Robust Mendelian Randomization Methods Based on Constrained Maximum Likelihood for Causal Inference

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Zhaotong Lin

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Advised by Dr. Wei Pan

June 2023

Abstract

Mendelian randomization (MR) has been increasingly applied for causal inference among traits, e.g. between potential risk factors and diseases, with observational data by using genetic variants as instrumental variables (IVs). Despite many successful MR applications, there are several gaps in the current literature to be filled. For example, only few (if any) MR methods can handle the violation of all IV assumptions, sample overlap in the GWAS data and/or linkage disequilibrium among IVs. And most of the MR applications only consider the total causal effect of one trait on the other. In this dissertation, we consider these important aspects to improve the robustness and effectiveness of MR. For the first project, we propose a two-step approach called Graph-MRcML, where we first apply an extended MR method to infer a causal network of total effects among multiple traits, then we modify a graph deconvolution algorithm to infer the corresponding network of direct effects. For the second project, we take a different route to consider multivariable MR, which includes multiple exposures in the model and estimates the direct effect of each exposure on the outcome while adjusting for possible mediating effects of other exposures. We propose an efficient and robust MVMR method based on constrained maximum likelihood, called MVMR-cML. For the third project, we move from polygenic MR to *cis*-MR, which uses correlated cis-variants from a single genomic region, compared to independent variants across the whole genome. A major difference is the need for taking into account linkage disequilibrium among cis-variants, for which we propose a robust cisMR-cML method. We conduct theoretical investigations, extensive simulations and real data applications to showcase the advantages of the three proposed methods in this work.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Mendelian randomization (MR) is an instrumental variable (IV) method which uses genetic variants as IVs (Davey Smith and Ebrahim, 2003; Zhu, 2021; Carter et al., 2021) to investigate the causal effects of (modifiable) risk factors on diseases. For an IV to be valid, it must satisfy the following three key assumptions:

(i) The IV is associated with the exposure;

(ii) The IV is independent of the unmeasured confounder;

(iii) The IV is independent of the outcome conditional on the exposure and the confounder.

Despite promising opportunities of applying MR for causal inference with the increasing availability of large genome-wide association study (GWAS) data, there are several gaps to be filled in MR applications. First, falling into the IV method framework, the validity of MR causal inference relies crucially on the three IV assumptions. Only the first IV assumption can be tested empirically and relatively easy to satisfy by using genome-wide significant single-nucleotide polymorphisms (SNPs) as IVs; while the second and the third assumptions are more likely to be violated in practice due to the prevalent genetic pleiotropy, including the so-called horizontal pleiotropy (i.e., the genetic effect on the outcome not mediated through the exposure). Second, most

of the existing MR methods assume the independence between the exposure GWAS samples and the outcome GWAS samples (Bowden et al., 2016; Qi and Chatterjee, 2019; Burgess et al., 2020; Xue et al., 2021). However, as more international consortia and large-scale biobanks emerging, it is inevitable to have overlapping samples between some GWAS datasets. Hence, there is an urgent need to develop new MR methods to handle the potential violation of IV assumptions as well as sample overlap between the two GWAS datasets. Third, the current practice of MR has been largely restricted to investigating the *total* causal effect of one trait on the other, while it would be useful to infer the *direct* causal effect after accounting for indirect or mediating effects through other traits). To fill these gaps, in Chapter 2, we propose a two-step framework called **Graph-MRcML** to infer a causal network among multiple traits. In the first step, we apply a robust MR method called MR-cML-C, which is an extension of the existing method proposed in Xue et al. (2021), on each pair of traits to infer a causal network of total effects among multiple traits. In particular, MR-cML-C is robust to the violation of all three IV assumptions, as well as the presence of overlapping samples in the two GWAS datasets. Then in the second step, we modify a graph deconvolution algorithm to infer the corresponding network of direct effects. By reconstructing both a total and a direct causal networks of 17 traits, including 11 common cardiometabolic risk factors and 6 diseases, we demonstrate the usefulness of our method.

As an alternative to Graph-MRcML, we take another route of multivariable MR (MVMR) in Chapter 3. Instead of having only one exposure and one outcome in the analysis as in univariable MR (UVMR) (as used in the first step in Graph-MRcML), MVMR includes multiple exposures and one outcome in the model. By accounting for multiple potential pathways (from the genetic variant to the outcome), MVMR is more robust to the notorious horizontal pleiotropy to a certain degree, and at the same time, the direct effect of one exposure on the outcome conditional on the

remaining exposures is estimated in the MVMR analysis (Burgess and Thompson, 2015). In Chapter 3, we propose a multivariable version of UVMR-cML (which is used in the first step of Graph-MRcML), called **MVMR-cML**. As its univariable counterpart, MVMR-cML is robust to the violation of IV assumptions, as well as the sample overlap among multiple GWAS datasets used in the analysis. Through extensive simulation studies as well as a real data application on 8 cardiometabolic risk factors and coronary artery disease, we demonstrate the difference between UVMR and MVMR, as well as the effectiveness of our new approach.

While conventional MR applications use independent IVs selected from the whole genome, including the proposed UVMR-cML and MVMR-cML, there has been a growing interest in MR studies focusing on a small genomic region using some local and correlated cis-SNPs as IVs, known as *cis*-MR. One of the most promising applications of *cis*-MR is for drug target discovery, including drug target prioritization, validation or drug repositioning (Schmidt et al., 2020; Zhao et al., 2022; Gkatzionis et al., 2023). In Chapter 4, we propose a *cis*-MR method called **cisMR-cML**, which extends the UVMR-cML to allow for correlated IVs. It again enjoys the nice statistical properties and the robustness to the presence of invalid IVs. Through simulation studies and a drug-target application for coronary artery disease, we demonstrate the proposed method to be a valuable addition to the toolbox of *cis*-MR analysis.

The work presented in Chapter 2 was accepted by *PLoS Genetics*, the work presented in Chapter 3 was just published in *American Journal of Human Genetics* (Lin et al., 2023), and the research in Chapter 4 is still ongoing.

Chapter 2

Combining Mendelian randomization and network deconvolution for inference of causal networks with GWAS summary data

2.1 Introduction

A fundamental task in science is to understand causal pathways among various risk factors and diseases. This is particularly challenging with observational data due to the likely presence of hidden confounding, implying that an observed association is not equivalent to a causation. In our real data example, we'd like to infer which of some known risk factors are causal to coronary artery disease (CAD). While many previous studies have established for example that obesity is associated with CAD (Khan et al., 2018), whether it is causal, especially independent of other known risk factors, is still debatable with conflicting results from observational studies (Powell-Wiley et al., 2021). Mendelian randomization (MR) is a powerful tool to infer causal relationship between two traits in the presence of unmeasured confounding, by using single nucleotide polymorphisms (SNPs) as instrumental variables (IVs) (Sanderson

et al., 2022; Zhu, 2021; Lawlor et al., 2008). A distinct and useful feature of MR is its applicability when the two traits come from two different genome-wide association study (GWAS) summary datasets. The conventional MR analysis usually assumes the causal direction is known from an exposure to an outcome. When the direction is not clear, bidirectional MR can be applied (Timpson et al., 2011; Brower et al., 2019). However, such a causal estimate only reflects the *total* causal effect from one trait to the other, which consists of possibly both a direct effect and an indirect effect mediated through other factors (Burgess et al., 2015; Burgess and Thompson, 2015; Carter et al., 2021; Wang, 2018). In our motivating real data example, we'd like to estimate causal relationships among multiple common risk factors and diseases; we are not only interested in a total effect of a risk factor, say obesity/BMI, on a disease, say CAD, but also its direct effect after accounting for possible mediating effects through other risk factors. In addition, in general we do not want to pre-specify any causal directions because, for example, there may be a bidirectional relationship between BMI and CAD. For this purpose, we propose a two-step framework to infer both total and direct causal networks, allowing bi-directional relationships (i.e. cycles). In the first step, we apply bidirectional MR on every pair of traits to construct a *total* causal (effect) graph, depicting the total causal effect from one node to the other. In the second step, we apply network deconvolution (Feizi et al., 2013) to the (estimated) total causal network to estimate the *direct* causal (effect) graph, each edge of which measures the direct effect of one node on the other after accounting for mediating effects through other nodes in the graph.

In principle, any bidirectional MR method could be used in the first step. However, the inference of the direct causal graph depends crucially on the validity of the estimated total causal effects in the first step, which relies on the three key IV assumptions in MR: (i) Relevance assumption - IVs are associated with the exposure; (ii) Independence assumption - IVs are independent of unmeasured confounding; (iii)

Exclusion restriction - IVs affect the outcome only through the exposure. However, these assumptions may be violated due to the pervasive horizontal pleiotropy (Cheng et al., 2022b; Dai et al., 2018). Under the plurality assumption (that the valid IVs form the largest group of IVs sharing the same causal parameter value), MR-cML is robust to the presence of some invalid IVs violating any or all of three IV assumptions and has been shown to perform better than many existing methods under various scenarios Xue et al. (2021). Furthermore, as shown before Xue and Pan (2022), with a simple IV screening procedure, MR-cML achieves good performance in inferring both causal directions and effect sizes between two traits while allowing bidirectional relationships (i.e. either trait is causal to the other at the same time). Thus, we will apply MR-cML in our causal graph framework, called **Graph-MRcML**.

One limitation of the original MR-cML is its implementation only for two-sample MR (i.e., assuming two independent GWAS summary datasets) (Xue et al., 2021). However, in practice, multiple traits may come from the same study, as several lipid traits from the Global Lipids Genetics Consortium GWAS data to be used in our real data example (Willer et al., 2013). More generally, as more international consortia and large-scale biobanks emerging, it is inevitable to have overlapping samples between some GWAS datasets. It has been shown that sample overlap may lead to biased estimates and inflated type-I errors in MR Burgess et al. (2016a). To address this, we first extend MR-cML to the overlapping-sample set-up, which turns out to be non-trivial, especially with respect to valid statistical inference. In addition, we establish theory that, perhaps surprisingly, the bias of the causal parameter estimator under the incorrect independence assumption (i.e. ignoring sample overlap) will disappear asymptotically (as the sample size increases); however, the usual (model-based) variance will be biased, thus we propose a robust/sandwich estimator. More importantly, the causal parameter estimator fully accounting for sample overlap is more efficient than the one under the working independence assumption. It is em-

phasized that, as a distinct feature, our proposed method not only estimates causal networks, but also can assess the statistical significance of any estimated causal effects. For this purpose, in addition to developing statistical theory for large-sample inference, we also develop a novel and effective data perturbation scheme for more accurate finite-sample inference by accounting for model fitting uncertainties (e.g. in selecting out invalid IVs). The latter task is technically challenging mainly because of the presence of some complex dependencies among the traits and the SNPs: we need to fully take into account of not only possible correlations among the traits (due to overlapping samples), but also each trait's being used multiple times across many pairs of traits (thus inducing dependencies among the resulting estimates in a causal network) and linkage disequilibrium (LD) among the SNPs/IVs across all the traits (even if the SNPs/IVs are selected as independent for each trait). In particular, it would be impractical to restrict the SNPs/IVs to be independent across all the traits, leading to no or few SNPs.

There are several approaches in the MR literature aiming to estimate the direct causal effects among multiple traits. Brown and Knowles (2020) proposed a two-step framework similar to ours, which used MR-Egger to construct a causal network of total effects, then under the sparsity assumption approximately invert it by penalized regression to infer the corresponding direct causal network. Besides the difference of our using more robust and efficient MR-cML versus their (modified) MR-Egger, no theory of their method is established; in particular, it is unclear how their proposed statistical inference would perform, partly due to technical challenges imposed by their using penalized regression. Another related method is two-step MR (Relton and Davey Smith, 2012) or network MR (Burgess et al., 2015), which focuses on the set-up with a candidate mediator between an exposure and an outcome. Our proposed method can be regarded as a generalization of this approach to inferring a more complex causal network of multiple traits without pre-specifying causal directions

and mediators. Finally, multivariable MR (MVMR) (Burgess and Thompson, 2015; Carter et al., 2021) can be used to estimate direct effects of multiple exposures on an outcome. However, first, our method depends only on the validity of univariable MR (UVMR) (and the corresponding valid IV assumptions), while there are additional assumptions required for MVMR (Lin et al., 2023). For example, a valid IV for UVMR may not be valid for MVMR , and there is a potential issue of multicollinearity in MVMR, leading to weak IV biases (Sanderson et al., 2021). Second, existing MVMR methods all require the use of independent IVs for all exposures, sometimes leading to no or only few IVs for some exposures if the number of exposures is not too small. More generally, application of any existing MVMR method would reduce the number of the IVs, leading to loss of estimation efficiency and the possible issue of multicollinearity as to be confirmed in the real data example.

To summarize, we have two main contributions in methods development. First, we propose a general framework for inferring (including estimating and testing) both total and direct causal graphs among multiple traits of interest. Second, for better performance of the proposed framework, we extend the MR-cML method Xue et al. (2021) to accommodate overlapping samples, and modify the network deconvolution algorithm, either of which can be useful in their own applications. Through extensive simulation studies, we show that the extended MR-cML performed better than the original one and other widely-used MR methods in the presence of sample overlap. We also show improved performance of our modified network deconvolution algorithm over that of the original one. Finally, we applied the proposed framework to 17 large-scale GWAS summary datasets (with median sample size of 254892 and median 45 IVs) to infer causal networks among 11 common cardiometabolic risk factors and 6 diseases, including 4 cardiometabolic diseases (coronary artery disease, stroke, type 2 diabetes, atrial fibrillation), Alzheimer's disease (AD) (for its associations with some cardiometabolic risk factors/diseases (Arega and Shao, 2022)) and asthma (more as

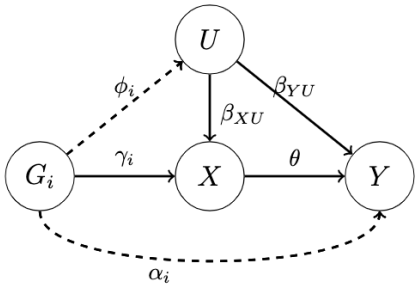


Figure 2.1: A causal model for the exposure X and the outcome Y.

a negative control), identifying some interesting causal pathways.

2.2 Methods

2.2.1 Causal model

Based on Figure 2.1, we have

$$U = \phi_i G_i + \epsilon_U, \quad X = \gamma_i G_i + \beta_{XU} U + \epsilon_X, \quad Y = \theta X + \beta_{YU} U + \alpha_i G_i + \epsilon_Y,$$

where $\epsilon_U, \epsilon_X, \epsilon_Y$ are independent random errors. We can further express the exposure X and the outcome Y as

$$\begin{aligned} X &= (\gamma_i + \beta_{XU}\phi_i)G_i + (\beta_{XU}\epsilon_U + \epsilon_X) := b_{Xi}G_i + \epsilon_X^*, \\ Y &= (\theta b_{Xi} + \beta_{YU}\phi_i + \alpha_i)G_i + (\theta\epsilon_X^* + \beta_{YU}\epsilon_U + \epsilon_Y) := b_{Yi}G_i + \epsilon_Y^*. \end{aligned}$$

Accordingly, we have

$$b_{Xi} = \gamma_i + \beta_{XU}\phi_i, \tag{2.1}$$

$$b_{Yi} = \theta b_{Xi} + \beta_{YU}\phi_i + \alpha_i := \theta b_{Xi} + r_i, \tag{2.2}$$

where b_{Xi} and b_{Yi} are the marginal effects of SNP G_i on the exposure X and the outcome Y respectively, θ is the causal effect of interest, and r_i represents all pleiotropic effects on Y (but not through X), including both uncorrelated pleiotropy (i.e. α_i) and correlated pleiotropy (i.e. $\beta_{YU}\phi_i$ that is correlated with b_{Xi}). A SNP i is an invalid IV if $r_i \neq 0$ (or $b_{Xi} = 0$).

Given the GWAS summary data $\{(\hat{\beta}_{Xi}, \hat{\sigma}_{Xi}, \hat{\beta}_{Yi}, \hat{\sigma}_{Yi}) : i = 1, \ldots, m\}$, where $\hat{\beta}_{Xi}$ and $\hat{\beta}_{Yi}$ are the estimates of b_{Xi} and b_{Yi} with the corresponding standard errors $\hat{\sigma}_{Xi}$ and $\hat{\sigma}_{Yi}$ respectively, the central goal of robust MR is to infer θ in the presence of invalid IVs.

2.2.2 Extension of MR-cML to overlapping samples

In this section, we consider the situation when there are overlapping samples between two GWAS datasets, and the m IVs/SNPs are independent. It is reasonable to assume

$$\begin{pmatrix} \hat{\beta}_{Xi} \\ \hat{\beta}_{Yi} \end{pmatrix} \sim \mathcal{N}\Big(\begin{pmatrix} b_{Xi} \\ \theta b_{Xi} + r_i \end{pmatrix}, \begin{pmatrix} \sigma_{Xi}^2 & \rho\sigma_{Xi}\sigma_{Yi} \\ \rho\sigma_{Xi}\sigma_{Yi} & \sigma_{Yi}^2 \end{pmatrix} \Big),$$

where ρ captures the correlation between $\hat{\beta}_{Xi}$ and $\hat{\beta}_{Yi}$ due to overlapping samples (and possibly other reasons) in the two GWAS datasets. We assume it is known or can be well-estimated; we will discuss its estimation later. We also assume σ_{Xi}^2 and σ_{Yi}^2 are known or well-estimated as $\hat{\sigma}_{Xi}^2$ and $\hat{\sigma}_{Yi}^2$ respectively. Then the log-likelihood (up to some constants) is

$$\begin{aligned} l(\theta, \{b_{Xi}, r_i\}; \{\hat{\beta}_{Xi}, \sigma_{Xi}, \hat{\beta}_{Yi}, \sigma_{Yi}\}, \rho) = \sum_{i=1}^{m} l_i(\theta, b_{Xi}, r_i; \hat{\beta}_{Xi}, \sigma_{Xi}, \hat{\beta}_{Yi}, \sigma_{Yi}, \rho) = \\ -\frac{1}{2(1-\rho^2)} \sum_{i=1}^{m} \Big(\frac{(\hat{\beta}_{Xi} - b_{Xi})^2}{\sigma_{Xi}^2} + \frac{(\hat{\beta}_{Yi} - \theta b_{Xi} - r_i)^2}{\sigma_{Yi}^2} - 2\rho \frac{(\hat{\beta}_{Xi} - b_{Xi})}{\sigma_{Xi}} \frac{(\hat{\beta}_{Yi} - \theta b_{Xi} - r_i)}{\sigma_{Yi}} \Big), \end{aligned} \tag{2.3}$$

where we use $\{b_{Xi}, r_i\} = \{(b_{Xi}, r_i) : i = 1, \ldots, m\}$ to represent a set of the parameters, and similarly for $\{\hat{\beta}_{Xi}, \sigma_{Xi}, \hat{\beta}_{Yi}, \sigma_{Yi}\}$. With Eq (2.3), we obtain the constrained maximum likelihood estimator (cMLE) by solving

$$\min_{\theta, b_{Xi}, r_i} -l(\theta, \{b_{Xi}, r_i\}; \{\hat{\beta}_{Xi}, \sigma_{Xi}, \hat{\beta}_{Yi}, \sigma_{Yi}\}, \rho) \text{ subject to } \sum_{i=1}^{m} I(r_i \neq 0) = K, \tag{2.4}$$

where $I(\cdot)$ is the indicator function, K is a tuning parameter representing the unknown number of invalid IVs. As MR-cML requires the plurality condition (see Assumption A.2 in Appendix A.1.1), there should be at least two valid IVs, i.e., K can be ranged from 0 to $m-2$. We note that when $\rho = 0$, it becomes the method proposed by Xue et al. (2021). We refer this extended version of MR-cML to **MR-cML-C** (C for correlated samples), and the original version in Xue et al. (2021) as **MR-cML-I** (I for independence) obtained under the (incorrect) working independence assumption of $\rho = 0$.

To infer a possibly bi-directional causal relationship between a pair of traits, we will apply bi-directional MR with an extra step as proposed in Xue and Pan (2022) to screen for valid IVs (see Appendix A.4.1).

2.2.2.1 Estimation of causal parameter and its standard error with fixed K

Equating the first-order derivatives of log-likelihood (Eq (2.3)) to zero gives:

$$\hat{r}_i(\theta, b_{Xi}) = (\hat{\beta}_{Yi} - \theta b_{Xi}) - \rho(\hat{\beta}_{Xi} - b_{Xi})\sigma_{Yi}/\sigma_{Xi}, \tag{2.5a}$$

$$\hat{b}_{Xi}(\theta, r_i) = \frac{\hat{\beta}_{Xi}/\sigma_{Xi}^2 - \rho(\hat{\beta}_{Yi} - r_i + \theta\hat{\beta}_{Xi})/(\sigma_{Xi}\sigma_{Yi}) + \theta(\hat{\beta}_{Yi} - r_i)/\sigma_{Yi}^2}{1/\sigma_{Xi}^2 - 2\rho\theta/(\sigma_{Xi}\sigma_{Yi}) + \theta^2/\sigma_{Yi}^2}, \tag{2.5b}$$

$$\hat{\theta}(\{b_{Xi}, r_i\}) = \frac{\sum_{i=1}^m \left[b_{Xi}(\hat{\beta}_{Yi} - r_i)/\sigma_{Yi}^2 - \rho b_{Xi}(\hat{\beta}_{Xi} - b_{Xi})/(\sigma_{Xi}\sigma_{Yi})\right]}{\sum_{i=1}^m b_{Xi}^2/\sigma_{Yi}^2}. \tag{2.5c}$$

For a given number of invalid IVs ($0 < K \leq m-2$), we use a coordinate descent-like algorithm to iteratively solve Eq (2.4). At the $(t+1)$th iteration:

Step 1: Calculate $r_i^{(t+1)} = (\hat{\beta}_{Yi} - \theta^{(t)} b_{Xi}^{(t)}) - \rho(\hat{\beta}_{Xi} - b_{Xi}^{(t)})\sigma_{Yi}/\sigma_{Xi}$; In order to select out K invalid IVs, we choose them as the ones with the largest

$$d_i^{(t+1)} := l_i(\theta^{(t)}, b_{Xi}^{(t)}, r_i^{(t+1)}; \hat{\beta}_{Xi}, \sigma_{Xi}, \hat{\beta}_{Yi}, \sigma_{Yi}, \rho) - l_i(\theta^{(t)}, b_{Xi}^{(t)}, 0; \hat{\beta}_{Xi}, \sigma_{Xi}, \hat{\beta}_{Yi}, \sigma_{Yi}, \rho),$$

so that the log-likelihood is maximally increased. Specifically, we order $d_i^{(t+1)}$ decreasingly, then for $i = 1, \ldots, K$, let $r_{(i)}^{(t+1)} = \hat{r}_{(i)}(\theta^{(t)}, b_{X(i)}^{(t)})$ (Eq (2.5a)); for $j = K+1, \ldots, m$, let $r_{(i)}^{(t+1)} = 0$.

