.

# Data and Code Availability

This study used the GWAS summary datasets that are all publicly available as indicated in their corresponding references. The proposed methods are implemented in R package `MRcML`, which is publicly available to download on GitHub at `https://github.com/xue-hr/MRcML`. All other MR methods used for comparison are in publicly available R packages with links given in the Web Resources section.

# Acknowledgments

We thank the two reviewers for many helpful and insightful comments, leading to much im-

provement. This research was supported by NIH grants R01 AG069895, RF1 AG067924, R01 AG065636, R01 HL116720, R01 GM113250 and R01 GM126002, by NSF grant DMS 1711226, and by the Minnesota Supercomputing Institute at the University of Minnesota.

# Declaration of Interests

The authors declare no competing interests.

# Web Resources

Software/R package for MR-cML, `https://github.com/xue-hr/MRcML`
CAUSE, `https://github.com/jean997/cause`
MR-Mix, `https://github.com/gqi/MRMix`
MR-ContMix, `https://cran.r-project.org/web/packages/MendelianRandomization`
MR-PRESSO, `https://github.com/rondolab/MR-PRESSO`
MR-Lasso, `https://onlinelibrary.wiley.com/doi/full/10.1002/gepi.22295`
MR-IVW, MR-Egger, MR-Weighted-Median, MR-Weighted-Mode, MR-RAPS, `https://github.com/MRCIEU/TwoSampleMR`
LD Hub, `http://ldsc.broadinstitute.org/lookup`
OMIM, `https://www.omim.org/`

# References

1. Verbanck, M., Chen, C.Y., Neale, B. and Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. Nature genetics *50*, 693-698.

2. Watanabe, K., Stringer, S., Frei, O., Mirkov, M.U., de Leeuw, C., Polderman, T.J., van der Sluis, S., Andreassen, O.A., Neale, B.M. and Posthuma, D. (2019). A global overview of pleiotropy and genetic architecture in complex traits. Nature genetics *51*, 1339-1348.

3. Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016). Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. Genetic Epidemiology *40*, 304-314.

4. Bowden, J., Del Greco, M. F., Minelli, C., Davey Smith, G., Sheehan, N., and Thompson, J. (2017). A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. Statistics in Medicine *36*, 1783-1802.

5. Zhao, J., Ming, J., Hu, X., Chen, G., Liu, J. and Yang, C. (2020). Bayesian weighted Mendelian randomization for causal inference based on summary statistics. Bioinformatics *36*, 1501-1508.

6. Zhao, Q., Wang, J., Hemani, G., Bowden, J., and Small, D. S. (2020). Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. Annals of Statistics *48*, 1742-1769.

7. Kang, H., Zhang, A., Cai, T. T., and Small, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. Journal of the American Statistical Association, *111*, 132-144.

8. Guo, Z., Kang, H., Cai, T. T., and Small, D.S. (2018). Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. Journal of the Royal Statistical Society: Series B (Statistical Methodology) *80*, 793-815.

9. Burgess, S., Foley, C. N., Allara, E., Staley, J. R., and Howson, J. M. (2020). A robust and efficient method for Mendelian randomization with hundreds of genetic variants. Nature communications *11*, 1-11.

10. Slob, E. A., and Burgess, S. (2020). A comparison of robust Mendelian randomization methods using summary data. Genetic Epidemiology *44*, 313-329.

11. Zhu, X. (2020). Mendelian randomization and pleiotropy analysis. Quantitative Biology 2020, 1-11.

12. Bowden, J., Del Greco M, F., Minelli, C., Davey Smith, G., Sheehan, N.A. and Thompson, J.R. (2016). Assessing the suitability of summary data for mendelian randomization analyses using MR-Egger regression: the role of the $I^2$ statistic. International Journal of Epidemiology *45*, 1961-1974.

13. Qi, G., and Chatterjee, N. (2020). Mendelian randomization analysis using mixture models for robust and efficient estimation of causal effects. Nature Communications *10*, 1-10.

14. Morrison, J., Knoblauch, N., Marcus, J. H., Stephens, M., and He, X. (2020). Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. Nature genetics *52*, 740-747.

15. Burgess, S., Bowden, J., Dudbridge, F., and Thompson, S. G. (2016). Robust instrumental variable methods using multiple candidate instruments with application to Mendelian randomization. arXiv, 1606.03279.

16. Windmeijer, F., Farbmacher, H., Davies, N., and Davey Smith, G. (2016). On the use of the lasso for instrumental variables estimation with some invalid instruments. Journal of the American Statistical Association *114*, 339-350.

17. Hartwig, F. P., Davey Smith, G., and Bowden, J. (2017). Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. International Journal of Epidemiology *46*, 1985-1998.

18. Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. International Journal of Epidemiology *44*, 512-525.

19. Didelez, V., and Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. Statistical Methods in Medical Research *16*, 309-330.

20. Burgess, S., Butterworth, A. S., and Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. Genetic Epidemiology *37*, 658-665.

21. Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model selection: an integral part of inference. Biometrics, pp. 603-618.

22. Shen, X., and Ye, J. (2002). Adaptive model selection. Journal of the American Statistical Association *97*, 210-221.

23. Mood, A.M., Graybill, F.A. and Boes, D.C. (1974). Introduction to the Theory of Statistics 1974 (McGraw-Hill Kogakusha)

24. Zheng, J., Erzurumluoglu, A.M., Elsworth, B.L., Kemp, J.P., Howe, L., Haycock, P.C., Hemani, G., Tansey, K., Laurin, C., Pourcain, B.S. (2017). LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. Bioinformatics *33*, 272-279.

25. Allen, H.L., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., et. al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*, 832-838.

26. Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Kutalik, Z., Amin, N., et. al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nature genetics *46*, 1173-86.

27. van der Harst, P. and Verweij, N. (2018). Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. Circulation research *122*, 433-443.

28. Malik, R., Chauhan, G., Traylor, M., Sargurupremraj, M., Okada, Y., Mishra, A., Rutten-Jacobs, L., Giese, A.K., Van Der Laan, S.W., Gretarsdottir, S., et. al. (2018). Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. Nature genetics *50*, 524-537.

29. Morris, A.P., Voight, B.F., Teslovich, T.M., Ferreira, T., Segre, A.V., Steinthorsdottir, V., Strawbridge, R.J., Khan, H., Grallert, H., Mahajan, A., et. al. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nature genetics *44*, 981.

30. Demenais, F., Margaritte-Jeannin, P., Barnes, K.C., Cookson, W.O., Altmüller, J., Ang, W., Barr, R.G., Beaty, T.H., Becker, A.B., Beilby, J., et. al. (2018). Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. Nature genetics *50*, 42-53.

31. Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., Mora, S., et. al. (2013). Discovery and refinement of loci associated with lipid levels. Nature genetics *45*, 1274.

32. Liu, M., Jiang, Y., Wedow, R., Li, Y., Brazel, D.M., Chen, F., Datta, G., Davila-Velderrain, J., McGuire, D., Tian, C., et. al. (2019). Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nature genetics *51*, 237-244.

33. Lu, Y., Day, F.R., Gustafsson, S., Buchkovich, M.L., Na, J., Bataille, V., Cousminer, D.L., Dastani, Z., Drong, A.W., Esko, T., et. al. (2016). New loci for body fat percentage reveal link between adiposity and cardiometabolic disease risk. Nature communications *7*, 1-15.

34. Horikoshi, M., Beaumont, R.N., Day, F.R., Warrington, N.M., Kooijman, M.N., Fernandez-Tajes, J., Feenstra, B., Van Zuydam, N.R., Gaulton, K.J., Grarup, N., et. al. (2016). Genome-wide associations for birth weight and correlations with adult disease. Nature *538*, 248-252.

35. Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich, M.L., Yang, J., et. al. (2015). Genetic studies of body mass index yield new insights for obesity biology. Nature *518*, 197-206.

36. Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A.U., Wheeler, E., Glazer, N.L., Bouatia-Naji, N., Gloyn, A.L., et. al. (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nature genetics *42*, 105-16.

37. Evangelou, E., Warren, H.R., Mosen-Ansorena, D., Mifsud, B., Pazoki, R., Gao, H., Ntritsos, G., Dimou, N., Cabrera, C.P., Karaman, I., et. al. (2018). Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. Nature genetics *50*, 1412-1425.