Step 2: Update b_{Xi} and θ using Eq (2.5b) and Eq (2.5c): $b_{Xi}^{(t+1)} = \hat{b}_{Xi}(\theta^{(t)}, r_i^{(t+1)})$, $\theta^{(t+1)} = \hat{\theta}(\{b_{Xi}^{(t+1)}, r_i^{(t+1)}\})$.

We repeat the above two steps until convergence, obtaining the final estimates $\hat{\theta}(K)$ and $\{\hat{r}_i(K), \hat{b}_{Xi}(K)\}_{i=1}^m$. It is noted that at the convergence the (estimated) invalid IVs (with $\hat{r}_i \neq 0$) do not contribute to estimating θ. We use the observed Fisher information to estimate the standard error (SE) of $\hat{\theta}(K)$.

2.2.2.2 Model selection and data perturbation

Following Xue et al. (2021), we use BIC to select the set of invalid IVs. We denote $B_0 = \{i|r_i \neq 0, i = 1, \ldots, m\}$ the set of truly invalid IVs, with size $|B_0| = K_0$. Denote the cMLEs obtained from Eq (2.4) as $\hat{\theta}(K)$, $\hat{b}_{Xi}(K)$, and $\hat{r}_i(K)$ for $i = 1, \ldots, m$, and $\hat{B}_K = \{i|\hat{r}_{Xi}(K) \neq 0, i = 1, \ldots, m\}$ the estimated set of invalid IVs. We estimate the K from a candidate set $\mathcal{K}$ based on the following Bayesian information criterion (BIC):

$$\text{BIC}(K) = -2l(\hat{\theta}(K), \{\hat{b}_{Xi}(K), \hat{r}_i(K)\}; \{\hat{\beta}_{Xi}, \sigma_{Xi}, \hat{\beta}_{Yi}, \sigma_{Yi}\}, \rho) + log(N) \cdot K, \tag{2.6}$$

where $N = \min(N_1, N_2)$. We select $\hat{K} = \arg\min_{K \in \mathcal{K}} \text{BIC}(K)$ and $\hat{B}_{\hat{K}} = \{i|\hat{r}_{Xi}(\hat{K}) \neq 0, i = 1, \ldots, m\}$. The final estimate $\hat{\theta} = \hat{\theta}(\hat{K})$ and its estimated standard error $\widehat{\text{SE}}(\hat{\theta}(\hat{K}))$ are used to perform inference on θ. We refer this method to **MR-cML-BIC-C**.

The proposed MR-cML-BIC-C is based on the (consistently) selected set of valid IVs, ignoring inherent uncertainty in model selection, thus tending to underestimate standard errors for finite samples. To better account for model selection uncertainty, especially with a small (to medium) sample size, we adopt the data perturbation approach (Xue et al., 2021; Xue and Pan, 2022), which is equivalent to bootstrapping the corresponding GWAS individual-level data (Lin et al., 2021). Briefly, for the $b-$th perturbation, $b = 1, \ldots, B$, we generate perturbed samples

$$\begin{pmatrix} \hat{\beta}_{Xi}^{(b)} \\ \hat{\beta}_{Yi}^{(b)} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \hat{\beta}_{Xi} \\ \hat{\beta}_{Yi} \end{pmatrix}, \begin{pmatrix} \sigma_{Xi}^2 & \rho\sigma_{Xi}\sigma_{Yi} \\ \rho\sigma_{Xi}\sigma_{Yi} & \sigma_{Yi}^2 \end{pmatrix} \right),$$

for $i = 1, \ldots, m$ independently. Then the remaining steps follow: we apply MR-cML-BIC-C on each perturbed dataset to obtain $\hat{\theta}^{(b)}$, and we use the sample mean and

standard deviation over the B estimates, $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}$, from the B perturbed datasets as the final estimate and its standard error respectively. We call this method **MR-cML-DP-C**.

We use a generic notation **MR-cML-C**, referring either MR-cML-BIC-C or MR-cML-DP-C; we use similar notations for **MR-cML-I**. We also use MR-cML to denote either MR-cML-C or MR-cML-I.

2.2.2.3 Estimation of ρ

The correlation between the GWAS estimates of two (continuous) traits X and Y is given by $\rho = \frac{N_0}{\sqrt{N_1 N_2}} r(\mathbf{x}, \mathbf{y})$, where N_0 is the sample size of the overlapping samples and $r(\mathbf{x}, \mathbf{y})$ is the phenotypic correlation between the two traits, which can be estimated based on completely overlapped individual-level data (Eq.(4) and (5) in Li et al. (2021b), Eq. (7) in LeBlanc et al. (2018)). Without individual-level data, two commonly-used strategies to estimate ρ are (i) using the correlation between the two sets of GWAS null Z-scores (Kim et al., 2015); and (ii) using the intercept from a fitted bivariate LD-score regression model (LDSC) (Turley et al., 2018). We note that, while our motivation is to take into account of sample overlap between the two GWAS studies, other relevant sources for correlations can also be captured, including not only sample overlap but also population stratification and cryptic relatedness (Bulik-Sullivan et al., 2015).

2.2.3 Graph-MRcML

Now instead of considering two traits, suppose we have T traits/diseases, say $Y_1, \ldots, Y_T$, and the goal is to construct a causal network among them. This can be done by first constructing an *total* causal graph ($\mathbf{G}_{tot}$), and second deconvoluting $\mathbf{G}_{tot}$ into the *direct* causal graph ($\mathbf{G}_{dir}$). Briefly, in the first step, we apply bi-directional MR-cML-C

on *every* pair of traits and obtain the total causal graph $\mathbf{G}_{tot}$. However, such a graph may contain both direct and indirect causal effects, and we'd like to distinguish them to better understand the causal paths among the T traits. Therefore, in the second step, we use a network deconvolution method Feizi et al. (2013) to estimate the direct causal graph $\mathbf{G}_{dir}$.

We also extend a data perturbation scheme for statistical inference on such graphs. Briefly, we perturb the GWAS summary data for the T traits multiple times and obtain the estimated total and direct causal (effect) graphs with each perturbed dataset. Then the empirical distribution of such estimates from multiple perturbed samples is used for inference. To account for multiple testing, we use the Bonferroni adjustment with the effective number of independent tests estimated as in Li et al. (2011) (see Appendix A.4.2).

2.2.3.1 Using MR-cML for estimation and inference of $\mathbf{G}_{tot}$

Let $\widehat{\mathbf{B}}$ and $\mathbf{S}$ denote the $m \times T$ matrices of GWAS summary statistics for the T traits: each entry $\widehat{\mathbf{B}}_{i,j}$ and $\mathbf{S}_{i,j}$ are the estimated association effect size and standard error between the i-th SNP and the trait Y_j respectively. Note that some $\widehat{\mathbf{B}}_{i,j}$ (and $\mathbf{S}_{i,j}$) can be missing if the i-th SNP is never used in the analyses involving the j-th trait. We use $\widehat{\mathbf{B}}_{i,\cdot}$ and $\widehat{\mathbf{B}}_{\cdot,j}$ to denote the i-th row and the j-th column of matrix $\widehat{\mathbf{B}}$ respectively. Let $\mathbf{P}$ denote the $T \times T$ correlation matrix, i.e. $\mathbf{P}_{j,j} = 1$ and $\mathbf{P}_{j,k} = \mathbf{P}_{k,j} = \rho_{jk}$ for $j \neq k$, where ρ_{jk} is the correlation between the GWAS estimates for traits j and k.

The algorithm consists of two parts. In the first part we perform data perturbation on the GWAS data, and in the second part we estimate the total causal (effect) network $\mathbf{G}_{tot}$ with the perturbed data. Instead of using the data perturbation scheme mentioned in Section 2.2.2.2 for each pair separately, here we will perturb the summary statistics for *all* traits together (i.e. the whole matrix $\widehat{\mathbf{B}}$). The reason is that, there might be correlations among the SNPs/IVs (i.e. the rows of $\widehat{\mathbf{B}}$) besides the

correlations among the GWAS traits (i.e. the columns of $\widehat{\mathbf{B}}$), and we need to take the correlations into account. For example, let's say we have GWAS summary data for three traits, HDL, LDL and TG, which may come from the same consortium (i.e. with overlapping samples). Then *between* the set of IVs for HDL→LDL and the set of IVs for LDL→TG, there might be SNPs in linkage disequilibrium (LD), though *within* the two sets, IVs were independent as selected for MR in practice. For this purpose, we use a matrix normal distribution (Kwak and Pan, 2017) to model and perturb the data. Let $\mathbf{Z} = \widehat{\mathbf{B}}/\mathbf{S}$ denote the matrix of Z-scores, then for the b-th perturbed dataset, $\mathbf{Z}^{(b)} = \mathbf{Z} + \mathbf{E}^{(b)}, \widehat{\mathbf{B}}^{(b)} = \mathbf{Z}^{(b)} * \mathbf{S}$, where $*$ is the element-wise multiplication and $\mathbf{E}^{(b)}$ follows a matrix normal distribution:

$$\mathbf{E}^{(b)} \sim \mathcal{MN}_{m,T}(\mathbf{0}_{m\times T}, \mathbf{R}, \mathbf{P}),$$

where $\mathbf{R}$ is the LD matrix of the m SNPs. Or equivalently, we have $\text{vec}(\mathbf{E}^{(b)}) \sim \mathcal{N}(\mathbf{0}_{mT}, \mathbf{P} \otimes \mathbf{R})$, where $\otimes$ denotes the Kronecker product and $\text{vec}(\mathbf{E}^{(b)})$ denotes the vectorization of $\mathbf{E}^{(b)}$. To generate $\mathbf{E}^{(b)}$, we first generate an mT-vector $\mathbf{v}$ from a standard normal distribution, and $\text{vec}(\mathbf{E}^{(b)}) = \mathbf{A}\mathbf{v}$, where $\mathbf{A}\mathbf{A}^T = \mathbf{P} \otimes \mathbf{R}$ and the matrix decomposition is done with eigen decomposition. Then we convert $\text{vec}(\mathbf{E}^{(b)})$ back to $\mathbf{E}^{(b)}$ and obtain the perturbed GWAS data $\widehat{\mathbf{B}}^{(b)} = \widehat{\mathbf{B}} + \mathbf{S} * \mathbf{E}^{(b)}$. In practice, we use the 1703 approximately independent LD blocks (Berisa and Pickrell, 2016), and extract the LD matrix $\mathbf{R}$ using the 1000 Genomes Phase 3 EUR population as the reference panel in `TwoSampleMR` (Hemani et al., 2018b). The algorithm is summarized as follows:

Step 1: Generate the perturbed data $\widehat{\mathbf{B}}^{(b)}$ as described above.

Step 2: Apply bidirectional MR-cML-BIC-C on every pair of traits to obtain (all the non-diagonal entries in) $\mathbf{G}_{tot}^{(b)}$.

We repeat the above steps B times and use the element-wise mean of $\{\mathbf{G}_{tot}^{(b)}\}_{b=1}^{B}$ as the final estimate for $\mathbf{G}_{tot}$, i.e., $\widehat{\mathbf{G}}_{tot} = \sum_{b=1}^{B} \mathbf{G}_{tot}^{(b)}/B$. We also use the element-wise standard deviation of $\{\mathbf{G}_{tot}^{(b)}\}_{b=1}^{B}$ to estimate the standard error for each entry in $\widehat{\mathbf{G}}_{tot}$. Lastly, the p-value for each (off-diagonal) entry of $\widehat{\mathbf{G}}_{tot}$ (or the edge in the corresponding network) (to test for the entry being 0 or the edge is absent) is calculated based on the standard normal distribution.

2.2.3.2 Using network deconvolution for estimation and inference of $\mathbf{G}_{dir}$

Under the assumption that the causal relationships between variables are linear, there is a relationship between $\mathbf{G}_{tot}$ and $\mathbf{G}_{dir}$ (Feizi et al., 2013):

$$\mathbf{G}_{tot} = \mathbf{G}_{dir} + \mathbf{G}_{dir}^2 + \mathbf{G}_{dir}^3 + \cdots = \mathbf{G}_{dir}(\mathbf{I} + \mathbf{G}_{dir} + \mathbf{G}_{dir}^2 + \mathbf{G}_{dir}^3 + \dots) = \mathbf{G}_{dir}(\mathbf{I} - \mathbf{G}_{dir})^{-1}, \tag{2.7}$$

where $\mathbf{I}$ is the identity matrix. The first equality in Eq (2.7) is by definition with an intuitive interpretation: a total effect represented by an element (i.e. edge) in $\mathbf{G}_{tot}$ can be decomposed into a direct effect represented by the corresponding edge in $\mathbf{G}_{dir}$ and the sum of all indirect effects mediated through one, two, $\cdots$, up to (infinitely) many nodes (due to possible cycles) as represented by the corresponding edges in $\mathbf{G}_{dir}^2, \mathbf{G}_{dir}^3, \cdots$ (Bentler and Freeman, 1983). An illustrative example is given in Appendix A.4.3. If and only if the spectral radius (i.e. the largest absolute value of all real/complex eigenvalues) of $\mathbf{G}_{dir}$ is less than 1, the third equality in Eq (2.7) holds Young (1981); Bentler and Freeman (1983). Then it is easy to show

$$\mathbf{G}_{dir} = \mathbf{G}_{tot}(\mathbf{I} + \mathbf{G}_{tot})^{-1}. \tag{2.8}$$

Note that in $\mathbf{G}_{tot}$, only the off-diagonal elements are estimated by bidirectional MR. For the diagonal elements, we may follow the practice in Feizi et al. (2013) of

setting them to zeros, and we refer this approach to **Graph-MRcML-d0**. This is correct if there is no cycle in the underlying direct graph. However, with cycles in $\mathbf{G}_{dir}$, in general the corresponding diagonal elements in $\mathbf{G}_{tot}$ are not zeros. To see this, we rewrite Eq (2.7) as

$$\mathbf{G}_{tot} = \mathbf{G}_{dir} + \mathbf{G}_{dir}(\mathbf{G}_{dir} + \mathbf{G}_{dir}^2 + \mathbf{G}_{dir}^3 + \dots) = \mathbf{G}_{dir} + \mathbf{G}_{dir}\mathbf{G}_{tot} = \mathbf{G}_{dir}(\mathbf{I} + \mathbf{G}_{tot}).$$

Denote the adjacency matrices $\mathbf{G}_{tot} = (T_{ij})$ and $\mathbf{G}_{dir} = (D_{ij})$, then the i-th diagonal element of $\mathbf{G}_{tot}$ can be expressed as

$$T_{ii} = \sum_j D_{ij}(I(j = i) + T_{ji}) = \sum_{j \neq i} D_{ij} T_{ji}, \tag{2.9}$$

where the second equality follows from the assumption that there is no self-loop in the direct graph (i.e. $D_{ii} = 0$). Accordingly, we propose a heuristic approach to specify the diagonal elements of $\mathbf{G}_{tot}$. We obtain the initial estimate of $\mathbf{G}_{dir}$ by setting $\hat{T}_{ii} = \sum_{j \neq i} T_{ij} T_{ji}$. Then we update $\hat{T}_{ii}$ and $\widehat{\mathbf{G}}_{dir}$ iteratively based on Eq (2.9) and Eq (2.8) till convergence. If it fails to converge, we will use the initial estimate (by setting $\hat{T}_{ii} = \sum_{j \neq i} T_{ij} T_{ji}$, which can serve as a good approximation in some scenarios, e.g. when T_{ij} is close to D_{ij}). We refer this approach to **Graph-MRcML-d1**. As in Section 2.2.3.1, we leverage data perturbation to obtain a final estimate and perform statistical inference of the direct causal graph. More discussions and illustrative examples are given in Appendix A.4.4.

There are some criticisms of the original network deconvolution paper (Feizi et al., 2013); see https://liorpachter.wordpress.com/2014/02/11/the-network-nonsense-of-manolis-kellis/. First of all, the key idea of network deconvolution is in Eq (2.7), based on a well-established and widely-used definition of the total effects in terms of the direct effects in the literature of linear structural equation modeling for directed causal

graphs (Bentler and Freeman, 1983). Second, we only consider smaller directed graphs with the total effects estimated by UVMR with much larger sample sizes (Burgess and Thompson, 2015; Carter et al., 2021). Hence the concern on relatively poor performance of network deconvolution compared to Gaussian graphical modeling (i.e. using the correlation and partial correlation to estimate the total and direct effects in the undirected graphs respectively for high-dimensional data) is not relevant here. Third, due to the differences between our and the original implementations, other main criticisms are not applicable here. Specifically, we do not scale, threshold or symmetrize the total effect graph $\mathbf{G}_{tot}$, hence there are no corresponding parameters (and their tuning). Instead of applying eigen-decomposition to the total graph, we invert the matrix in (Eq (2.8)) directly. We do acknowledge that the method requires the assumption that the spectral radius of the direct causal graph is less than 1, which may be violated in practice. However, as to be shown in the numerical examples, our method performed well (without encountering the spectral radius issue).

2.2.4 Theory

As shown in Appendix A.1, our proposed methods enjoy some desirable statistical properties: under mild conditions, the BIC can consistently select valid IVs; both the MR-cML-BIC-I and MR-cML-BIC-C estimators are consistent and asymptotically normal for the true causal parameter θ; and the Graph-MRcML estimators are consistent and asymptotically normal for the true total and direct causal effect graphs.

2.2.5 Simulation for MR with sample overlap

In this simulation, we investigated the performance of different MR methods in the presence of overlapping samples. We simulated data as follows:

$$\mathbf{U} = \mathbf{G}\boldsymbol{\phi} + \boldsymbol{\epsilon}_U, \quad \mathbf{X} = \gamma \cdot \mathbf{G} + \mathbf{U} + \boldsymbol{\epsilon}_X, \quad \mathbf{Y} = \theta \cdot \mathbf{X} + \mathbf{G}\boldsymbol{\alpha} + \mathbf{U} + \boldsymbol{\epsilon}_Y,$$

where $\boldsymbol{\epsilon}_U, \boldsymbol{\epsilon}_X, \boldsymbol{\epsilon}_Y$ were generated from $\mathcal{N}(0,1)$ independently. 20 IVs were generated independently from a binomial distribution with minor allele frequency (MAF) 0.3 and the IV strength γ was set to 0.08 for all IVs. We considered 0% and 30% invalid IVs. In the case of 0% invalid IVs, the set-up was the same as that in Burgess et al. (2016a). In the case of 30% invalid IVs, we generated the direct effect $\boldsymbol{\alpha}$ iid from $\mathcal{N}(0.04, 0.05^2)$, and considered (i) uncorrelated pleiotropy (i.e., $\boldsymbol{\phi} = 0$) and (ii) correlated pleiotropy, where $\boldsymbol{\phi}$ was generated iid from Unif(0,0.01). The causal effect θ was set to be 0 or 0.2. Additionally, we generated 1000 null SNPs with the same MAF, of which we used the sample correlation of the z-scores to estimate ρ. The GWAS sample sizes for the exposure and the outcome were set as $N_1 = N_2 = N = 25\,000$. We varied the proportion of sample overlap as one of {0%, 50%, 80%, 100%}. For example, the exposure GWAS summary statistics were calculated using the first 25 000 individuals, and the outcome GWAS summary statistics were calculated using the next 25 001-50 000 individuals for 0% overlap, or using individuals 12 501-37 500 for 50% overlap.

For each simulation set-up, we ran 1000 replicates and compared some popular MR methods, MR-IVW (Burgess et al., 2013), weighted-median (Bowden et al., 2016), weighted-mode (Hartwig et al., 2017), MR-Mix (Qi and Chatterjee, 2019), MR-RAPS (Zhao et al., 2020), MR-cML-I (Xue et al., 2021), with the proposed MR-cML-C.

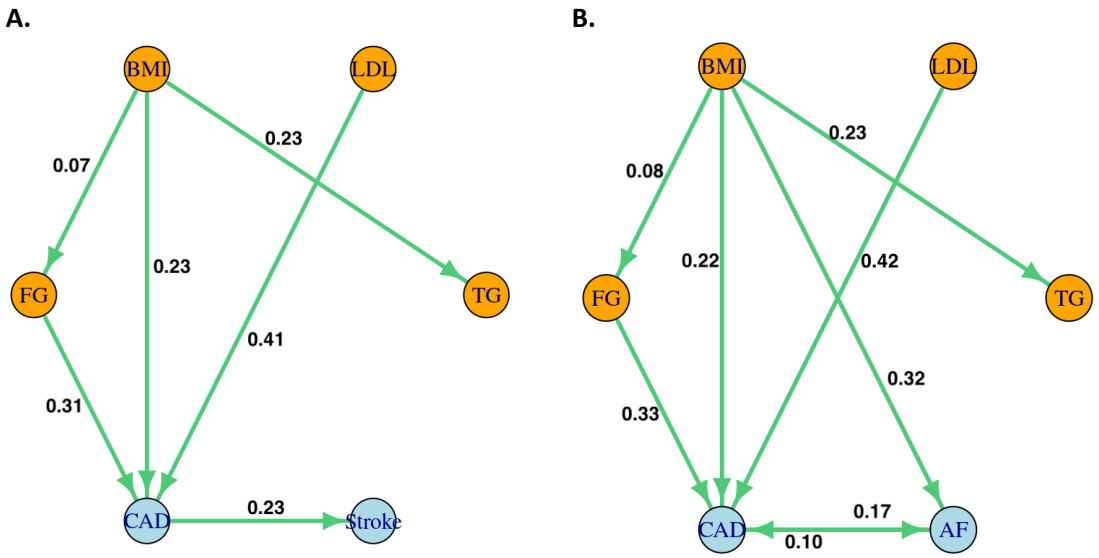


Figure 2.2: Estimated direct causal graphs for 6 traits.

2.2.6 Simulation for direct causal network inference

We conducted simulations based on real GWAS summary statistics to study the performance of Graph-MRcML. We generated GWAS summary statistics based on the direct graph among 6 traits as shown in Figure 2.2. We considered scenarios with and without cycles in the direct graph, corresponding to Figure 2.2A and Figure 2.2B respectively. Briefly, we first generated the true GWAS effect sizes $\mathbf{B}$ based on the underlying direct graph, and then generated the GWAS estimates $\widehat{\mathbf{B}}$ based on the matrix normal distribution introduced in Section 2.2.3.1 to capture the LD structure and sample structure among GWAS data as in the real data analysis. We also considered different GWAS sample sizes by varying the standard error matrix $\mathbf{S}$. Details are given in Appendix A.4.5.

We repeated the simulation 100 times for each set-up, and applied Graph-MRcML with 200 data perturbations on each simulated dataset $\widehat{\mathbf{B}}$ and $\mathbf{S}$, with other inputs remaining the same as in the real data analysis, including the LD matrix $\mathbf{R}$ and the

correlation matrix $\mathbf{P}$ of the GWAS summary data, and the set of IVs used in each MR analysis after the screening process. We also performed a simulation to study the influence of the number of IVs used in an analysis.

2.2.7 Real data analysis

We applied the proposed Graph-MRcML framework to study the causal relationships among 17 traits, including 11 cardiometabolic risk factors and 6 diseases. The 11 risk factors were triglycerides (TG), low-density lipoprotein cholesterol (LDL), high-density lipoprotein cholesterol (HDL), Height, body-mass index (BMI), birth weight (BW), diastolic blood pressure (DBP), systolic blood pressure (SBP), fasting glucose (FG), Smoke (cigarette per day) and Alcohol (alcoholic drinks per week). The 6 diseases were coronary artery disease (CAD), stroke, type 2 diabetes (T2D), asthma (more as a negative control), atrial fibrillation (AF) and Alzheimer's disease (AD). The sample sizes for the 17 GWAS datasets ranged from 10 083 to 1 030 836, with a median of 256 879. We followed the same data pre-process steps described in Section 2.3 of (Xue and Pan, 2022) to prepare the data using `TwoSampleMR` package (Hemani et al., 2018b).

2.2.8 Data and code availability

All GWAS summary data used in the real data example are publicly available as listed in Table A.33. R code to pre-process the data and implement the proposed methods is available on GitHub at https://github.com/ZhaotongL/GraphMRcML.

2.3 Results

2.3.1 Simulation for MR with sample overlap: better type-I error control and higher power of MR-cML-C than other methods

We compared the data perturbation version of MR-cML-I and MR-cML-C, and other commonly used MR methods in this section. Here we discuss main findings while more results are given in the Appendix A.2.

In the case of no invalid IVs, most of the methods were able to control the type-I error reasonably well in the presence of sample overlap; only MR-IVW and MR-RAPS had slightly inflated type-I errors. MR-cML-DP-C, MR-cML-DP-I, IVW, Weighted-Median and MR-RAPS had comparably higher power than Weighted-Mode and MR-Mix (Figure A.8). However, as shown in Figure 2.3, only MR-cML-DP-C gave unbiased causal estimates across all the scenarios, while other methods might remain biased, even more so in some situations, as the proportion of sample overlap increased.

In the case of 30% invalid IVs with correlated pleiotropy, Figure 2.4 shows the empirical type-I error (left) and power (right) for different methods. Only MR-cML-DP-C, Weighted-Mode and MR-Mix could control the type-I error as the proportion of sample overlap increased, and MR-cML-DP-C had the highest power among these three methods. As shown in Appendix A.2, MR-cML-DP-C yielded smaller biases and MSEs than many other methods, especially when the proportion of sample overlap was high. In particular, with the working independence assumption, MR-cML-DP-I was able to control the type-I error when there was no sample overlap, but had slightly inflated type-I errors as the sample overlapping proportion increased. In addition, as shown in Appendix A.2, while MR-cML-C and MR-cML-I yielded biased

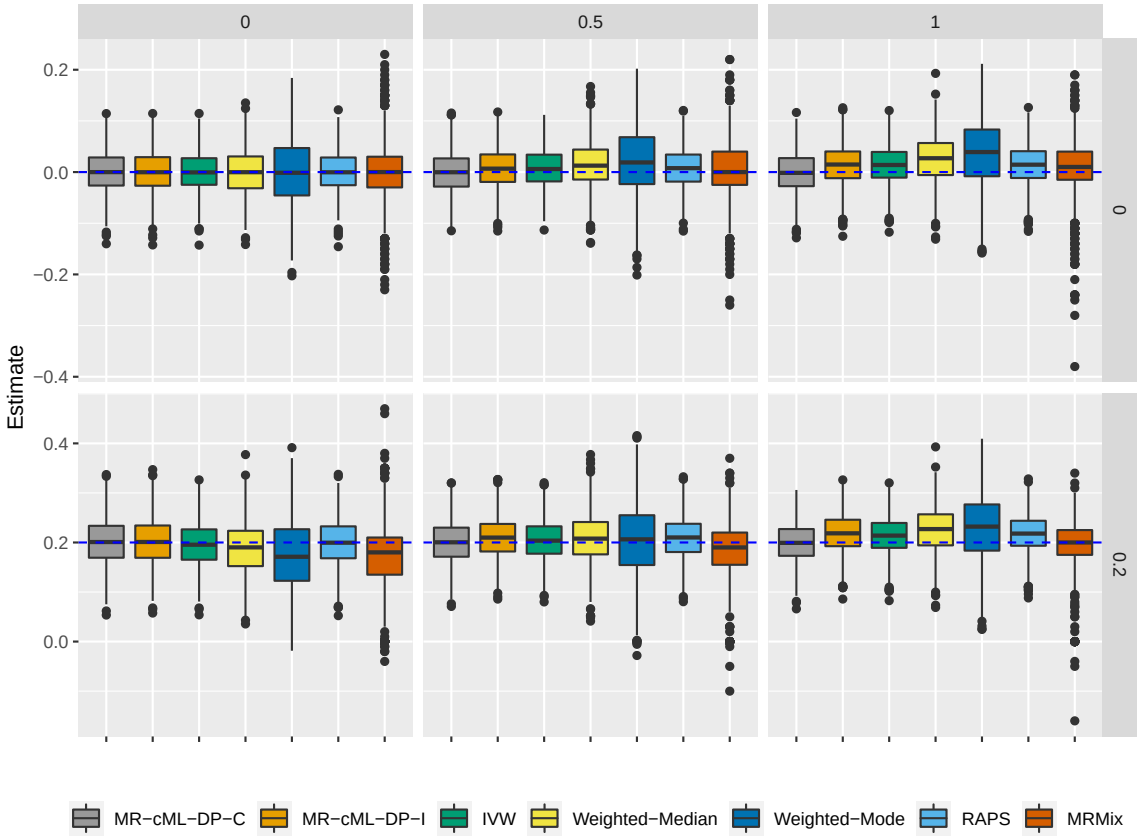


Figure 2.3: Estimates of the causal effect θ with 0% invalid IVs across 1000 replicates. From left to right correspond to 0%, 50% and 100% overlapping samples. Top panel:$\theta = 0$ and bottom panel: $\theta = 0.2$.

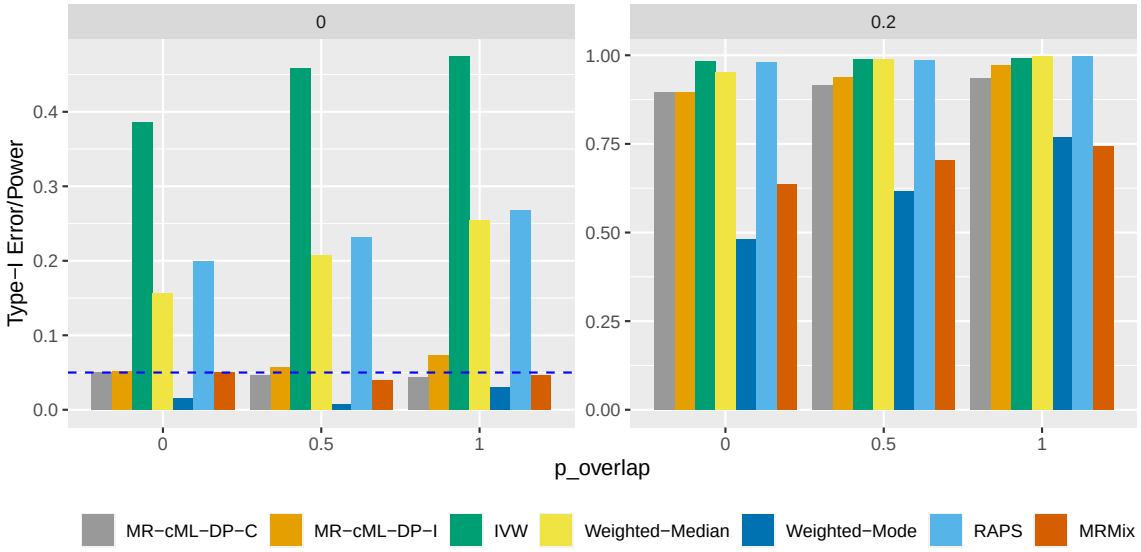


Figure 2.4: Empirical type-I error and power in the presence of 30% invalid IVs with correlated pleiotropy. X-axis represents different proportions of sample overlap (0%, 50% and 100%). Left: $\theta = 0$ (type-I error) and right: $\theta = 0.2$ (power).

estimates because both methods sometimes failed to identify all invalid IVs (perhaps due to the small sample size and/or small effects of some invalid IVs), the bias of MR-cML-C was smaller than that of MR-cML-I in presence of sample overlap. A likely reason was that MR-cML-C performed better in identifying invalid IVs than MR-cML-I (Figure A.15). It is also notable that as the proportion of the overlapping samples increased, even as the total sample size decreased, the estimates of the causal parameter θ became more precise with smaller variances, and the estimate by MR-cML-C was less biased (Figure A.9 and Figure A.12). This suggests an advantage of using the overlapping-sample or only one-sample design over using the two-sample design in MR for causal inference (when the sample structure is correctly accounted for).

2.3.1.1 Other simulation results

We performed simulations to study the consistency of MR-cML-BIC-I when we misspecified the model by ignoring the correlation between the two GWAS summary datasets. First, BIC based on the incorrectly specified model was able to select the correct set of invalid IVs with increasing probabilities as the sample size being increasing. Second, we studied the performance of MR-cML-BIC-I when assuming it correctly selected the set of invalid IVs. In this case, the bias of MR-cML-BIC-I was going to zero as the sample size increased, but the standard error was overestimated using the usual (naive or model-based) variance estimator. On the other hand, the robust sandwich variance estimator was consistent in estimating the standard error but the confidence interval based on it had a low coverage rate mainly due to the finite-sample bias of the causal estimate by MR-cML-BIC-I. We also point out that, although the selection consistency of BIC used in MR-cML-BIC-I and the estimation consistency of MR-cML-BIC-I still hold in the presence of sample overlap, MR-cML-BIC-C still outperformed MR-cML-BIC-I, especially when the sample size was not large enough. Details are given in Appendix A.1.4.

A model averaging (MA) approach was proposed in Xue et al. (2021) (that is often combined with data perturbation) to achieve better inferential performance for finite samples. We also adopted the MA approach in MR-cML-C, called MR-cML-MA-C (and MR-cML-MA-DP-C with data perturbation). We found that MR-cML-MA-C performed better than MR-cML-BIC-C, but it was still unsatisfactory: it might yield inflated type-I errors. However, with data perturbation, there was no additional benefit from model averaging; MR-cML-MA-DP-C performed similarly to MR-cML-DP-C. For this reason, we skip the discussion of MA in the Methods section, and we no longer recommend the use of MA. The detailed results are given in Appendix A.3.1.

Table 2.1: Empirical type-I error and power by Graph-MRcML-d1 for (a) Set-up (a) and (b) Set-up (b). Numbers underlined correspond to power.

	BMI	LDL	FG	TG	CAD	Stroke
BMI		0	0.59	1	0.9	0
LDL	0		0	0.01	1	0
FG	0	0		0	0.89	0
TG	0	0.01	0		0.06	0.01
CAD	0	0	0	0		1
Stroke	0	0.01	0	0	0	

(a) Set-up (a)

	BMI	LDL	FG	TG	CAD	AF
BMI		0	0.8	1	0.87	1
LDL	0.02		0	0	1	0
FG	0	0		0	0.96	0
TG	0	0	0		0.01	0
CAD	0	0	0	0		1
AF	0	0	0	0	1	

(b) Set-up (b)

2.3.2 Simulation for direct causal network inference: recovery of the direct causal network by Graph-MRcML

In this section, we studied the performance of Graph-MRcML in recovering the direct causal network. We summarize the main findings here while more detailed results and discussions are provided in the Supplementary.

We first applied Graph-MRcML on the simulated GWAS summary data. The iterative algorithm in Graph-MRcML-d1 converged successfully in all simulations for both set-ups. As shown in Table 2.1, Graph-MRcML-d1 was able to control the type-I error reasonably well and yielded high power. Graph-MRcML-d0 (i.e., the diagonal elements of the total graph were set to zero) also had similar type-I error and power (Table A.14). Both methods gave only small biases for some entries in the direct graph as shown in Tables A.15 and A.16.

We also applied Graph-MRcML to the simulated GWAS summary data of various sample sizes. Detailed results are given in Appendix A.4.5.2. To summarize, there were slightly inflated type-I errors with a smaller sample size, probably because of MR-cML-BIC-C failing to identify all invalid IVs. But as the sample size increased, type-I error was well controlled. Furthermore, as there was no cycle in Set-up (a), the diagonal elements of the total graph were all zeros. As a result, the diagonal elements of the total graph in both Graph-MRcML-d0 and Graph-MRcML-d1 were consistent, and the resulting direct graph estimates approached the true values as the sample

size increased (Tables A.17-A.19). On the other hand, in the presence of cycles, the true diagonal elements of the total graph in Set-up (b) were not all zeros, hence the direct graph estimate by Graph-MRcML-d0 might be off. In contrast, if it converged successfully, Graph-MRcML-d1 yielded almost unbiased estimates of the direct graph as the sample size increased (Tables A.23-A.25). In general, based on our experience, in the absence of cycles in a direct graph, Graph-MRcML-d0 and Graph-MRcML-d1 performed similarly; otherwise, Graph-MRcML-d1 would have better performance than Graph-MRcML-d0.

As shown in the real data analysis to be discussed next, the number of IVs used in each MR analysis had a wide range. To study the impact of the number of IVs on the performance of the proposed method, we conducted an additional simulation with more IVs for FG, which had the least IVs in the real data analysis. We found that when we had more (valid) IVs for FG, the precision of the direct effect estimates of FG on the other traits increased, while the precision for other direct effect estimates might or might not change much, depending on the underlying relationship among the traits. Details are given in Appendix A.4.5.3.

2.3.3 Real data analysis

We applied the proposed Graph-MRcML framework to study the causal relationships among 17 traits, including 11 cardiometabolic risk factors and 6 diseases. We first estimated the correlation matrix $\mathbf{P}$ using both the null Z-scores approach and bivariate LDSC regression as discussed in Section 2.2.2.3. As shown in Figure A.18, most of the GWAS traits were not highly correlated with each other except for a few such as {SBP, DBP}, {TG, HDL, LDL}, the GWAS datasets of which were collected from the same study respectively. Two approaches generally gave very similar results with only slight differences, and we used the LDSC estimates in the subsequent analysis. For the pairs with small correlations ($|\rho| < 0.1$), we set them to be 0 in the subsequent

analysis because, as shown by our numerical and theoretical studies, the results would be robust to ignoring such small correlations.

The numbers of IVs used in all pairwise bi-directional MR analyses (after the IV screening procedure) ranged from 6 to 374, with a median of 48. We performed $B = 2000$ data perturbations and applied bidirectional MR-cML-BIC-C on each perturbed dataset to obtain the total graph. Due to the temporal order of birth weight and the other traits, the total causal effect of each trait to birth weight was set to zero. Given the possible presence of cycles in the underlying (unknown) direct graph (e.g. between SBP and DBP), we applied Graph-MRcML-d1 to obtain the direct graph. We will discuss the results by Graph-MRcML-d1 based on the Bonferroni-adjusted significance level with 228 effective tests, i.e. $0.05/228 \approx 2.2\text{e-}4$. Results by Graph-MRcML-d0 are discussed in Section Appendix A.5.2, where we might end up with an estimated direct network with a spectral radius greater than one.

Additional investigations on the relationship among lipid traits and glycemic traits (Zhu et al., 2022) are discussed in Appendix A.5.4. Our findings are similar to that in Zhu et al. (2022), suggesting that fasting insulin has plausible direct effects on TG and HDL.

2.3.3.1 Total causal effect network identifies many causal relationships among risk factors and complex diseases

Figure 2.5A shows the inferred total causal graph ($\widehat{\mathbf{G}}_{tot}$) for the 17 traits. In a total causal graph, an edge $A \to B$ represents the total causal effect of A on B; its presence or effect size does not depend on whether or what other variables are included in the graph. In other words, all or a part of a total causal effect may be the sum of all mediating effects through other variables included or not included in the graph. First, it suggested that BMI had a positive causal effect on CAD. It also identified many well-accepted causal relationships from the risk factors to diseases as discussed in

Morrison et al. (2020), such as DBP → CAD, LDL → CAD, FG → T2D and so on. As a negative control, no causal path towards asthma was identified. It is also noted that a negative causal effect between Height and AD was suggested. Many observational studies have found that height is inversely associated with the risk of AD (Beeri et al., 2005; Petot et al., 2007; Russ et al., 2014), and a recent longitudinal study that analyzed data from hundreds of thousands of men also found a link between height and the likelihood of developing dementia (Jørgensen et al., 2020).

Second, some bidirectional relationships between a risk factor and a disease were identified: BMI ↔ T2D, FG ↔ T2D. BMI and FG are both well-accepted causal risk factors for T2D. For direction FG ← T2D, it is possible that the pancreas makes more insulin to make up for insulin resistance in T2D, and blood sugar levels build up overtime. For BMI ← T2D, T2D may cause weight loss since the cells cannot get the energy they need from glucose, and the body breaks down fat to use for energy instead, but more studies are needed.

Third, there were many interesting links between the risk factors. For example, our method inferred a positive causal link BMI → Smoke (cigarette per day). Previous observational studies have reported a positive correlation between BMI and smoking intensity (Dare et al., 2015), and a common biological basis for nicotine addiction and obesity was also suggested with genetic evidence (Thorgeirsson et al., 2013). Recently, an MR study suggested that obese individuals are more likely to smoke and with a higher smoking intensity in both the discovery and replication samples (Carreras-Torres et al., 2018).

Lastly, some links among the diseases were identified, such as CAD ↔ AF, AF → Stroke and CAD → Stroke. Common heart disorders are risk factors for stroke (Arboix, 2015), for example, CAD increases the risk for stroke, because plaque builds up in the arteries and blocks the flow of oxygen-rich blood to the brain. Also AF can cause blood clots that may break loose and travel to another part of the body,

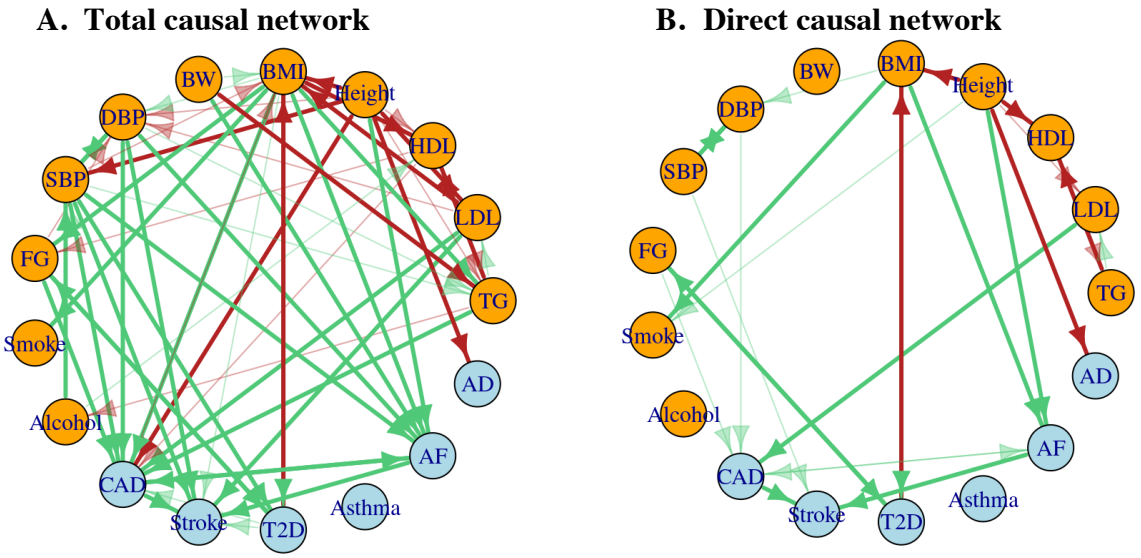


Figure 2.5: The estimated total (A) and direct (B) causal graphs for the 11 risk factors and 6 diseases. The edges in green represent positive effects and those in red are negative ones. The nodes in blue are diseases and those in orange are risk factors. The dark-solid edges are identified at the Bonferroni-adjusted significance level, while the light-colored ones are marginally significant at a less stringent level of 6.5e-3.

cutting off blood supply to the brain (Alshehri, 2019). Studies have shown that there is a vicious cycle between CAD and AF (Liang and Wang, 2021).