38. Zhu, X., Li, X., Xu, R. and Wang, T. (2020). An iterative approach to detect pleiotropy and perform mendelian randomization analysis using GWAS summary statistics. Bioinformatics, https://doi.org/10.1093/bioinformatics/btaa985.

39. Chen J., and Chen Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. Biometrika *95*, 759-771.

40. Zhang, Y., and Shen, X. (2010). Model selection procedure for high-dimensional data. Statistical Analysis and Data Mining: The ASA Data Science Journal *3*, 350-358.

41. Wang, L., Kim, Y. and Li, R. (2013). Calibrating non-convex penalized regression in ultra-high dimension. Annals of statistics *41*, 2505-2536.

42. Shen, X., Pan, W., Zhu, Y., and Zhou, H. (2013). On constrained and regularized high-dimensional regression. Annals of the Institute of Statistical Mathematics *65*, 807-832.

43. Zhu, J., Wen, C., Zhu, J., Zhang, H. and Wang, X. (2020). A polynomial algorithm for best-subset selection problem. Proceedings of the National Academy of Sciences *117*, 33117-33123.

44. Hu, X., Zhao, J., Lin, Z., Wang, Y., Peng, H., Zhao, H., Wan, X. and Yang, C. (2021). MR-APSS: a unified approach to Mendelian Randomization accounting for pleiotropy and sample structure using genome-wide summary statistics. bioRxiv. https://doi.org/10.1101/2021.03.11.434915

45. Sanderson, E., Richardson, T., Hemani, G. and Smith, G.D. (2021). The use of negative control outcomes in Mendelian randomization to detect potential population stratification. International Journal of Epidemiology. doi: 10.1093/ije/dyaa288.

46. Zhao, Q., Chen, Y., Wang, J. and Small, D.S. (2019). Powerful three-sample genome-wide design and robust statistical inference in summary-data Mendelian randomization. International Journal of Epidemiology *48*, 1478-1492.

47. Wang, K. and Han, S. (2021). Effect of selection bias on two sample summary data based Mendelian randomization. Scientific reports *11*, 1-8.

48. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., Nicolae, D.L., GTEx Consortium, et. al. (2015). A gene-based association method for mapping traits using reference transcriptome data. Nature genetics *47*, 1091-1098.

49. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W., Jansen, R., De Geus, E.J., Boomsma, D.I., Wright, F.A., et. al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nature genetics *48*, 245-252.

50. Hu, Y., Li, M., Lu, Q., Weng, H., Wang, J., Zekavat, S.M., Yu, Z., Li, B., Gu, J., Muchnik, S., et. al. (2019). A statistical framework for cross-tissue transcriptome-wide association analysis. Nature genetics *51*, 568-576.

51. Xu, Z., Wu, C., Wei, P. and Pan, W. (2017). A Powerful Framework for Integrating eQTL and GWAS Summary Data. Genetics *207*, 893-902.

52. Zhu, Z., Zheng, Z., Zhang, F., Wu, Y., Trzaskowski, M., Maier, R., Robinson, M.R., McGrath, J.J., Visscher, P.M., Wray, N.R., et. al. (2018). Causal associations between risk factors and common diseases inferred from GWAS summary data. Nature communications *9*, 224.

53. Yang, Y., Shi, X., Jiao, Y., Huang, J., Chen, M., Zhou, X., Sun, L., Lin, X., Yang, C. and Liu, J. (2020). CoMM-S2: a collaborative mixed model using summary statistics in transcriptome-wide association studies. Bioinformatics *36*, 2009-2016.

54. Yuan, Z., Zhu, H., Zeng, P., Yang, S., Sun, S., Yang, C., Liu, J. and Zhou, X. (2020). Testing and controlling for horizontal pleiotropy with probabilistic Mendelian randomization in transcriptome-wide association studies. Nature communications *11*, 1-14.

55. Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A.N., Knowles, D.A., Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., et. al. (2019). Opportunities and challenges for transcriptome-wide association studies. Nature genetics *51*, 592-599.