2.3.3.2 Direct causal effect network suggests direct causal pathways

Figure 2.5B shows the inferred direct causal graph ($\widehat{\mathbf{G}}_{dir}$) for the 17 traits. It is notable that this $\widehat{\mathbf{G}}_{dir}$ and all the direct graph estimates from the 2000 perturbed datasets had a spectral radius smaller than one, while the iterative algorithm in Graph-MRcML-d1 converged successfully in all cases. In a direct causal graph, an edge $A \to B$ represents the direct causal effect of A to B after conditioning on all the other variables included in the graph; in other words, it is the remaining causal effect of A to B after accounting for (i.e. removing) all possible mediating effects of A to B through any of the other variables included in the graph. Accordingly, the direct

effect size of A to B also depends on what other variables are included in the graph.

First, we can see that several independent risk factors of diseases were identified after accounting for other traits in the graph. For example, LDL was an independent risk factor for CAD. In the Cooper Clinic Longitudinal Study Abdullah et al. (2018), LDL was found to be associated with the CAD mortality in a multivariable model after adjusting for atherosclerotic CVD risk factors such as HDL, current tobacco use, hypertension, BMI and glucose. Another example is BMI $\rightarrow$ AF. Many observational cohort studies have highlighted BMI as an independent risk factor for AF after adjusting for traditional risk factors. In the Women's Health study, BMI was found to be associated with elevations in AF risk after accounting for a wide range of covariates including diabetes, hypertension, history of hypercholesterolemia, alcohol consumption and smoking (Tedrow et al., 2010). And in the Danish Diet, Cancer, and Health Study Frost et al. (2005), BMI was found to be significantly associated with AF after adjusting for height, smoking, alcohol consumption, hypertension, diabetes, heart diseases, etc. Similarly, height is also a well-known independent risk factor for AF (as shown by Height $\rightarrow$ AF) after adjusting for many traditional risk factors as evidenced by many studies (Sohail et al., 2021; Persson et al., 2018; Andersen et al., 2018). Overall, some of our findings here have lent strong support for some existing causal hypotheses drawn from previous observational cohort studies.

Perhaps as expected, the direct causal graph was less dense than the total causal graph in Figure 2.5A. First, multiple edges among the risk factors in the total causal graph were removed from the direct causal graph. For example, the edge from BMI to FG disappeared in the direct graph, which was probably because T2D served as an important mediator: BMI $\rightarrow$ T2D $\rightarrow$ FG. At the same time, the reverse path FG $\rightarrow$ T2D $\rightarrow$ BMI was also significant, suggesting the mechanisms underlying obesity, T2D and fasting glucose might be complicated, and more studies are needed (Karpe et al., 2011).

Moreover, some edges between the risk factors and diseases were also removed. For example, LDL → Stroke is a known causal pair, however, as shown in Figure 2.5B, CAD might act as a mediator: LDL → CAD → Stroke. Another important question of interest is the role of BMI on CAD. While the total graph suggested BMI as a causal risk factor for CAD, it has been controversial about whether BMI is an independent risk factor for CAD (Powell-Wiley et al., 2021). Obesity is associated with many pathophysiological mechanisms involved in the development of CAD, such as preposition to insulin resistance and type 2 diabetes mellitus, lipid abnormalities and hypertension. In our analysis, the direct effect of BMI on CAD was not significant (p-value ≈ 0.16) after accounting for other factors; the direct causal effect estimate of BMI on CAD was attenuated from an estimated total causal effect of 0.31 (in the logOR scale) $\widehat{\mathbf{G}}_{tot}$ to 0.09 in $\widehat{\mathbf{G}}_{dir}$. Furthermore, at a less stringent significance level (p-value<6.5e-3), there were indirect causal pathways from BMI to CAD via blood pressures, T2D and fasting glucose. Overall, our results suggested that BMI may be considered as a 'minor' independent risk factor for CAD after accounting for other factors or comorbidities (Ades and Savage, 2017).

For comparison, we also applied MVMR-IVW and MVMR-Robust (Grant and Burgess, 2021) to investigate the causal effects of the 15 traits on CAD; we did not include Alzheimer's disease (mainly because we did not expect to detect its causal effect on CAD). We used the `mv_extract_exposures` function in `TwoSampleMR` package to obtain the candidate IVs that were (nearly) independent across all the 15 traits (as required by current MVMR methods), leading to fewer IVs: 6 traits only had 6 or fewer IVs, and T2D only had one IV (with p-value<5e-8); this is a downside of using any current MVMR methods as discussed earlier. As shown in Table A.34, 13 out of the 15 traits as exposures had a conditional F-statistic smaller than 10 (4.74 for BMI), suggesting that an MVMR analysis may suffer from weak instrument biases (Sanderson et al., 2021). At the end, MVMR gave a similar conclusion: the esti-

mated direct effect of BMI on CAD by MVMR-IVW was 0.07 with p-value 0.37, and it was 0.12 with p-value 0.06 by MVMR-robust. A recent study found a significant direct effect of BMI on CAD using MVMR-Robust, but with a smaller set of 6 risk factors, including Height, BMI, LDL, TG, SBP and HbA1c (Wang et al., 2022b). We further applied MVMR with the 5 exposures (excluding HbA1c), and the conditional F-statistics for BMI and SBP were still smaller than 10 (Table A.35). Nevertheless, with this smaller set of risk factors, the estimated direct effect of BMI on CAD by MVMR-Robust was 0.20 with a significant p-value 6e-4, while by MVMR-IVW it was 0.13 with p-value 0.08. Lastly, our approach also suggested a significant direct effect of BMI on CAD (with an effect estimate 0.23 and p-value 1e-4) in this smaller network with the five traits and CAD.

2.4 Discussion

It is always of interest to disentangle the causal relations among multiple traits, in which, for example, one can distinguish direct versus indirect/mediating causal effects. In this paper, we have proposed a general framework called Graph-MRcML to infer both a total causal (effect) network and a direct causal (effect) network. This framework has several merits. First, it allows for bidirectional edges or cycles in a directed graph, which is more likely to reflect true biological processes (Zhu et al., 2007). Second, using bidirectional MR to infer a total causal network alleviates the issue of unmeasured confounding and reverse causation with observational data. It does not require users to specify causal directions in advance. Third, many current methods of estimating causal networks require that the data for all the traits come from the same sample without hidden confouding (Li et al., 2021a; Yuan et al., 2019); in contrast, our proposed framework can use GWAS data of traits from different (and possibly overlapping) samples with hidden confounding. Fourth, besides estimating

both a total and a direct causal networks, the proposed data perturbation scheme allows for robust inference; that is, in addition to reconstructing a causal network, it allows testing the presence of an edge and constructing a confidence interval for any causal effect. Moreover, our proposed data perturbation scheme is both novel and effective by using a matrix normal distribution to effectively account for possible correlations among the SNPs (due to linkage disequilibrium) and among the traits simultaneously.

While our proposed framework is flexible in that it can use any bidirectional MR method to construct a total causal graph, we have focused on using MR-cML for its superior and robust performance. In particular, MR-cML is robust to both correlated and uncorrelated pleiotropic effects, while it possesses some nice statistical properties (e.g. estimation consistency and asymptotic normality) with impressive numerical performance. Furthermore, as shown in Xue and Pan (2022) both numerically and theoretically, with a simple IV screening procedure, MR-cML achieves good performance in inferring bi-directional causal directions under different scenarios with the exposure and the outcome being continuous and/or binary, as well as with some SNPs associated with a confounder of the two traits. However, it was originally proposed as a two-sample MR method. Nowadays, many large-scale GWAS were/are conducted by various consortia formed by many smaller studies with overlapping samples to maximize the total sample size and statistical power. Therefore, some SNP-exposure associations and SNP-outcome associations might be estimated from overlapping samples. To our best knowledge, except for a few newly proposed methods (Hu et al., 2022; Mounier and Kutalik, 2021; Cheng et al., 2022b), most of the widely-used MR methods are based on two independent samples (Qi and Chatterjee, 2019; Xue et al., 2021; Gleason et al., 2021). In this work, we have extended MR-cML to a more general set-up allowing sample overlap - it can be applied to two-sample, overlapping-two-sample and even one-sample set-ups. We have shown that

all desirable statistical properties in the original version (Xue et al., 2021) carry over in this extended version of MR-cML. It is notable that we have also shown in the Supplementary that both the BIC selection consistency and the estimation consistency based on the original MR-cML-I in Xue et al. (2021) (under the possibly incorrect assumption of no sample overlap) still hold in the presence of sample overlap; however, with realistic finite sample sizes, the extended MR-cML-C performed much better than MR-cML-I in the presence of sample overlap as shown in our simulation studies. Nevertheless, our numerical and theoretical results on MR-cML-I in the presence of sample overlap does suggest its (asymptotic) robustness, explaining why in practice it might be fine to ignore the issue of sample overlap if the proportion of overlapping is small.

There are a few limitations in this work. First, our proposed method is based on the classic statistical theory for a large sample size (N) and a fixed/small number of both IVs (m) and traits. This is suitable for a typical MR analysis with the sample sizes of GWAS data in tens to hundreds of thousands, while the number of IVs is often from tens to hundreds and that of traits in in low tens. However, if the number of (valid) IVs is small and/or the number of traits is large relative to the sample sizes, the finite-sample performance may go down with less precise estimates and loss of power. As shown in the simulations, in the presence of invalid IVs, the proposed method may yield biased estimates due to finite sample size; however, as shown by the theory, we expect it to give consistent selection of invalid IVs and consistent estimates as the sample size increases. Moreover, with a large number of traits, as in multiple regression, an issue similar to multicollinearity may appear or become more severe. In the future, it would be useful to incorporate variable selection to select and use only a subset of necessary traits to be included in the graph, though it may be then necessary to address the issue of post-selection inference, e.g. via data perturbation. Second, as in typical MR applications, we used the same GWAS sample

to select significant SNPs as IVs and to estimate their association effects with the exposure. This will lead to the well-known "winner's curse" or selection bias (Wang and Han, 2021). As in Hu et al. (2022), we may account for the selection process by suitably adjusting the likelihood, which will be a future topic. Third, although the graph deconvolution algorithm is straightforward with a closed-form solution, the key assumption is that a direct causal graph has a spectral radius smaller than one. In practice, this assumption may be violated for an estimated direct causal graph, e.g. due to errors in estimating the corresponding total causal graph. However, when such a violation is not severe, the estimated direct causal graph might still be useful. Furthermore, because our proposed estimators for both the total and direct graphs are consistent (when the assumptions for MR-cML and graph deconvolution hold), it is expected that increasing sample sizes of GWAS will alleviate the potential problem. Fourth, in the real data analysis, due to the fact that one's birth weight cannot be affected by any traits developed/measured in a later time, the (true) total causal effects of any latter traits to birth weight should be zeros, and we set them as zeros in the estimated total causal graph. While such a practice is not necessary when applying our proposed method, we expect that, by taking advantage of this prior knowledge, doing so would perform better for finite samples. Finally, as different MR methods rely on their own assumptions as well as the quality of genetic variants as IVs, more applications to real data, including applying alternative MR methods would be warranted as a means of causal triangulation.

Chapter 3

Robust multivariable Mendelian randomization based on constrained maximum likelihood

3.1 Introduction

Mendelian randomization (MR), including its default version, univariable MR (UVMR), is an instrumental variable (IV) method that utilizes genetic variants as IVs to infer the causal relationship between an exposure and an outcome (Davey Smith and Ebrahim, 2003; Zhu, 2021; Boehm and Zhou, 2022). With numerous publicly available large-scale genome-wide association study (GWAS) summary data, MR has recently become popular and powerful to infer causal relationships even in the presence of unmeasured confounding and reverse causation (Sleiman and Grant, 2010). In UVMR a valid IV is required to satisfy the following three assumptions:

UV-A1: the IV is (marginally) associated with the exposure;

UV-A2: the IV is independent of the unmeasured confounder;

UV-A3: the IV is independent of the outcome conditional on the exposure and confounder.

Despite promising and wide applications of UVMR to observational data for causal

inference, in reality, these assumptions may not always hold. In particular, the widespread pleiotropy is a major concern, violating UV-A2 or UV-A3; that is, a genetic variant is associated with the outcome other than through the exposure of interest (Hemani et al., 2018a). Screening procedures can be applied to avoid using pleiotropic variants, but there may be only few or even no variants solely associated with the exposure, leading to loss of power, in addition to likely biases. A number of UVMR methods robust to pleiotropy have been proposed, but under different untestable assumptions (Hemani et al., 2018a; Lin et al., 2021; Xue et al., 2021; Morrison et al., 2020; Qi and Chatterjee, 2019; Cheng et al., 2022b; Hu et al., 2022). Alternatively, one can alleviate the problem by including other associated risk factors as multiple exposures in the model, motivating the use of multivariable MR (MVMR) (Burgess and Thompson, 2015). MVMR includes multiple exposures in the model and allows the genetic variants to be associated with one or more of them without violating the IV assumption. This is useful especially when a set of related risk factors for the outcome of interest share many commonly associated genetic variants, e.g., various lipids (Waterworth et al., 2010). Another scenario where MVMR would be useful is in mediation analysis (Carter et al., 2021). In MVMR, the direct effect of one exposure on the outcome not mediated through the rest of the exposures is estimated, while in UVMR, only the total effect of the exposure on the outcome, including that mediated through other exposures, is estimated. Therefore, MVMR would be useful when the aim is to study the causal mechanism of a set of risk factors on the outcome after accounting for potential causal pathways among the risk factors.

In spite of these exciting advantages of MVMR over UVMR, some assumptions are still required to have valid IVs for MVMR. Paralleling with that in UVMR, in MVMR a valid IV must satisfy that

MV-A1: the IV is associated with at least one exposure conditional on the other exposures included in the model;

MV-A2: the IV is independent of any confounder of each exposure-outcome pair;

MV-A3: the IV is independent of the outcome conditional on all exposures included in the model and the confounders.

Since a marginal association does not imply a conditional association, while the reverse is also true but less likely, in that sense Assumption MV-A1 is stronger than UV-A1 while MV-A3 is weaker than UV-A3. We give an example on each of the two cases in the simulation results for mediation analysis. We note that the above assumptions are for one IV, instead of a set of IVs as considered by others (Sanderson, 2021). The main reason is that, differing from the previous MVMR approaches, we would like to consider robust MVMR in the presence of some invalid IVs violating one or more of the above three assumptions. While there are many robust UVMR methods in the literature, however, only few MVMR methods exist, which may not be sufficiently robust and efficient (as to be shown) (Grant and Burgess, 2021; Sanderson et al., 2021; Rees et al., 2017). In this work, we propose a robust and efficient MVMR method by extending the univariable MR-cML (Xue et al., 2021) (referred to UVMR-cML) to the multivariable setting, called MVMR-cML. Based on constrained maximum likelihood (cML), UVMR-cML is robust to the violation of all three IV assumptions (in the univariable case): it allows the presence of invalid IVs violating any or all of the three IV assumptions as long as the plurality condition and several other mild conditions hold. Under some mild conditions, it can consistently identify invalid IV(s) with either or both of correlated and uncorrelated pleiotropy, yielding a consistent and asymptotic normal estimator of the causal effect. It has been shown to have robust performance under various simulation setups and real data analyses (Xue et al., 2021). Under a similar constrained maximum likelihood framework, MVMR-cML is expected to enjoy the same good properties as its univariable version. In particular, as its UV counterpart, it is robust to the presence of invalid IVs violating some or all of the three IV assumptions, including those with correlated or uncorre-

lated pleiotropy as to be confirmed in simulations. The flexible likelihood framework also allows to account for possible correlations among any sets of GWAS summary statistics (e.g. due to overlapping samples), and thus the method can be widely applied in different scenarios, including one-sample, two-sample, or even mixed-sample designs. We develop an efficient R package with Rcpp to integrate different versions of MR-cML for various uses.

The rest of this chapter is organized as follows. First, we extend the plurality condition for model identification from the univariable case (Guo et al., 2018) to the current multivariable context. Then we introduce MVMR-cML and its two variants based on model selection and data perturbation respectively. We show the superior performance of the proposed method over other existing MVMR methods through extensive simulations. Lastly, we consider an application to study the direct effects of 8 cardio-metabolic risk factors on coronary artery disease (CAD [MIM: 608320]).

3.2 Methods

3.2.1 Causal model and its interpretation

Suppose we have m independent SNPs, L exposures X_l's with $l = 1, \cdots, L$, an unmeasured confounder (ensemble) U and an outcome Y. We consider the following true causal model (Figure 3.1A), which can be viewed as a multivariable version of the true causal model presented in Xue et al. (2021):

$$b_{Yi} = \theta_1 b_{X_1 i} + \cdots + \theta_L b_{X_L i} + \alpha_i + \beta_{YU}\phi_i := \theta_1 b_{X_1 i} + \cdots + \theta_L b_{X_L i} + r_i, \qquad (3.1)$$

where $b_{X_l i}$ and b_{Yi} are the (true) marginal associations between SNP/IV G_i and the l-th exposure and the outcome respectively, θ_l is the direct causal effect of X_l on Y, $r_i = \alpha_i + \beta_{YU}\phi_i$ represents the total pleiotropic effect of G_i on Y (not through

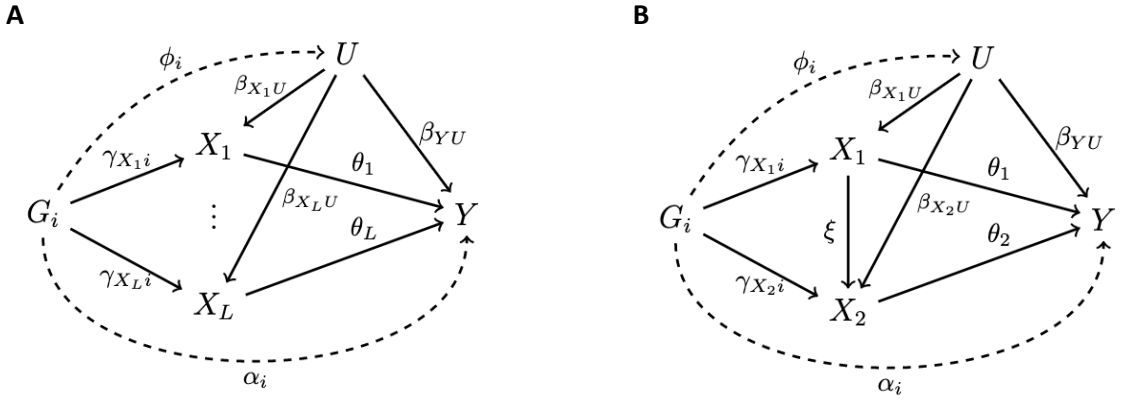


Figure 3.1: (A) A general and (B) a specific causal graphs showing the relationships among one IV (G_i), multiple exposures ($X_1, ..., X_L$), an unmeasured confounder (U) and the outcome (Y).

X_l's). For example, in the absence of any causal relationship among the exposures in Figure 3.1A, we have $b_{X_li} = \gamma_{X_li} + \beta_{X_lU}\phi_i$, for $l = 1, \ldots, L$; in Figure 3.1B with a causal relation from X_1 to X_2, we have $b_{X_1i} = \gamma_{X_1i} + \beta_{X_1U}\phi_i$ and $b_{X_2i} = \gamma_{X_2i} + \beta_{X_2U}\phi_i + \xi b_{X_1i}$.

The three valid IV assumptions for a valid IV G_i require respectively $\gamma_{X_li} \neq 0$ (for some $1 \leq l \leq L$), $\phi_i = 0$ and $\alpha_i = 0$, the last two of which imply no pleiotropic effects with $r_i = 0$. On the other hand, $\alpha_i \neq 0$ and $\phi_i \neq 0$ lead to so-called *uncorrelated* and *correlated* pleiotropy respectively. It is a major goal here to consider robust MVMR analysis in the possible presence of some invalid IVs with $r_i \neq 0$.

In the true causal model depicted in Figure 3.1A, for MVMR we do not need to specify causal relationships among the L exposures, the presence of which is likely and, importantly, implies that the direct causal effect θ_l is in general different from the total causal effect of X_l on Y. For illustration, we consider a simple example with $L = 2$ exposures, one of which mediates the effect of the other on the outcome (Figure 3.1B); more examples can be found elsewhere (Sanderson, 2021; Sanderson

et al., 2019). The corresponding linear regression model for MVMR is

$$Y = X_1\theta_1 + X_2\theta_2 + U\beta_{YU} + G_i\alpha_i + \epsilon,$$

where ϵ is a random error. Accordingly, we can interpret the direct (causal) effect θ_1 of X_1 on Y as that of changing X_1 while holding X_2 fixed. In contrast, by replacing $X_2 = \xi X_1 + G_i\gamma_{X_2 i} + U\beta_{X_2 U} + e_2$ (with e_2 being a random error), we obtain the corresponding linear regression model for UVMR for X_1 as

$$Y = X_1(\theta_1 + \xi\theta_2) + U(\beta_{YU} + \beta_{X_2 U}\theta_2) + G_i(\alpha_i + \gamma_{X_2 i}\theta_2) + (\epsilon + e_2\theta_2),$$

giving the total effect of X_1 on Y as $\theta_1 + \xi\theta_2$, including both the direct effect θ_1 and the indirect effect $\xi\theta_2$ mediated through X_2. In addition, this example also illustrates an advantage of MVMR over UVMR: if $\alpha_i = 0$ and $\phi_i = 0$, G_i is likely a valid IV with no pleiotropic effect (i.e. $r_i = 0$) in MVMR; however, in UVMR for X_1, G_i is an invalid IV with pleiotropic effect $r_i = \gamma_{X_2 i}\theta_2 \neq 0$.

3.2.2 Model identification

Suppose the ground truth to Eq (3.1) is given by $b_{Yi} = \theta_1^* b_{X_1 i} + \cdots + \theta_L^* b_{X_L i} + r_i^*$, and the set of valid IVs is $\mathcal{V}^* = \{i : r_i^* = 0\}$ with $m_0 = |\mathcal{V}^*|$. Here we use the asterisk to denote the true value of a parameter, and only consider relevant IVs with $\mathbf{b}_{Xi} = (b_{X_1 i}, \ldots, b_{X_L i})^T \neq \mathbf{0}$. We say that the true parameters $\mathbf{r}^* = (r_1^*, \ldots, r_m^*)^T$ and $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_L^*)^T$ are identifiable if, given the (true) marginal associations b_{Yi} and $\mathbf{b}_{Xi} \neq \mathbf{0}$ for $i = 1, \cdots, m$, there is a unique solution $(\mathbf{r}^*, \boldsymbol{\theta}^*)$ (as the ground truth) to

$$b_{Yi} = r_i + \theta_1 b_{X_1 i} + \cdots + \theta_L b_{X_L i}, \quad i = 1, \ldots, m, \tag{3.2}$$

under the constraint that any solution is obtained from the *largest* set(s) of IVs with their corresponding $r_i = 0$ (as a part of the solution). Note that the constraint is needed because otherwise we would have many solutions with $m + L$ unknown parameters (r_i's and θ_j's) in m linear equations.

There are two aspects to consider for model identification. First, unique to MVMR (i.e. different from UVMR), even in the ideal case of using only m_0 valid IVs (with $r_i = 0$), in order to have a unique solution to the system of linear equations in Eq (3.2), we require that the marginal association matrix $\mathbf{B}_{\mathcal{V}^*} = (\mathbf{b}_{X1}, \mathbf{b}_{X2}, ..., \mathbf{b}_{Xm_0})^T$ to be of full column rank, which implies at least one valid IV for each exposure and thus $m_0 \geq L$. This condition ensures the marginal associations between the IVs and exposures are not multicollinear, as pointed out by others; a conditional F-test has been proposed to test for possible violation of this condition (Sanderson et al., 2021). The other aspect, similar to that in UVMR, is a plurality condition in the presence of invalid IVs with $r_i \neq 0$: the (true) valid IVs form the largest group to give the same causal parameter estimate. The two aspects are combined into the following assumption.

Assumption 3.1

Suppose the matrix of $\mathbf{b}_{Xi}$'s for $i \in \mathcal{V}^*$, $\mathbf{B}_{\mathcal{V}^*} = [\mathbf{b}_{Xi}]_{i\in\mathcal{V}^*} \in \mathbb{R}^{m_0\times L}$, has full column rank L. Moreover, the following multivariable plurality condition holds:

$$|\mathcal{V}^*| > \max_{\{\mathbf{c}\neq\mathbf{0},\mathbf{c}\in\mathbb{R}^L\}} \left|\{i : r_i^* = \mathbf{b}_{Xi}^T\mathbf{c}\}\right|. \tag{3.3}$$

Theorem 3.1

Given b_{Yi} and $\mathbf{b}_{Xi}$ for $i = 1, \ldots, m$, the true parameters $\boldsymbol{\theta}^*$ and $\mathbf{r}^*$ in Eq (3.2) are identifiable if and only if Assumption 3.1 holds.

The proof of Theorem 3.1 is given in Appendix B.3.1. We first give some intuitive explanation on why the plurality condition is needed. If it is violated, it means that

some invalid IVs form the largest group with a solution $\mathbf{c}+\boldsymbol{\theta}^* \neq \boldsymbol{\theta}^*$ to Eq (3.2). Next we note that the plurality condition in Eq (3.3) is a generalization of that for the univariable case with $L = 1$. Theorem 1 in Guo et al. (2018) states that, given b_{Yi} and $b_{X_1 i}$, the model parameters θ_1 and $r_1, \ldots, r_m$ are identifiable if and only if the following plurality rule condition holds:

$$|\mathcal{V}^*| > \max_{c \neq 0} |\{i : r_i^*/b_{X_1 i} = c\}|, \tag{3.4}$$

where $\mathcal{V}^*$ is the set of valid IVs with $r_i = 0$ and $b_{X_1 i} \neq 0$. It is clear that Eq (3.3) reduces to Eq (3.4) with $L = 1$.

3.2.3 New method: multivariable MRcML (MVMR-cML)

We extend the UVMR-cML proposed in Xue et al. (2021) to the multivariable case. The goal of MVMR-cML is to estimate the direct causal effect of each exposure on the outcome (while relaxing the no-pleiotropy condition for valid IVs). Denote $\{\hat{\beta}_{X_1 i}, \ldots, \hat{\beta}_{X_L i}, \hat{\beta}_{Yi}, \hat{\sigma}^2_{X_1 i}, \ldots, \hat{\sigma}^2_{X_L i}, \hat{\sigma}^2_{Yi}\}_{i=1}^m$ as the GWAS summary statistics of m (nearly) independent SNPs/IVs, each of which is selected based on its significant marginal association with at least one of the exposures (e.g. at the usual genome-wide significance level of p-value $<$ 5e-8). We have

$$\hat{\beta}_{X_l i} = b_{X_l i} + \epsilon_{X_l i}, \quad l = 1, \ldots, L,$$
$$\hat{\beta}_{Yi} = b_{Yi} + \epsilon_{Yi},$$

with $\text{var}(\epsilon_{X_l i}) = \sigma^2_{X_l i}$ and $\text{var}(\epsilon_{Yi}) = \sigma^2_{Yi}$. We assume $\sigma^2_{X_l i}$ and σ^2_{Yi} are known or well estimated as $\hat{\sigma}^2_{X_l i}$ and $\hat{\sigma}^2_{Yi}$ respectively. Then we have the model for SNP i:

$$\hat{\boldsymbol{\beta}}_i = (\hat{\beta}_{X_1 i}, \ldots, \hat{\beta}_{X_L i}, \hat{\beta}_{Yi})^T \sim \mathcal{N}\left(\mathbf{b}_i = (b_{X_1 i}, \ldots, b_{X_L i}, \sum_{l=1}^{L} \theta_l b_{X_l i} + r_i)^T, \boldsymbol{\Sigma}_i\right), \tag{3.5}$$

where $\boldsymbol{\Sigma}_i = \begin{pmatrix} \sigma^2_{X_1 i} & \rho_{12}\sigma_{X_1 i}\sigma_{X_2 i} & \cdots & \rho_{1L}\sigma_{X_1 i}\sigma_{X_L i} & \rho_{1Y}\sigma_{X_1 i}\sigma_{Yi} \\ & \sigma^2_{X_2 i} & \cdots & \rho_{2L}\sigma_{X_2 i}\sigma_{X_L i} & \rho_{2Y}\sigma_{X_2 i}\sigma_{Yi} \\ \vdots & & \ddots & & \vdots \\ & \ddots & & \sigma^2_{X_L i} & \rho_{LY}\sigma_{X_L i}\sigma_{Yi} \\ & & \cdots & & \sigma^2_{Yi} \end{pmatrix}$, $\rho_{ll'}$ $(l, l' = 1, \ldots, L$ and $l \neq l')$ is the correlation between the two GWAS summary estimates for exposures X_l and $X_{l'}$, and ρ_{lY} $(l = 1, \ldots, L)$ is the correlation between the two GWAS summary datasets for exposure X_l and outcome Y. When the $L+1$ GWAS summary datasets are calculated from $L+1$ sets of non-overlapping samples respectively, $\rho_{ll'} = \rho_{lY} = 0$ and $\boldsymbol{\Sigma}_i$ is diagonal. Otherwise, we can use bivariate LDSC to estimate all the ρ's (Bulik-Sullivan et al., 2015), or using the correlation between the two sets of GWAS null Z-scores (Kim et al., 2015).

Since the m IVs are independent, the log-likelihood of the observed GWAS data (up to some constants) is

$$l(\boldsymbol{\theta}, \{\mathbf{b}_{Xi}\}, \{r_i\}; \{\hat{\boldsymbol{\beta}}_i, \boldsymbol{\Sigma}_i\}) = \sum_{i=1}^{m} l_i(\boldsymbol{\theta}, \mathbf{b}_{Xi}, r_i; \hat{\boldsymbol{\beta}}_i, \boldsymbol{\Sigma}_i) = -\frac{1}{2}\sum_{i=1}^{m} (\hat{\boldsymbol{\beta}}_i - \mathbf{b}_i)^T \boldsymbol{\Sigma}_i^{-1} (\hat{\boldsymbol{\beta}}_i - \mathbf{b}_i), \tag{3.6}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_L)^T$, and we use $\{\mathbf{b}_{Xi}\} = \{(b_{X_1 i}, \ldots, b_{X_L i})^T, i = 1, \ldots, m\}$ to represent a set of the parameters, and similarly for $\{r_i\}$ and $\{\hat{\boldsymbol{\beta}}_i, \boldsymbol{\Sigma}_i\}$.

Under the constraint that the number of invalid IVs is K, we estimate the unknown

parameters by solving the following constrained maximum likelihood:

$$(\hat{\boldsymbol{\theta}}, \{\widehat{\mathbf{b}}_{Xi}\}, \{\hat{r}_i\}) = \underset{\boldsymbol{\theta}, \{\mathbf{b}_{Xi}\}, \{r_i\}}{\arg\max}\ l(\boldsymbol{\theta}, \{\mathbf{b}_{Xi}\}, \{r_i\}; \{\hat{\boldsymbol{\beta}}_i, \boldsymbol{\Sigma}_i\}) \quad \text{subject to} \sum_{i=1}^{m} I(r_i \neq 0) = K.$$

For a given number of invalid IVs, K, a coordinate descent-like algorithm is implemented as follows: at the $(t+1)$th iteration,

Step 1: calculate $\hat{r}_i^{(t+1)}$ by solving $\frac{\partial l_i}{\partial r_i}\Big|_{\boldsymbol{\theta}^{(t+1)}, \mathbf{b}_{Xi}^{(t)}} = 0$;
order $d_i^{(t+1)} = l_i(\boldsymbol{\theta}^{(t)}, \mathbf{b}_{Xi}^{(t)}, \hat{r}_i^{(t+1)}; \hat{\boldsymbol{\beta}}_i, \boldsymbol{\Sigma}_i) - l_i(\boldsymbol{\theta}^{(t)}, \mathbf{b}_{Xi}^{(t)}, 0; \hat{\boldsymbol{\beta}}_i, \boldsymbol{\Sigma}_i)$ decreasingly, then for $i = 1, \ldots, K$, update $r_{(i)}^{(t+1)} = \hat{r}_{(i)}^{(t+1)}$; for $i = K+1, \ldots, m$, update $r_{(i)}^{(t+1)} = 0$;

Step 2: update $\mathbf{b}_{Xi}^{(t+1)}$ by solving $\frac{\partial l_i}{\partial \mathbf{b}_{Xi}}\Big|_{\boldsymbol{\theta}^{(t)}, r_i^{(t+1)}} = 0$ for $i = 1, \ldots, m$;

Step 3: update $\boldsymbol{\theta}^{(t+1)}$ by solving $\frac{\partial l}{\partial \boldsymbol{\theta}}\Big|_{\{\mathbf{b}_{Xi}^{(t+1)}, r_i^{(t+1)}\}} = 0$.

We repeat the above three steps until convergence, obtaining the final estimates $\widehat{\boldsymbol{\theta}}(K)$ and $\{\widehat{\mathbf{b}}_{Xi}(K), \hat{r}_i(K)\}_{i=1}^m$. As in UVMR-cML (Xue et al., 2021), it is notable that at the convergence the (estimated) invalid IVs (with $\hat{r}_i \neq 0$) do not contribute to estimating $\boldsymbol{\theta}$, and the resulting cMLE of $\boldsymbol{\theta}$ is the same as the maximum (profile) likelihood estimator being applied to all (selected) valid IVs.

We select the number of invalid IVs, K, from a candidate set $\mathcal{K}$ based on the following Bayesian information criterion (BIC):

$$\text{BIC}(K) = -2l(\widehat{\boldsymbol{\theta}}(K), \{\widehat{\mathbf{b}}_{Xi}(K), \hat{r}_i(K)\}; \{\hat{\boldsymbol{\beta}}_i, \boldsymbol{\Sigma}_i\}) + log(N) \cdot K, \tag{3.7}$$

where N is the minimum sample size of all GWAS datasets used in the model. We select $\hat{K} = \arg\min_{K \in \mathcal{K}} \text{BIC}(K)$ and $\widehat{\mathcal{V}} = \{i | \hat{r}_i(\hat{K}) = 0, i = 1, \ldots, m\}$. Then the final cMLE of $\boldsymbol{\theta}$ is $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\hat{K})$. The standard errors are calculated based on the observed

Fisher information matrix from the (profile) likelihood with all selected valid IVs (in $\widehat{\mathcal{V}}$) (Xue et al., 2021; Zhao et al., 2020). With $\widehat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_L)^T$ and corresponding standard errors, we draw inference based on the asymptotic normal distribution. We call this method **MVMR-cML-BIC**. We note that our proposed cMLE $\widehat{\boldsymbol{\theta}}$ also enjoys the nice statistical properties of estimation and selection consistency as its univariable counterpart UVMR-cML (Xue et al., 2021). Here we state the main conclusions with the proofs relegated to the Appendix.

Assumption 3.2

For every SNP $i = 1, \ldots, m$,

$$\hat{\boldsymbol{\beta}}_i = (\hat{\beta}_{X_1 i}, \ldots, \hat{\beta}_{X_L i}, \hat{\beta}_{Y i})^T \sim \mathcal{N}\left(\mathbf{b}_i = (b_{X_1 i}, \ldots, b_{X_L i}, \sum_{l=1}^{L} \theta_l b_{X_l i} + r_i)^T, \boldsymbol{\Sigma}_i\right),$$

with known covariance matrix $\boldsymbol{\Sigma}_i$. Furthermore, the m vectors $\{\hat{\boldsymbol{\beta}}_i\}_{i=1}^m$ are mutually independent.

Assumption 3.3

Let $N = \min(N_{X_1}, \ldots, N_{X_L}, N_Y)$. There exist positive constants c_1 and c_2 such that we have $c_1/N \leq \sigma^2_{X_l i} \leq c_2/N$, and $c_1/N \leq \sigma^2_{Y i} \leq c_2/N$, for $l = 1, \ldots, L$, $i = 1, \ldots, m$.

Theorem 3.2

With Assumption 3.1 to Assumption 3.3 satisfied, our proposed BIC consistently selects valid IVs, i.e. $P(\hat{K} = m - m_0) \to 1$ and $P(\widehat{\mathcal{V}} = \mathcal{V}^*) \to 1$ as $N \to \infty$. Furthermore, the proposed constrained maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$, combined with the use of the BIC, is consistent for the true causal parameter $\boldsymbol{\theta}^*$, and asymptotically normal with

$$\mathbf{V}^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

where $\mathbf{V} = \mathbb{E}[-\partial^2 \tilde{l}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}']$ is the expected Fisher information matrix for the profile log-likelihood that can be consistently estimated by its sample version.

For typical GWAS summary data with large sample sizes, Assumption 3.2 and Assumption 3.3 are reasonable. Based on the plurality condition, the range of K can be varied from 0, i.e. no invalid IV, up to $m - (L + 1)$. However in practice, we suggest first try a smaller range of K, for example from 0 to $m/2$, or based on other methods like MVMR-Lasso (Grant and Burgess, 2021), and one can keep expanding the range if the best K selected is on or close to the upper bound. One reason is that the proportion of invalid IVs is relatively low in many real data examples using UVMR (Xue et al., 2021; Lin et al., 2021), which is expected to be lower in MVMR when we explicitly include other exposures in the model. Moreover, this can speed up the computation of MVMR-cML dramatically, especially for the data perturbation version to be described next.

To better account for the uncertainty in model selection described above, we adopt the data perturbation approach (Xue et al., 2021); in a UVMR context, the data perturbation on GWAS summary data is shown to be equivalent to bootstrapping GWAS individual-level data (Lin et al., 2021). For the b-th perturbation, $b = 1, \ldots, B$, we generate the perturbed GWAS summary data:

$$\hat{\boldsymbol{\beta}}_i^{(b)} = (\hat{\beta}_{X_1 i}^{(b)}, \ldots, \hat{\beta}_{X_L i}^{(b)}, \hat{\beta}_{Yi}^{(b)})^T \sim \mathcal{N}\left(\hat{\boldsymbol{\beta}}_i = (\hat{\beta}_{X_1 i}, \ldots, \hat{\beta}_{X_L i}, \hat{\beta}_{Yi})^T, \boldsymbol{\Sigma}_i\right),$$

for $i = 1, \ldots, m$ independently. Then we apply MVMR-cML-BIC on the b-th perturbed sample and obtain $\widehat{\boldsymbol{\theta}}^{(b)}$. And we use the (element-wise) sample mean and sample covariance of $\widehat{\boldsymbol{\theta}}^{(1)}, \ldots, \widehat{\boldsymbol{\theta}}^{(B)}$ as the final estimate of $\boldsymbol{\theta}$ and its covariance matrix respectively. The number of perturbations B is suggested to be at least 100. We call this method **MVMR-cML-DP**.

3.2.4 Simulations

3.2.4.1 Comparison of MVMR-cML and other MVMR methods in the presence of pleiotropy

We first compare the performance of our proposed method MVMR-cML with other existing MVMR methods in the presence of pleiotropy. Following Grant and Burgess (2021), we simulated data as follows:

$$\begin{aligned}
\mathbf{U} &= \mathbf{G}\boldsymbol{\phi} + \mathbf{e}_U, \\
\mathbf{X}_l &= \mathbf{G}\boldsymbol{\gamma}_{X_l} + 0.25\mathbf{U} + \mathbf{e}_{X_l}, \quad l = 1, \ldots, 4, \\
\mathbf{Y} &= \sum_{l=1}^{4} \theta_l \cdot \mathbf{X}_l + \mathbf{G}\boldsymbol{\alpha} + \mathbf{U} + \mathbf{e}_Y,
\end{aligned}$$

where each component of $\mathbf{e}_U, \mathbf{e}_{X_l}, \mathbf{e}_Y$ was independently and identically distributed (iid) as the standard normal $\mathcal{N}(0,1)$, $\mathbf{G}$ was the genotype matrix with 20 IVs, each generated independently from a binomial distribution with minor allele frequency (MAF) 0.3 and $\gamma_{X_l i} \sim$ Uniform$(0, 0.22)$ iid. Two sets of values for the causal effects were considered: (1) $(\theta_1, \theta_2, \theta_3, \theta_4) = (0.2, 0.1, 0.3, 0.4)$ and (2) $(\theta_1, \theta_2, \theta_3, \theta_4) = (0, -0.1, 0.1, 0.2)$. We considered three scenarios with different patterns of pleiotropy, and considered 30% and 50% invalid IVs for each scenario:

S1: Balanced and uncorrelated pleiotropy (with InSIDE satisfied): $\boldsymbol{\phi}$ was set to 0 and the α_i's corresponding to invalid IVs were generated from $\mathcal{N}(0, 0.2^2)$.

S2: Directional and uncorrelated pleiotropy (with InSIDE satisfied): $\boldsymbol{\phi}$ was set to 0 and the α_i's corresponding to invalid IVs were generated from $\mathcal{N}(0.1, 0.2^2)$.

S3: Directional and correlated pleiotropy (with InSIDE violated): for invalid IVs, $\boldsymbol{\phi}_i$'s were generated from Uniform$(0, 0.1)$ and the α_i's were generated from $\mathcal{N}(0, 0.2^2)$.

We also considered a scenario (S4) where some IVs violated assumption MV-A1 and some had uncorrelated pleiotropy ($\boldsymbol{\phi} = 0$). Specifically, we randomly selected 30% SNPs and set $\gamma_{X_1 i} = \gamma_{X_2 i} = \gamma_{X_3 i} = \gamma_{X_4 i} = 0$. We also randomly selected 30% SNPs that had pleiotropic effect with $\alpha_i \sim \mathcal{N}(0.1, 0.1^2)$. Note that these two sets of (30%) invalid IVs could have overlap.

The GWAS sample sizes for all traits were set as $N = 50\,000$. The GWAS summary statistics for the 4 risk factors were calculated using the same 50 000 individuals, and the outcome GWAS summary statistics were calculated using the other non-overlapping 50 000 individuals. We calculated $\rho_{ll'}$ as the sample correlation between $\mathbf{X}_l$ and $\mathbf{X}_{l'}$, and ρ_{lY} was set to zero in MVMR-cML.

For each simulation set-up, we ran 500 replications and compared MVMR-cML-BIC, MVMR-cML-DP and some existing MVMR methods including MVMR-IVW (Burgess and Thompson, 2015), MVMR-Egger (Rees et al., 2017), MVMR-median, MVMR-robust and MVMR-Lasso (Grant and Burgess, 2021).

3.2.4.2 Comparison of MVMR-cML and other MVMR methods in the presence of weak IVs

In this section, we simulated a MVMR model with two exposures and 45 IVs. We considered two different ways to generate the strengths of SNP-exposure associations. In the first scenario, some *weak* IVs were simulated with relatively small $\gamma_{X_1 i}$ and $\gamma_{X_2 i}$, i.e., the exposures were marginally weakly associated with the IVs. In the second scenario, some marginally strong but *conditionally weak* IVs were generated by introducing a strong correlation between $\gamma_{X_1 i}$ and $\gamma_{X_2 i}$; that is, each exposure was strongly associated with the IVs marginally, but only weakly associated conditional on the other exposure. In both scenarios, we maintained the two-sample conditional F-statistic proposed in Sanderson et al. (2021) smaller than 10, the conventional cut-off for detecting weak IVs in MVMR. Details of simulation set-ups are given in the

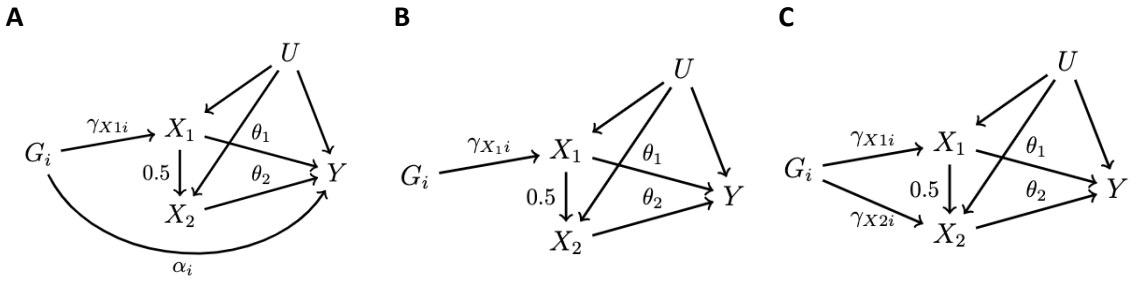


Figure 3.2: Three scenarios of simulated genetic instruments for (A) $i = 1, \ldots, K$;(B) $i = K+1, \ldots, K+m_1$; and(C) $i = K+m_1+1, ..., 20.\gamma_{X_1 i}, \gamma_{X_2 i} \overset{iid.}{\sim} \mathcal{U}(0, 0.22)$, $\alpha_i \sim \mathcal{N}(0.1, 0.2^2)$, and $(\theta_1, \theta_2) = (0.1, 0.2)$.

Appendix B.4. For each set-up, we ran 500 replications and compared the proposed MVMR-cML with other existing MVMR methods.

3.2.4.3 Mediation analysis: MVMR versus UVMR

In this simulation, we illustrate two advantages of using MVMR over UVMR. First, when there are causal relationships among the risk factors, MVMR can distinguish a direct effect from a total effect of one risk factor on the outcome, while UVMR estimates the total effect on the outcome. Second, ignoring causal pathways via other risk factors may lead to the violation of the IV assumptions for some IVs, and thus of the plurality condition in UVMR.

We considered $L = 2$ risk factors X_1 and X_2 in Figure 3.1B, with a causal effect $X_1 \to X_2$, and simulated the GWAS summary statistics according to the three scenarios in Figure 3.2. More details are given in the Appendix B.5.

We applied MVMR-cML and other MVMR methods with both X_1 and X_2 as exposures, and applied UVMR-cML and UVMR-IVW with only X_1 as the exposure. In particular, the SNPs in Figure 3.2A had direct/pleiotropic effects on Y, not going through either X_1 or X_2, thus they were invalid IVs in both UVMR and MVMR analyses; the SNPs in Figure 3.2B did not have direct effects on Y conditional on X_1,

so they were valid IVs in both UVMR (for X_1 only) and MVMR analyses; the SNPs in Figure 3.2C were directly associated with X_2 in addition to a mediating route via X_1, thus they were invalid IVs in UVMR but not in MVMR. We considered different combinations of (K, m_1) from $\{(0,1),(0,14),(0,18),(3,11),(3,15),(10,4)\}$. We note that when $K = 0$ (i.e., no SNPs generated from scenario Figure 3.2A), all IVs were valid in MVMR analysis; when $(K, m_1) = (0, 1)$, there was only one valid IV for UVMR, and thus the plurality condition for UMVR (Xue et al., 2021) was violated.

3.2.5 GWAS data

In the real data application, we focused on assessing the causal effects of 8 cardiometabolic risk factors on CAD (van der Harst and Verweij, 2018). The 8 risk factors were triglyceride (TG), low-density lipoprotein cholesterol (LDL), high-density lipoprotein cholesterol (HDL) (Willer et al., 2013), body-mass index (BMI) (Locke et al., 2015), height (Wood et al., 2014), fasting glucose (FG) (Scott et al., 2012), systolic blood pressure (SBP) and diastolic blood pressure (DBP) (Evangelou et al., 2018). The sample size of the 9 GWAS datasets ranges from around 130 000 to around 750 000. We extracted the SNPs as IVs using the `mv_extract_exposures` function in R package `TwoSampleMR` (Hemani et al., 2018b). After harmonizing data, we retained 201 IVs; for each of the 8 exposures, the number of significantly associated IVs ranged from 14 to 79.

3.3 Results

3.3.1 Simulations: better performance of MVMR-cML over other MVMR methods

3.3.1.1 Robustness to pleiotropy

Here we show some representative results for estimation and inference of the direct causal effect of exposure X_1 on the outcome, θ_1, while the results for $\theta_2, \theta_3, \theta_4$ are provided in the Supplementary with the same conclusion. Table 3.1 and Table 3.2 show the results of different MVMR methods under scenarios S1-S3 when $\theta_1 = 0.2$ and $\theta_1 = 0$ respectively. And Table 3.3 shows the results under scenario S4.

First, across all considered scenarios, our proposed method MVMR-cML had the smallest bias and mean squared error (MSE) among all the methods, followed by MVMR-Lasso and MVMR-robust. The advantage of our methods over other methods was more pronounced when there were a higher proportion of invalid IVs at 50%. MVMR-Egger and MVMR-IVW had much less precise estimates than other methods, leading to much lower power and larger MSEs in general. Second, MVMR-cML-DP was the only method that could control the type-I error below the nominal level 5% (Table 3.2), followed by MVMR-IVW, MVMR-Egger and MVMR-robust, while at the same time MVMR-cML-DP had much higher power than the other three methods (Table 3.1). On the other hand, MVMR-Lasso had the largest inflated type-I error rate. Lastly, though MVMR-cML-BIC had good performance in estimation in terms of a small bias and MSE, its inference might not be satisfactory. In the presence of invalid IVs, due to the uncertainty in model selection, its mean standard error was lower than the sample standard deviation of the estimates, leading to an anti-conservative coverage rate and an inflated type-I error rate (Table 3.2). And the issue was more severe when 50% IVs were invalid. Nevertheless, the data perturbation

approach was able to alleviate this problem, giving a satisfactory coverage rate close to the nominal level 95% with a well controlled type-I error rate. In addition, MVMR-cML-DP had a smaller MSE than MVMR-cML-BIC in the presence of 50% invalid IVs. Hence, overall, MVMR-cML-DP performed best and would be recommended.

Table 3.1: Mean and standard deviation (SD) of estimates, mean standard error (SE) and coverage rate (Cov), power, mean squared error (MSE) when $m = 20$ and $\theta_1 = 0.2$.

	30% invalid						50% invalid					
Method	Mean	SD	SE	Cov	Power	MSE	Mean	SD	SE	Cov	Power	MSE
Scenario 1: Balanced pleiotropy, InSIDE met												
MVMR-cML-BIC	0.200	0.070	0.054	0.872	0.912	0.005	0.195	0.227	0.064	0.678	0.794	0.052
MVMR-cML-DP	0.199	0.073	0.084	0.974	0.706	0.005	0.193	0.182	0.192	0.968	0.322	0.033
MVMR-Egger	0.206	0.440	0.428	0.950	0.092	0.194	0.156	0.599	0.559	0.940	0.070	0.360
MVMR-IVW	0.207	0.372	0.361	0.944	0.114	0.138	0.163	0.502	0.475	0.938	0.080	0.253
MVMR-Lasso	0.194	0.095	0.058	0.894	0.870	0.009	0.205	0.255	0.084	0.686	0.714	0.065
MVMR-median	0.198	0.114	0.083	0.922	0.730	0.013	0.209	0.279	0.121	0.740	0.580	0.078
MVMR-robust	0.200	0.085	0.096	0.921	0.661	0.007	0.191	0.381	0.413	0.927	0.156	0.145
Scenario 2: Directional pleiotropy, InSIDE met												
MVMR-cML-BIC	0.200	0.076	0.055	0.860	0.896	0.006	0.203	0.223	0.067	0.694	0.796	0.050
MVMR-cML-DP	0.199	0.075	0.085	0.958	0.712	0.006	0.204	0.190	0.173	0.930	0.370	0.036
MVMR-Egger	0.204	0.509	0.484	0.946	0.086	0.259	0.207	0.649	0.601	0.926	0.098	0.422
MVMR-IVW	0.267	0.432	0.409	0.926	0.130	0.191	0.297	0.555	0.510	0.914	0.128	0.317
MVMR-Lasso	0.203	0.113	0.063	0.840	0.844	0.013	0.244	0.389	0.093	0.580	0.714	0.153
MVMR-median	0.210	0.134	0.090	0.904	0.686	0.018	0.249	0.402	0.142	0.658	0.568	0.164
MVMR-robust	0.204	0.109	0.102	0.899	0.662	0.012	0.267	0.429	0.474	0.919	0.136	0.188
Scenario 3: Directional pleiotropy, InSIDE violated												
MVMR-cML-BIC	0.198	0.069	0.055	0.872	0.908	0.005	0.203	0.201	0.065	0.698	0.780	0.041
MVMR-cML-DP	0.200	0.069	0.083	0.974	0.714	0.005	0.202	0.159	0.180	0.950	0.354	0.025
MVMR-Egger	0.221	0.507	0.452	0.922	0.138	0.258	0.265	0.609	0.569	0.912	0.098	0.376
MVMR-IVW	0.211	0.433	0.384	0.916	0.128	0.187	0.268	0.508	0.484	0.926	0.108	0.263
MVMR-Lasso	0.202	0.141	0.059	0.878	0.902	0.020	0.221	0.334	0.089	0.664	0.718	0.112
MVMR-median	0.206	0.163	0.087	0.924	0.688	0.027	0.227	0.337	0.133	0.762	0.576	0.114
MVMR-robust	0.202	0.089	0.100	0.907	0.659	0.008	0.243	0.404	0.405	0.929	0.190	0.165

Table 3.2: Mean and standard deviation (SD) of estimates, mean standard error (SE) and coverage rate (Cov), type-I error, mean squared error (MSE) when $m = 20$ and $\theta_1 = 0$.

	30% invalid						50% invalid					
Method	Mean	SD	SE	Cov	Type I	MSE	Mean	SD	SE	Cov	Type I	MSE
Scenario 1: Balanced pleiotropy, InSIDE met												
MVMR-cML-BIC	0.001	0.056	0.044	0.882	0.118	0.003	-0.004	0.149	0.054	0.700	0.300	0.022
MVMR-cML-DP	0.001	0.056	0.069	0.974	0.026	0.003	-0.003	0.127	0.143	0.952	0.048	0.016
MVMR-Egger	0.008	0.441	0.427	0.946	0.054	0.194	-0.043	0.601	0.559	0.930	0.070	0.363
MVMR-IVW	0.008	0.371	0.361	0.944	0.056	0.138	-0.037	0.504	0.475	0.936	0.064	0.256
MVMR-Lasso	-0.003	0.085	0.049	0.906	0.094	0.007	0.007	0.243	0.074	0.684	0.316	0.059
MVMR-median	0.000	0.096	0.070	0.920	0.080	0.009	0.005	0.276	0.108	0.736	0.264	0.076
MVMR-robust	0.003	0.077	0.079	0.907	0.093	0.006	-0.006	0.381	0.417	0.921	0.079	0.145
Scenario 2: Directional pleiotropy, InSIDE met												
MVMR-cML-BIC	-0.002	0.055	0.045	0.880	0.120	0.003	0.009	0.156	0.055	0.746	0.254	0.024
MVMR-cML-DP	-0.001	0.057	0.069	0.970	0.030	0.003	0.006	0.144	0.144	0.954	0.046	0.021
MVMR-Egger	0.006	0.508	0.483	0.942	0.058	0.258	0.009	0.648	0.601	0.930	0.070	0.420
MVMR-IVW	0.067	0.431	0.408	0.918	0.082	0.190	0.096	0.553	0.510	0.910	0.090	0.315
MVMR-Lasso	0.002	0.104	0.052	0.868	0.132	0.011	0.042	0.380	0.084	0.590	0.410	0.146
MVMR-median	0.008	0.122	0.075	0.916	0.084	0.015	0.046	0.392	0.128	0.640	0.360	0.156
MVMR-robust	0.000	0.078	0.076	0.906	0.094	0.006	0.060	0.422	0.478	0.921	0.079	0.182
Scenario 3: Directional pleiotropy, InSIDE violated												
MVMR-cML-BIC	0.002	0.062	0.045	0.876	0.124	0.004	0.007	0.197	0.054	0.706	0.294	0.039
MVMR-cML-DP	0.001	0.056	0.067	0.966	0.034	0.003	0.007	0.136	0.145	0.938	0.062	0.019
MVMR-Egger	0.024	0.506	0.451	0.918	0.082	0.257	0.066	0.610	0.569	0.908	0.092	0.377
MVMR-IVW	0.012	0.433	0.383	0.914	0.086	0.187	0.068	0.508	0.483	0.926	0.074	0.263
MVMR-Lasso	0.004	0.116	0.051	0.874	0.126	0.014	0.029	0.326	0.079	0.674	0.326	0.107
MVMR-median	0.007	0.154	0.074	0.922	0.078	0.024	0.024	0.329	0.119	0.740	0.260	0.109
MVMR-robust	0.003	0.070	0.078	0.887	0.113	0.005	0.049	0.396	0.413	0.925	0.075	0.159

Table 3.3: Results for scenario S4. Mean and standard deviation (SD) of estimates, mean standard error (SE) and coverage rate (Cov), power/type-I error, mean squared error (MSE).

	$\theta_1 = 0.2$						$\theta_1 = 0$					
Method	Mean	SD	SE	Cov	Power	MSE	Mean	SD	SE	Cov	Type I	MSE
MVMR-cML-BIC	0.205	0.142	0.067	0.802	0.814	0.020	0.013	0.216	0.058	0.806	0.194	0.047
MVMR-cML-DP	0.207	0.129	0.137	0.960	0.496	0.017	0.014	0.137	0.140	0.970	0.030	0.019
MVMR-Egger	0.259	0.313	0.325	0.948	0.152	0.101	0.054	0.324	0.330	0.936	0.064	0.108
MVMR-IVW	0.266	0.294	0.309	0.946	0.176	0.091	0.062	0.297	0.315	0.936	0.064	0.092
MVMR-Lasso	0.213	0.179	0.078	0.762	0.756	0.032	0.019	0.208	0.069	0.786	0.214	0.044
MVMR-median	0.226	0.181	0.112	0.836	0.618	0.033	0.022	0.226	0.099	0.812	0.188	0.052
MVMR-robust	0.204	0.169	0.118	0.851	0.632	0.029	0.023	0.166	0.095	0.814	0.186	0.028

3.3.1.2 Robustness to weak IVs

In this section, we considered a scenario with marginally (and conditionally) weak IVs (Table 3.4), and a scenario with marginally strong but conditionally weak IVs (Table 3.5), while all IVs were valid with no pleiotropic effects. The average conditional F-statistics across 500 simulation replicates were 6.73 and 6.70 for the two exposures respectively in the first scenario, and 9.35 and 9.38 in the second scenario. We can see that, only MVMR-cML yielded (almost) unbiased estimates (with the smallest MSE) in both scenarios, suggesting its robustness to weak instrument bias. We note that data perturbation might produce a slightly conservative confidence interval as observed in the previous section. In the null causal effect case (Table B.7 and B.8), all methods gave unbiased results.

Table 3.4: Simulation results for weak IVs. Mean and standard deviation (SD) of estimates, mean standard error (SE) and coverage rate (Cov), power, mean squared error (MSE).

	$\theta_1 = 0.5$						$\theta_2 = -0.3$					
Method	Mean	SD	SE	Cov	Power	MSE	Mean	SD	SE	Cov	Power	MSE
MVMR-cML-BIC	0.501	0.081	0.073	0.924	1.000	0.007	-0.301	0.082	0.073	0.922	0.972	0.007
MVMR-cML-DP	0.500	0.081	0.102	0.984	1.000	0.007	-0.301	0.081	0.103	0.984	0.878	0.007
MVMR-Egger	0.449	0.097	0.094	0.904	0.992	0.012	-0.225	0.078	0.077	0.824	0.826	0.012
MVMR-IVW	0.433	0.068	0.066	0.824	1.000	0.009	-0.233	0.068	0.067	0.820	0.920	0.009
MVMR-Lasso	0.430	0.080	0.158	0.868	0.764	0.011	-0.233	0.080	0.165	0.872	0.502	0.011
MVMR-median	0.430	0.083	0.090	0.906	1.000	0.012	-0.233	0.083	0.091	0.894	0.748	0.011
MVMR-robust	0.432	0.069	0.069	0.828	1.000	0.009	-0.233	0.069	0.069	0.818	0.900	0.009

* The average conditional F-statistics across 500 replicates for X_1 and X_2 are 6.73 and 6.70 (with SD 1.44 and 1.41) respectively.

Table 3.5: Simulation results for conditionally weak IVs. Mean and standard deviation (SD) of estimates, mean standard error (SE) and coverage rate (Cov), power, mean squared error (MSE).

	$\theta_1 = 0.5$						$\theta_2 = -0.3$					
Method	Mean	SD	SE	Cov	Power	MSE	Mean	SD	SE	Cov	Power	MSE
MVMR-cML-BIC	0.501	0.064	0.059	0.940	1.000	0.004	-0.301	0.045	0.042	0.932	1.000	0.002
MVMR-cML-DP	0.503	0.065	0.081	0.984	1.000	0.004	-0.302	0.046	0.057	0.982	1.000	0.002
MVMR-Egger	0.450	0.057	0.057	0.854	1.000	0.006	-0.265	0.040	0.040	0.846	1.000	0.003
MVMR-IVW	0.450	0.055	0.056	0.846	1.000	0.006	-0.265	0.039	0.039	0.850	1.000	0.003
MVMR-Lasso	0.449	0.066	0.134	0.869	0.860	0.007	-0.265	0.047	0.094	0.869	0.801	0.003
MVMR-median	0.449	0.071	0.073	0.882	1.000	0.008	-0.265	0.050	0.051	0.882	1.000	0.004
MVMR-robust	0.450	0.058	0.058	0.822	1.000	0.006	-0.265	0.041	0.041	0.826	1.000	0.003

* The average conditional F-statistics across 500 replicates for X_1 and X_2 are 9.35 and 9.38 (with SD 1.95 and 1.96) respectively.

As shown in the Appendix B.4, we also considered other scenarios with both weak and pleiotropic IVs. Under these more challenging situations, all MVMR methods yielded biased estimates, but MVMR-cML-DP was least biased and controlled type-I error best.

3.3.2 Simulations: advantages of MVMR-cML for mediation analysis

In this simulation, we considered a scenario for mediation analysis with two exposures, one of which (X_2) mediated the effect of the other (X_1) on the outcome. We compared MVMR-cML with other MVMR methods and two representative UVMR methods. We show some representative results here while more results are given in the Supplementary. First, Table 3.6 shows the results with the number of invalid IVs $K = 10$ and the number of valid IVs (for both MVMR and UVMR) $m_1 = 4$ (while the remaining 6 IVs being valid for MVMR but not for UVMR). With 50% invalid IVs for MVMR, all other methods yielded severely biased estimates for the direct causal effect of each exposure on the outcome, while only MVMR-cML gave (almost) unbiased estimates. As shown in the Supplementary, when the number of invalid IVs K decreased, the performance of other robust MVMR methods improved and was more similar to that of MVMR-cML. In the UVMR analysis, since most of the IVs were invalid, UVMR-IVW yielded much more biased estimates for the total causal effect of X_1 than those of UVMR-cML. As shown in the Supplementary (Tables B.14-B.17), with more valid IVs, UVMR-cML yielded almost unbiased estimates for the total causal effect.

Second, Table 3.7 shows the results when $K = 0$ and $m_1 = 1$, in which case only one IV was valid in UVMR analysis, thus the plurality condition required by UVMR-cML was violated. MVMR-cML had similar performance to MVMR-IVW, which can be considered as the oracle estimator as all IVs were valid in MVMR analysis. All MVMR methods yielded (almost) unbiased estimates for the direct causal effect. On the other hand, as the plurality condition was violated in UVMR, both UVMR-cML and UVMR-IVW performed poorly in estimating the total effect of X_1.

Note that we did not apply UVMR with exposure X_2 because all of the IVs

Table 3.6: Mean and standard deviation (SD) of estimates, mean standard error (SE) and power when $K = 10$, $m_1 = 4$. The total causal effect of X_1 is $\theta_{1T} = 0.2$.

	$\theta_1 = 0.1$ ($\theta_{1T} = 0.2$)				$\theta_2 = 0.2$			
Method	Mean	SD	SE	Power	Mean	SD	SE	Power
MVMR-cML-DP	0.102	0.029	0.032	0.866	0.199	0.025	0.027	0.992
MVMR-Egger	0.389	0.580	0.638	0.062	-0.261	0.344	0.589	0.004
MVMR-IVW	0.699	0.438	0.495	0.270	-0.142	0.288	0.571	0.002
MVMR-Lasso	0.201	0.290	0.035	0.920	0.144	0.181	0.029	0.946
MVMR-median	0.227	0.290	0.055	0.900	0.130	0.180	0.042	0.892
MVMR-robust	0.269	0.339	0.491	0.123	0.105	0.204	0.287	0.319
UVMR-cML-DP	0.209	0.064	0.025	0.928	NA	NA	NA	NA
UVMR-IVW	0.591	0.249	0.266	0.628	NA	NA	NA	NA

Table 3.7: Mean and standard deviation (SD) of estimates, mean standard error (SE) and power when $K = 0$, $m_1 = 1$. The total causal effect of X_1 is $\theta_{1T} = 0.2$.

	$\theta_1 = 0.1$ ($\theta_{1T} = 0.2$)				$\theta_2 = 0.2$			
Method	Mean	SD	SE	Power	Mean	SD	SE	Power
MVMR-cML-DP	0.101	0.020	0.021	0.992	0.199	0.014	0.015	1.000
MVMR-Egger	0.100	0.021	0.020	0.988	0.199	0.017	0.017	1.000
MVMR-IVW	0.100	0.019	0.018	0.994	0.199	0.014	0.013	1.000
MVMR-Lasso	0.100	0.023	0.045	0.729	0.199	0.017	0.031	0.980
MVMR-median	0.101	0.023	0.025	0.980	0.199	0.017	0.018	1.000
MVMR-robust	0.100	0.020	0.019	0.994	0.199	0.014	0.014	1.000
UVMR-cML-DP	0.298	0.045	0.019	1.000	NA	NA	NA	NA
UVMR-IVW	0.344	0.029	0.031	1.000	NA	NA	NA	NA

were invalid for UVMR, though the total effect of X_2 on the outcome was the same as its direct effect. This example illustrates an advantage of MVMR over UVMR: the valid IV assumption MV-A3 is weaker than the corresponding UV-A3. On the other hand, in the case of $K = 0$, $m_1 = 20$ and $\theta_1 = 0$ with a true causal graph of $G_i \longrightarrow X1 \longrightarrow X2 \longrightarrow Y$ for all i, there is no valid IV for X_2 in MVMR while all the IVs are valid in UVMR for either X_1 or X_2, illustrating that the valid IV assumption MV-A1 is stronger than UV-A1.

3.3.3 Real data application: the causal effects of cardiometabolic risk factors on coronary artery disease

In this section, we studied the causal effects of the 8 cardio-metabolic risk factors on CAD. The correlation matrix for the SNP-trait association estimates was estimated using bivariate LDSC as discussed in Section 2.3 (with details provided in Appendix B.6). We first applied MVMR-cML (and MVMR-IVW as the standard method) using the 8 risk factors as exposures and CAD as the outcome. Then we applied UVMR-cML on each risk factor-CAD pair using the set of IVs significantly associated (p-value $<$ 5e-8) with the corresponding risk factor (Figure 3.3). We calculated the conditional F-statistics (Sanderson et al., 2021) for each of the eight exposures. As shown in Figure 3.3, all of the conditional F-statistics were larger than 10 except for BMI (8.15). This suggests that the weak instrument bias should not be severe in this analysis, and as shown in the previous simulations, our proposed method was expected to be robust. Furthermore, as discussed later, estimating the direct effect of BMI was potentially problematic and would not be a focus here.

3.3.3.1 Direct causal effects estimated by MVMR

The results are shown in Figure 3.3, and we summarize a few main findings here. First, MVMR methods suggested a null effect of HDL on CAD after adjustment for other 7 risk factors, while UVMR-cML suggested a protective effect. Second, the positive effect of DBP on CAD diminished in MVMR-cML, but stayed nearly significant in MVMR-IVW. There are several possible reasons for a risk factor to show an effect in UVMR analysis but not in MVMR. The first is that some genetic variants for that risk factor might have pleiotropic effects, leading to biased inference in UVMR. The second is that the effect of that risk factor on CAD is mediated through other risk factors in the model. The third is possible loss of power in MVMR. We will take

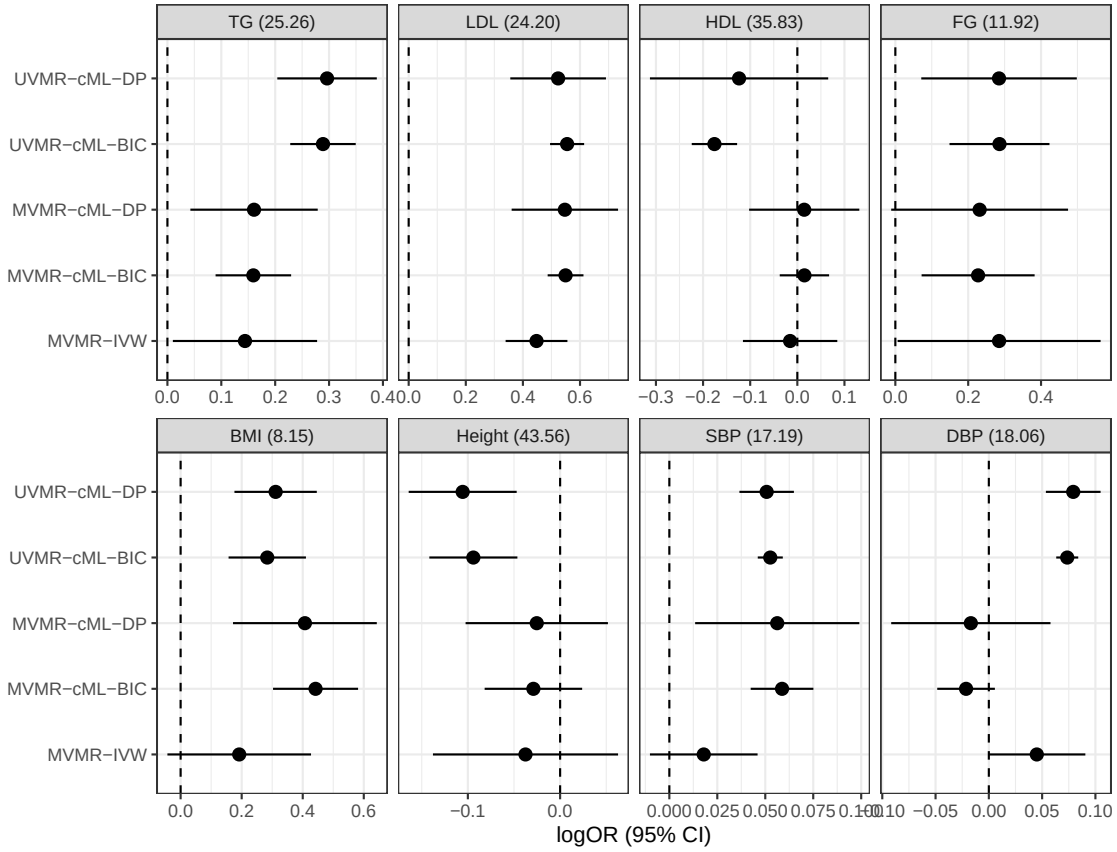


Figure 3.3: The estimated effects (and 95% confidence intervals) of each of the 8 risk factors on CAD by various UVMR and MVMR methods. The conditional F-statistic is given in the parentheses following each exposure name.

a closer look at these possibilities in the following sections. Moreover, we note that the GWAS summary statistics for SBP and DBP were obtained after adjusting for BMI. While using covariate-adjusted SNP-trait associations may lead to bias in MR analyses (Hartwig et al., 2021), as discussed in Gilbody et al. (2022), including the covariate, BMI in this case, as an additional exposure in an MVMR analysis can recover the direct causal effects of the exposures of interest, DBP and SBP in this case. However, the estimated direct effect of the covariate (BMI) on the outcome (CAD) was potentially biased and should be interpreted with caution.

MVMR-cML-BIC identified 10 invalid IVs out of the 201 IVs used in the MVMR

analysis. We calculated the conditional F-statistics based on the selected set of valid IVs. The results were similar to those of using the whole set of 201 IVs: only BMI had a conditional F-statistic smaller than 10 (Table B.19), suggesting no severe issue of multicollinearity. Lastly, to detect possible outlying or influential IVs, Cook's distance for each IV in the MVMR-IVW model was calculated. None of them exceeded the recommended threshold of the median of the corresponding F-distribution (Cook, 1977; Zuber et al., 2020), suggesting that there was no influential point in the analysis (Figure B.2). We also applied a leave-one-out analysis with MVMR-cML-DP, reaching the same conclusion (Figures B.3 and B.4).

3.3.3.2 Causal effects of lipids on CAD

The causal effect of various lipid fractions, including LDL, HDL and TG, on CAD is an important issue that has been studied widely. While increased LDL doubtlessly has a deleterious causal effect on CAD, the roles of TG and HDL are still under debate. Numerous MR analyses have been conducted to investigate this issue. In particular, in standard UVMR analyses using genome-wide significant variants for HDL, a protective role of HDL on CAD has been suggested by several UVMR methods (Xue et al., 2021). However, this could be due to the fact that many variants associated with HDL are also associated with other risk factors like LDL, TG and BMI, suggesting possible pleiotropic effects. For example, Holmes et al. (2015) showed that, when an unrestricted allele score was used, a protective effect of HDL was identified, while when a restricted allele score (by removing any SNPs associated with either of the other two lipid traits) was used, no significant effect was found.

In our example here, we reached a similar conclusion. First, in the univariable analysis, UVMR-cML-BIC suggested that HDL had a protective effect on CAD while the result of UVMR-cML-DP was not significant. As noted in Xue et al. (2021), model selection based on BIC might miss some invalid IVs due to their small effect

sizes, leading to inflated type-I errors, especially when pleiotropic effects were weak, while data perturbation could help. We found that out of the 26 IVs associated with HDL, 17 were also associated with at least one of the three other likely causal risk factors, including LDL, TG and BMI (at the significance level of 5e-8). However, UVMR-cML-BIC only identified 8 of them, leading to possibly biased inference. On the other hand, for UVMR-cML-DP, 12 of them were identified as invalid at least 10 times out of the 100 perturbed datasets. Though UVMR-cML-DP performed more robustly than UVMR-cML-BIC in the presence of many invalid IVs, it also yielded a much wider confidence interval, even than that of MVMR-cML. We also applied UVMR-cML using only 9 IVs for HDL by removing the 17 potentially invalid IVs. In this case, both UVMR-cML-BIC and UVMR-cML-DP suggested a null effect of HDL on CAD (though it could be due to the reduce power of using less SNPs). On the other hand, both MVMR-cML-BIC and MVMR-cML-DP suggested that both TG and LDL, but not HDL, had a causal effect on CAD after adjusting for other risk factors included in the model; this could be partly due to the relaxed non-pleiotropy assumption in MVMR.

3.3.3.3 Diminished causal effects of DBP on CAD after accounting for SBP

As shown in Figure 3.3, DBP had a significant positive effect on CAD in UVMR-cML, but the effect completely diminished in MVMR-cML. This was concordant with some previous studies finding that the effect of DBP on CAD disappeared after adjusting for SBP (Arvanitis et al., 2021; Levin et al., 2020). First, there were 79 SNPs associated with DBP, and 63 of them were also associated with at least one of the other 7 risk factors (at the significance level of 5e-8). We removed these 63 variants and performed UVMR-cML on DBP to CAD with the remaining set of 16 SNPs. Unlike what we observed for HDL, the UVMR-cML result still showed a significant positive effect,

suggesting that the diminished effect of DBP in MVMR (but significant in UVMR) might be due to other reasons besides the presence of pleiotropic variants.

Next, we applied MVMR-cML-DP with a subset of exposures to investigate some potential mediating effects. We separately used each of the other 7 risk factors along with DBP as exposures, and CAD as the outcome. As shown in the left panel in Figure 3.4, the effect of DBP on CAD after adjusting for SBP changed from that in UVMR-cML-DP, while adjusting for any one of the other 6 risk factors (Height, BMI, FG, HDL, LDL and TG) did not lead much changes. Alternatively, we started from the "full" model as shown in Figure 3.3 with all 8 risk factors and deleted each of the 7 risk factors (other than DBP) respectively. The right panel in Figure 3.4 told a similar story: after SBP was removed from the full model, DBP still had an effect on CAD, though slightly smaller than that in UVMR, after adjusting for the other 6 remaining risk factors. This suggested that, a mediating effect of DBP on CAD was likely via SBP. However, we should be cautious in interpreting any mediating effect as it relies on the correct inference by both UVMR and MVMR. It is also possible that, the total effect of DBP on CAD shown in UVMR was due to the fact that many of the IVs for DBP had pleiotropic effects on SBP (or a closely related trait), some of which may be too weak to be detectable, so the significant result with the restricted set of 16 IVs might still be biased, and the different results between UVMR-cML and MVMR-cML could be due to their differing adjustments for pleiotropy via SBP.

Lastly, we applied other robust MVMR methods. All methods gave results in line with that of MVMR-cML, suggesting a null direct effect of DBP on CAD, and their point estimates were also in the same direction (see Figure B.1).

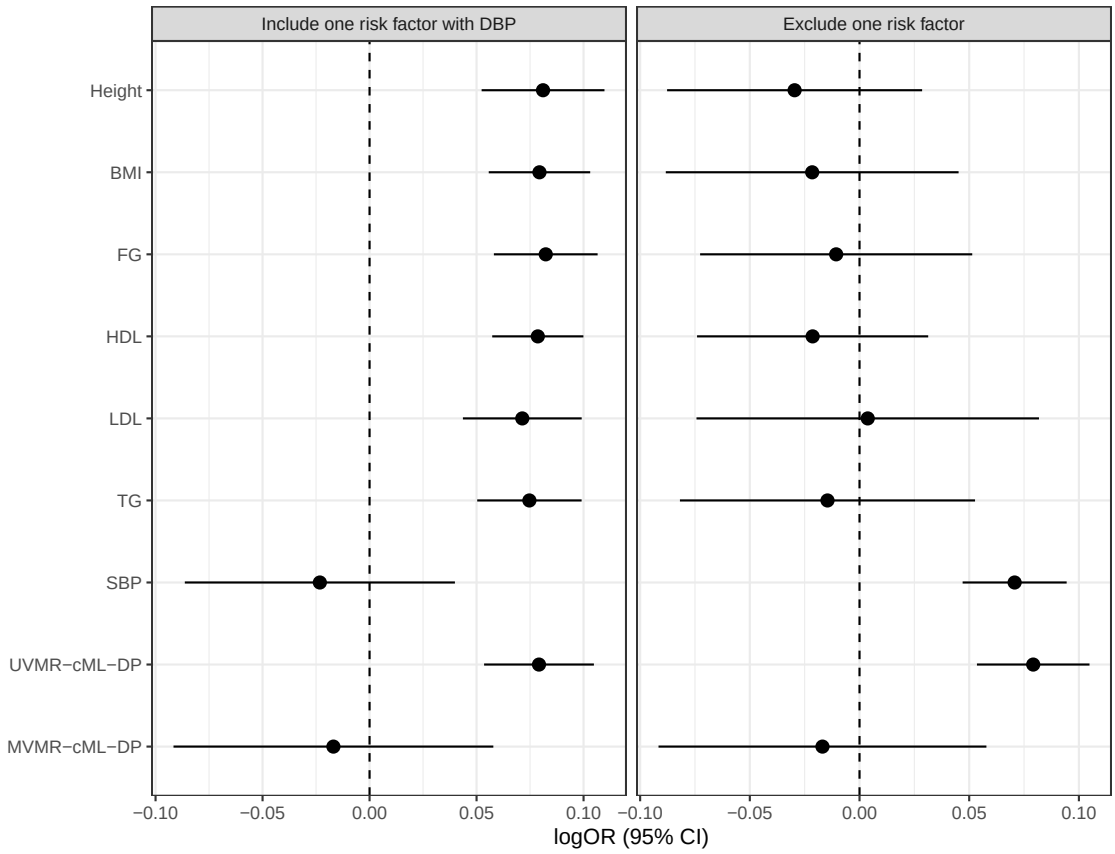


Figure 3.4: The estimated effects (and 95% confidence intervals) of DBP on CAD using various sets of exposures by MVMR-cML-DP. Left panel corresponds to the sets of 2 exposures (DBP plus one of the other 7 risk factors). Right panel corresponds to the sets of the 6 exposures after excluding one of 7 risk factors marked out in the left.Results from UVMR-cML-DP and MVMR-cML-DP in Figure 3.3 are also added at bottom for comparison.

3.4 Discussion

We have proposed a robust and efficient multivariable Mendelian randomization method based on constrained maximum likelihood, called MVMR-cML. It is an important and useful extension of the UVMR-cML (Xue et al., 2021). We have shown in both simulations and a real data application that, compared to its univariable counterpart (and other UVMR methods), MVMR-cML has two main advantages (while maintaining the major advantages of UVMR-cML, such as its estimation efficiency and robustness). First, MVMR estimates the direct effect of an exposure on the outcome after accounting for the other exposures included in the model, while UVMR only estimates the total effect. When there are causal pathways among the exposures, the direct effect of an exposure is in general different from its total effect on an outcome. Second, MVMR can account for some known pleiotropic effects through other exposures included in the MVMR model, making it more robust to pleiotropy than its univariable counterparts. This is important especially when (putative causal) risk factors, e.g. various lipids such as TG, LDL and HDL, share some genetic associations. Although UVMR-cML and some other UVMR methods allow for invalid IVs under the plurality condition, it is still possible that some invaid IVs with weak pleiotropic effects cannot be selected out, leading to incorrect inference. On the other hand, an application of MVMR requires some assumptions beyond those of UVMR; in particular, MVMR requires at least one valid IV for each exposure; that is, the matrix of the marginal IV-exposure associations has a full column rank. We have given an example in mediation analysis, in which UVMR, but not MVMR, would work. More generally, as in multiple regression, if the marginal association matrix is nearly singular, multicollinearity may lead to unstable estimates and thus inflated type I errors and loss of power in MVMR; a conditional F-test has been proposed to detect such a case (Sanderson et al., 2021). In addition, all the current MVMR

methods require the use of (nearly) independent IVs, leading to fewer IVs being used (than that in UVMR), and thus possibly exacerbating the issue of multicollinearity. An extreme example is when there is only one valid IV for one exposure in MVMR; removing this IV would lead to the non-identifiable MVMR model and thus unreliable estimates. These issues can be regarded as a price we pay for using MVMR,

There are a few advantages of our proposed method over other existing MVMR methods. First, MVMR-cML has nice statistical properties such as selection consistency, estimation consistency and asymptotic normality with strong theoretical support. Second, as highlighted in our simulation studies, MVMR-cML was shown to have the smallest bias and MSE under various pleiotropic and/or weak IV settings among all MVMR methods being compared. In particular, MVMR-cML-DP consistently performed best, especially in controlling type I errors. Compared to MVMR-cML-BIC, MVMR-cML-DP accounts for model selection uncertainties ignored by the former, often performs better for finite samples and thus is recommended. Third, most of the MVMR methods are based on the two-sample MR setup, which means that all the exposure and outcome GWAS were performed in non-overlapping (and unrelated) samples. Our likelihood framework can account for overlapping samples by taking into account of the correlations among genetic associations with the exposures and outcome. This avoids sample splitting into non-overlapping subsets for analysis with reduced power (e.g. in Davies et al. (2019); Sanderson et al. (2019)). Fourth, although in this paper we focused on testing for each direct effect separately, it is straightforward to test for a subset of multiple causal effects jointly in MVMR-cML. Furthermore, unlike some other MVMR methods, we do not need the assumption of no measurement errors of SNP-exposure associations (NOME) in our method by directly accounting for the variation of the estimated associations ($\hat{\beta}_{X_l i}$) in the likelihood framework. This contributes to its robustness to weak IV biases (Sanderson et al., 2021; Burgess and Thompson, 2011; Carter et al., 2021).

There are some limitations with our method. First, as in other MVMR methods, we assume a linear and homogeneous effect of each exposure on the outcome (Grant and Burgess, 2021; Carter et al., 2021). When a linear effect is absent, it does not necessarily imply no causal effects. Some exposures might have a non-linear effect on the outcome; for example, a U-shaped or J-shaped effect of DBP on CAD has been reported in previous studies (Beddhu et al., 2018; Liang and Wang, 2021). Second and importantly, as in multiple regression, if MVMR fails to detect a direct effect, it could be due to its low power with multicollinearity if many related exposures are included. Third, despite the two advantages of MVMR over UVMR mentioned earlier, when an estimated direct effect in MVMR differs from an estimated total effect in UVMR, it may be difficult to distinguish whether such a difference is due to mediating effects of other exposures, or to pleiotropic effects of genetic variants, on the outcome, or both. Finally, though emerging as a powerful tool for causal inference with observational data, various (MV)MR methods depend on their own assumptions as well as on the quality of the genetic variants as IVs, and more generally, on the GWAS data being used. MR cannot completely replace traditional experimental studies while triangulation through more applications of various MR methods to real data would be worthwhile and warranted.

Chapter 4

A constrained maximum likelihood-based cis-Mendelian randomization method robust to invalid instruments: application to drug target discovery

4.1 Introduction

Mendelian randomization is a widely-used method that uses genetic variants as instrumental variables to infer the causal relationship between a pair of traits, one called exposure and another as an outcome. Since genetic variants are randomly allocated and fixed at conception, it minimizes the risk of confounding and reverse causation with observational data (Lawlor et al., 2008; Sanderson et al., 2022). Within the instrumental variable (IV) regression framework, MR also requires the three valid IV assumptions: the IVs must be (1) associated with the exposure; (2) independent of any confounders of the exposure-outcome relationship; (3) not associated with the outcome conditional on the exposure and confounders. Subject to these assumptions, MR can provide evidence of a putative causal relationship between the exposure and the outcome, and the inverse variance weighting (IVW) method (Burgess et al., 2013)

can be applied. However, only the first IV assumption can be tested and is relatively easy to be satisfied in practice by using genome-wide significant SNPs associated with the exposure; in contrast, the second and third assumptions cannot be tested empirically and are likely to be violated due to the presence of wide-spread (horizontal) pleiotropy. Numerous MR methods have been proposed to handle the presence of horizontal pleiotropy (Slob and Burgess, 2020; Boehm and Zhou, 2022), but most of them require the use of independent IVs as conducted in most MR analyses.

Meanwhile, there has been a growing interest in MR studies focusing on a small genetic region using some local and correlated cis-SNPs as IVs, known as *cis*-MR. One of the most promising applications of *cis*-MR is for drug target discovery, including drug target prioritization, validation or repositioning (Schmidt et al., 2020; Zhao et al., 2022; Gkatzionis et al., 2023). Drug-target MR uses a protein (as potential drug target) or its downstream biomarker as the exposure, and corresponding cis-SNPs of the gene encoding the protein as IVs. Despite the significance of such an analysis, it still depends crucially on the three valid IV assumptions. Although one can first perform linkage disequilibrium (LD) clumping to obtain some (approximately) independent IVs before applying one or more of existing robust MR methods based on independent IVs, it would lead to possibly severe loss of power due to only one or few independent SNPs remaining; in fact, with only one or two SNPs, many robust MR methods cannot be applied. As an alternatively, we has to use multiple correlated IVs in *cis*-MR. However, only few cis-MR methods are robust to the violation of the IV assumptions. Perhaps the most widely used cis-MR method is the generalized MR-IVW (Burgess et al., 2016b), which uses generalized linear regression to account for LD (among correlated SNPs) but assumes all IVs are valid. Similarly, a generalized version of MR-Egger (Burgess and Thompson, 2017) and another closely related method, LDA-Egger (Barfield et al., 2018), have also been proposed, which require a stringent (INSIDE) assumption on the relationship between the unknown

IV strengths and pleiotropic effects; furthermore, more generally, MR-Egger is low powered and sensitive to the coding of the SNPs (Lin et al., 2022) There are several recently proposed Bayesian methods to account for both LD and horizontal pleiotropy, such as MR-LDP (Cheng et al., 2020), MR-Corr2 (Cheng et al., 2022a), MR-CUE (Cheng et al., 2022b), RBMR (Wang et al., 2022a). All these methods impose different modeling assumptions on the distribution of the latent/hidden pleiotropic effects, while some can only handle either correlated pleiotropy or uncorrelated pleiotropy, but not both.

In this work, we propose a robust *cis*-MR method called cisMR-cML, extending MR-cML (Xue et al., 2021) to allow for correlated SNPs as IVs. As its previous version with independent IVs, cisMR-cML is robust to violation of any one, two or all three IV assumptions, imposing minimum modeling assumptions with strong theoretical support. We point out two main differences between cisMR-cML and MR-cML. First, in cisMR-cML we model conditional/joint SNP effects, instead of marginal effects as directly available from GWAS summary data. Second, when selecting SNPs as IVs for the exposure, we include not only SNPs associated with the exposure, but also those associated with the outcome. These two differences are important: due to the use of correlated SNPs, failing to do so may lead to *all* invalid IVs. These two differences are largely neglected in the literature, but have significant implications for extensions of other robust MR methods to using correlated IVs, such as the median-based (Bowden et al., 2016) and mode-based (Hartwig et al., 2017) MR methods. We also implement a new computational algorithm to enhance the likelihood of obtaining a globally optimal solution to our non-convex (invalid) IV selection problem (Zhu et al., 2020). We show the robustness of the proposed method to the presence of invalid IVs in simulation studies, and illustrate the severe consequence of using only SNPs that are conditionally associated with the exposure. Lastly, we demonstrate the effectiveness of the proposed method in two real data

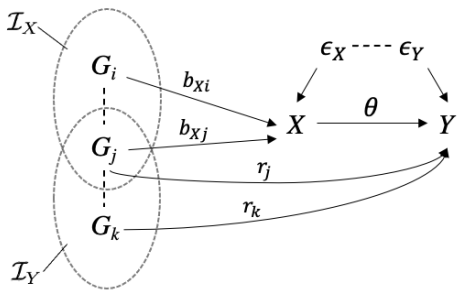


Figure 4.1: Causal diagram showing the relationships among multiple SNPs (G), the exposure (X) and the outcome (Y).

applications for drug target discovery for coronary artery disease (CAD). In the first application, we use downstream biomarkers to serve as a proxy of the perturbation of a drug target, while in the second one, we perform a proteome-wide analysis to identify some proteins as potential drug target for CAD.

4.2 Methods

4.2.1 Model

Based on Figure 4.1 with m SNPs (G_1,...,G_m), assuming that both the genotypes and the traits have been standardized, the true models for the exposure X and the outcome Y are

$$\mathbf{X} = \sum_{i=1}^{m} b_{Xi}\mathbf{G}_i + \boldsymbol{\epsilon}_X, \tag{4.1}$$

$$\mathbf{Y} = \theta\mathbf{X} + \sum_{i=1}^{m} r_i\mathbf{G}_i + \boldsymbol{\epsilon}_Y, \tag{4.2}$$

where $\boldsymbol{\epsilon}_X$ and $\boldsymbol{\epsilon}_Y$ are random error terms independent of SNPs $\{\mathbf{G}_i\}_{i=1}^m$. In general, $\boldsymbol{\epsilon}_X$ and $\boldsymbol{\epsilon}_Y$ are correlated due to the presence of hidden confounding. Plugging Eq (4.1)

in Eq (4.2), we have

$$\mathbf{Y} = \sum_{i=1}^{m}(\theta b_{Xi} + r_i)\mathbf{G}_i + (\theta\boldsymbol{\epsilon}_X + \boldsymbol{\epsilon}_Y) := \sum_{i=1}^{m} b_{Yi}\mathbf{G}_i + \boldsymbol{\epsilon}_Y^*, \tag{4.3}$$

$$b_{Yi} = \theta b_{Xi} + r_i, \tag{4.4}$$

where $\{b_{Xi}\}_{i=1}^m$ and $\{b_{Yi}\}_{i=1}^m$ are the **joint** effects of the m SNPs on the exposure and the outcome respectively. Note that in Figure 4.1, for SNPs $i \in \mathcal{I}_X$, $b_{Xi} \neq 0$; and for SNPs $i \in \mathcal{I}_Y$, $r_i \neq 0$. Based on the three valid IV assumptions, a valid IV is a SNP with $b_{Xi} \neq 0$ and $r_i = 0$, i.e. $i \in \mathcal{I}_X \setminus \mathcal{I}_Y$. As discussed in Theorem 1 in Guo et al. (2018), Eq (4.4) is identifiable if and only if the valid IVs form the largest group of IVs sharing the same causal parameter value (i.e., the plurality condition).

Eq (4.1) and Eq (4.2) give the joint association models of m SNPs on the exposure and on the outcome respectively, while typically in GWAS, the *marginal* associations of each SNP with the exposure and with the outcome are modeled:

$$\mathbf{X} = b_{Xi}^*\mathbf{G}_i + \epsilon_X^*, \qquad \mathbf{Y} = b_{Yi}^*\mathbf{G}_i + \epsilon_Y^*.$$

Accordingly, we denote such GWAS summary statistics of the exposure and the outcome as $\{\hat{\beta}_{Xi}^*, \hat{\beta}_{Yi}^*, \sigma_{Xi}^*, \sigma_{Yi}^*\}_{i=1}^m$. The goal in this paper is to infer the causal effect θ in the possible presence of invalid IVs using publicly available GWAS summary data.

Next we build a bridge between the conditional/joint SNP-effect estimates and the marginal estimates. From Eq (4.1), we have $\hat{\boldsymbol{\beta}}_X = (\mathbf{G}_X^T\mathbf{G}_X)^{-1}(\mathbf{G}_X^T\mathbf{X})$, where $\mathbf{G}_X$ is the $N_X \times m$ standardized genotype matrix and $\mathbf{X}$ is the standardized phenotype vector of length N_X. Then $\hat{\boldsymbol{\beta}}_X = \mathbf{R}_X^{-1}\hat{\boldsymbol{\beta}}_X^*$, where $\mathbf{R}_X$ is the LD matrix and $\hat{\boldsymbol{\beta}}_X^*$ is the GWAS estimate. Denote the (estimated) covariance matrix of the joint effect estimate ($\hat{\boldsymbol{\beta}}_X$) as $\boldsymbol{\Sigma}_X = (\mathbf{G}_X^T\mathbf{G}_X)^{-1}\tilde{\sigma}_X^2$, where $\tilde{\sigma}_X^2 = ||\mathbf{X} - \mathbf{G}_X\hat{\boldsymbol{\beta}}_X||^2/N_X = 1 - \hat{\boldsymbol{\beta}}_X^T\hat{\boldsymbol{\beta}}_X^*$. In practice, if the individual genotype matrix for calculating $\mathbf{R}_X$ is not available, the LD matrix

can be estimated using some publicly available reference panel denoted as $\mathbf{R}$, where $\mathbf{R} \approx \mathbf{R}_X$. Furthermore, if the GWAS estimates are not calculated on the standardized genotypes and phenotype, they can be approximated as $\hat{\beta}^*_{Xi}/\sqrt{\hat{\beta}^{*2}_{Xi} + (N_X - 2) \cdot \sigma^{*2}_{Xi}}$ (Xue and Pan, 2020). Similarly we can estimate $\hat{\boldsymbol{\beta}}_Y$ and $\boldsymbol{\Sigma}_Y$ from $\hat{\boldsymbol{\beta}}^*_Y$ and $\boldsymbol{\sigma}^*_Y$.

In the following sections, unless specified otherwise, we assume the asymptotic distribution of the joint-effect estimates $\hat{\boldsymbol{\beta}}_X \sim \mathcal{MVN}(\mathbf{b}_X, \boldsymbol{\Sigma}_X)$ and $\hat{\boldsymbol{\beta}}_Y \sim \mathcal{MVN}(\mathbf{b}_Y, \boldsymbol{\Sigma}_Y)$ with $\mathbf{b}_X = (b_{X1}, \ldots, b_{Xm})^T$ and $\mathbf{b}_Y = (b_{Y1}, \ldots, b_{Ym})^T$.

4.2.2 The MR-IVW and MR-Egger methods

Inverse-variance weighted method (MR-IVW) (Burgess et al., 2013) and MR-Egger regression (Bowden et al., 2015) are two of the most widely used MR methods. These two methods are often discussed in the context of independent SNPs/IVs, where MR-IVW and MR-Egger can be regarded as weighted linear regression (with weights equal to σ^{*-2}_{Yi}) of $\hat{\boldsymbol{\beta}}^*_Y$ on $\hat{\boldsymbol{\beta}}^*_X$, without and with the intercept term respectively. Both methods have previously been extended to account for correlated IVs and we will next give a brief overview of the existing methods, while details can be found in their corresponding references.

4.2.2.1 Generalized IVW and Egger

To account for the correlations among IVs, MR-IVW and MR-Egger have been extended based on generalized weighted linear regression (Burgess et al., 2016b; Burgess and Thompson, 2017), and we rfer them as GIVW and GEgger throughout the paper.

The GIVW and GEgger estimators are:

$$\hat{\theta}_{GIVW} = (\hat{\boldsymbol{\beta}}_X^{*T} \boldsymbol{\Omega}_Y^{-1} \hat{\boldsymbol{\beta}}_X^*)^{-1} (\hat{\boldsymbol{\beta}}_X^{*T} \boldsymbol{\Omega}_Y^{-1} \hat{\boldsymbol{\beta}}_Y^*), \tag{4.5}$$

$$\begin{pmatrix} \hat{\alpha}_{GEgger} \\ \hat{\theta}_{GEgger} \end{pmatrix} = [(\mathbf{1}, \hat{\boldsymbol{\beta}}_X^*)^T \boldsymbol{\Omega}_Y^{-1} (\mathbf{1}, \hat{\boldsymbol{\beta}}_X^*)]^{-1} [(\mathbf{1}, \hat{\boldsymbol{\beta}}_X^*)^T \boldsymbol{\Omega}_Y^{-1} \hat{\boldsymbol{\beta}}_Y^*],$$

$$\hat{\theta}_{GEgger} = \frac{(\mathbf{1}^T \boldsymbol{\Omega}_Y^{-1} \mathbf{1})(\hat{\boldsymbol{\beta}}_X^{*T} \boldsymbol{\Omega}_Y^{-1} \hat{\boldsymbol{\beta}}_Y^*) - (\mathbf{1}^T \boldsymbol{\Omega}_Y^{-1} \hat{\boldsymbol{\beta}}_X^*)(\mathbf{1}^T \boldsymbol{\Omega}_Y^{-1} \hat{\boldsymbol{\beta}}_Y^*)}{(\mathbf{1}^T \boldsymbol{\Omega}_Y^{-1} \mathbf{1})(\hat{\boldsymbol{\beta}}_X^{*T} \boldsymbol{\Omega}_Y^{-1} \hat{\boldsymbol{\beta}}_X^*) - (\mathbf{1}^T \boldsymbol{\Omega}_Y^{-1} \hat{\boldsymbol{\beta}}_X^*)(\mathbf{1}^T \boldsymbol{\Omega}_Y^{-1} \hat{\boldsymbol{\beta}}_X^*)}, \tag{4.6}$$

where $\boldsymbol{\Omega}_Y = \mathbf{R} \cdot \boldsymbol{\sigma}_Y^* \boldsymbol{\sigma}_Y^{*T}$ is the covariance matrix of $\hat{\boldsymbol{\beta}}_Y^*$, " $\cdot$ " is the element-wise multiplication and $\mathbf{R}$ is the LD matrix among the m SNPs. When the SNPs are independent, i.e., $\boldsymbol{\Omega}_Y$ becomes a diagonal matrix with the ith diagonal element σ_{Yi}^{*2}, GIVW and GEgger become the original MR-IVW and MR-Egger. The GIVW and GEgger methods are implemented in the R package `MendelianRandomization` (Yavorska and Burgess, 2017).

4.2.2.2 LD-Aware (LDA) IVW and Egger

LD-Aware (LDA) MR-IVW and MR-Egger are two other variants of MR-IVW and MR-Egger proposed by Barfield et al. (2018) to account for LD among IVs, and we refer them to LIVW and LEgger throughout. These LDA-estimators are very similar to the generalized MR-IVW (GIVW) and MR-Egger (GEgger), except that the input data is the conditional estimates $\{\hat{\beta}_X, \hat{\beta}_Y\}$, instead of the marginal estimates $\{\hat{\beta}_X^*, \hat{\beta}_Y^*\}$:

$$\hat{\theta}_{LIVW} = (\hat{\boldsymbol{\beta}}_X^T \boldsymbol{\Sigma}_Y^{-1} \hat{\boldsymbol{\beta}}_X)^{-1} (\hat{\boldsymbol{\beta}}_X^T \boldsymbol{\Sigma}_Y^{-1} \hat{\boldsymbol{\beta}}_Y), \tag{4.7}$$

$$\hat{\theta}_{LEgger} = \frac{(\mathbf{1}^T \boldsymbol{\Sigma}_Y^{-1} \mathbf{1})(\hat{\boldsymbol{\beta}}_X^T \boldsymbol{\Sigma}_Y^{-1} \hat{\boldsymbol{\beta}}_Y) - (\mathbf{1}^T \boldsymbol{\Sigma}_Y^{-1} \hat{\boldsymbol{\beta}}_X)(\mathbf{1}^T \boldsymbol{\Sigma}_Y^{-1} \hat{\boldsymbol{\beta}}_Y)}{(\mathbf{1}^T \boldsymbol{\Sigma}_Y^{-1} \mathbf{1})(\hat{\boldsymbol{\beta}}_X^T \boldsymbol{\Sigma}_Y^{-1} \hat{\boldsymbol{\beta}}_X) - (\mathbf{1}^T \boldsymbol{\Sigma}_Y^{-1} \hat{\boldsymbol{\beta}}_X)(\mathbf{1}^T \boldsymbol{\Sigma}_Y^{-1} \hat{\boldsymbol{\beta}}_X)}, \tag{4.8}$$

with $\hat{\boldsymbol{\beta}}_X = \mathbf{R}^{-1}\hat{\boldsymbol{\beta}}_X^*$, $\hat{\boldsymbol{\beta}}_Y = \mathbf{R}^{-1}\hat{\boldsymbol{\beta}}_Y^*$ and $\boldsymbol{\Sigma}_Y = \mathbf{R}^{-1}\boldsymbol{\Omega}_Y\mathbf{R}^{-1}$. Comparing Eq (4.5) with Eq (4.7), and Eq (4.6) with Eq (4.8), we see that $\hat{\theta}_{LIVW} = \hat{\theta}_{GIVW}$, but in general $\hat{\theta}_{LEgger} \neq \hat{\theta}_{GEgger}$. The LDA-Egger method can be implemented with the code provided by the original authors.

While the extensions of MR-IVW and MR-Egger allow for correlations among IVs, they inherit the same limitations in their corresponding original versions. For example, both IVW and Egger may yield biased causal inference unless all IVs are valid or under some stringent (so-called InSIDE) condition between the instrument strengths and theiir direct effects (Burgess and Thompson, 2017; Lin et al., 2022).

4.2.3 New method: cisMR-cML

In this section, we propose a robust cisMR method accounting for possible violations of any invalid IV assumptions. It is an extension of MR-cML (Xue et al., 2021) to allow for correlated SNPs. Suppose we have the estimated *joint/conditional* associations of the m SNPs with the exposure as $\hat{\boldsymbol{\beta}}_X = (\hat{\beta}_{X1}, \ldots, \hat{\beta}_{Xm})^T$ and its covariance matrix $\boldsymbol{\Sigma}_X$, and those with the outcome as $\hat{\boldsymbol{\beta}}_Y = (\hat{\beta}_{Y1}, \ldots, \hat{\beta}_{Ym})^T$ and $\boldsymbol{\Sigma}_Y$, which can be obtained from the GWAS summary statistics as described before. The model for the proposed **cisMR-cML** is

$$\begin{aligned} \hat{\boldsymbol{\beta}}_X &\sim \mathcal{MVN}(\mathbf{b}_X, \boldsymbol{\Sigma}_X), \\ \hat{\boldsymbol{\beta}}_Y &\sim \mathcal{MVN}(\mathbf{b}_Y = \theta\mathbf{b}_X + \mathbf{r}, \boldsymbol{\Sigma}_Y), \end{aligned} \tag{4.9}$$

where $\boldsymbol{\Sigma}_X$ and $\boldsymbol{\Sigma}_Y$ are the covariance of $\hat{\boldsymbol{\beta}}_X$ and $\hat{\boldsymbol{\beta}}_Y$ respectively, θ is the causal effect of interest, $\mathbf{b}_X$ is a vector of the unknown joint effects of m SNPs on the exposure, and $\mathbf{r}$ is a vector of the unknown direct effects on the outcome not mediated through the exposure. Note that $\mathbf{r}$ captures both the correlated and uncorrelated (horizontal) pleiotropic effects. Assuming the independence between the exposure GWAS dataset

and the outcome GWAS dataset, we have the log-likelihood for the proposed model Eq (4.9) (up to some constants):

$$l(\theta, \mathbf{b}_X, \mathbf{r}; \hat{\boldsymbol{\beta}}_X, \hat{\boldsymbol{\beta}}_Y, \boldsymbol{\Sigma}_X, \boldsymbol{\Sigma}_Y) = -\frac{1}{2}\Big[(\hat{\boldsymbol{\beta}}_X - \mathbf{b}_X)^T \boldsymbol{\Sigma}_X^{-1} (\hat{\boldsymbol{\beta}}_X - \mathbf{b}_X) + (\hat{\boldsymbol{\beta}}_Y - \theta \mathbf{b}_X - \mathbf{r})^T \boldsymbol{\Sigma}_Y^{-1} (\hat{\boldsymbol{\beta}}_Y - \theta \mathbf{b}_X - \mathbf{r})\Big]. \tag{4.10}$$

Under the constraint that the number of invalid IVs is $0 \leq K < m-1$, we obtain the constrained maximum likelihood estimator (cMLE) by solving

$$\min_{\theta, \mathbf{b}_X, \mathbf{r}} - l(\theta, \mathbf{b}_X, \mathbf{r}; \hat{\boldsymbol{\beta}}_X, \hat{\boldsymbol{\beta}}_Y, \boldsymbol{\Sigma}_X, \boldsymbol{\Sigma}_Y) \text{ subject to } \sum_{i=1}^{m} I(r_i \neq 0) = K, \tag{4.11}$$

where $I(\cdot)$ is the indicator function, and K is a tuning parameter representing the unknown number of invalid IVs to be determined by a model selection criterion as to be discussed. The pluraity condition implies $K < m - 1$.

To solve Eq (4.11), the coordinate-descent algorithm in the original MR-cML Xue et al. (2021) may not be directly applicable. In particular, in Step 1 in "Computation" Section in Xue et al. (2021) to select invalid IVs, the top K IVs with the largest increase of the log-likelihood (before and after being selected as invalid) are selected independently and sequentially. This is reasonable when all IVs are independent. But with correlated IVs, such an approach may be problematic, as the pleiotropic effects r_i's are intertwined with each other in the second term in Eq (4.10). We will propose a new computational algorithm to obtain the cMLE next.

4.2.3.1 Computation

To solve Eq (4.11) with a given K, we propose a new algorithm based on the idea of exchanging some variables in the selected set (called active set) with those in the

unselected set (called inactive set). We first define the active set $\mathcal{A}$ be the set of invalid IVs with $r_i \neq 0$, i.e. $\mathcal{A} = \{i : r_i \neq 0\}$, and the inactive set $\mathcal{I} = \mathcal{A}^C = \mathcal{U} \backslash \mathcal{A}$ as the set of remaining IVs with $r_i = 0$, where $\mathcal{U} = \{1, 2, \ldots, m\}$. Denote $\mathbf{a}^{\mathcal{A}}$ be the vector whose i-th entry is a_i if $i \in \mathcal{A}$, and zero otherwise.

Given an index set $\mathcal{A} \subsetneq \{1, \ldots, m\}$ with cardinality $|\mathcal{A}| = K$, we can compute $\{\hat{\theta}, \hat{\mathbf{b}}_X, \hat{\mathbf{r}}\} = \arg\max_{\theta, \mathbf{b}_X, \mathbf{r}} l(\theta, \mathbf{b}_X, \mathbf{r})$ under the constraint of $r_i = 0$ if $i \notin \mathcal{A}$. In order to decide which variables should be exchanged, we define the backward sacrifice and forward sacrifice as follows:

1. **Backward sacrifice**: For any $i \in \mathcal{A}$, the magnitude of discarding invalid IV i is:
$$\xi_i = l(\hat{\theta}, \hat{\mathbf{b}}_X, \hat{\mathbf{r}}^{\mathcal{A}}) - l(\hat{\theta}, \hat{\mathbf{b}}_X, \hat{\mathbf{r}}^{\mathcal{A}\backslash\{i\}})$$

2. **Forward sacrifice**: For any $i \in \mathcal{I}$, the magnitude of adding IV i in the invalid-IV set is:
$$\zeta_i = l(\hat{\theta}, \hat{\mathbf{b}}_X, \hat{\mathbf{r}}^{\mathcal{A}} + \hat{\mathbf{t}}^{\{i\}}) - l(\hat{\theta}, \hat{\mathbf{b}}_X, \hat{\mathbf{r}}^{\mathcal{A}}),$$
where $\hat{\mathbf{t}}^{\{i\}} = \arg\max_{\mathbf{t}} l(\hat{\theta}, \hat{\mathbf{b}}_X, \hat{\mathbf{r}}^{\mathcal{A}} + \mathbf{t}^{\{i\}})$.

Intuitively, for $i \in \mathcal{A}$ (or $i \in \mathcal{I}$), a larger ξ_i (or ζ_i) suggests that selecting the ith IV as invalid contributes more importantly to increasing the likelihood, or the ith IV is potentially invalid. While these two types of sacrifices are not comparable (in terms of their numeric value), we can exchange some IVs in $\mathcal{A}$ with small ξ_i and some IVs in $\mathcal{I}$ with large ζ_i.

Next, we introduce the algorithm for solving Eq (4.11) with a given number of invalid IVs $0 < K < m - 1$. Algorithm 1 demonstrates the procedure to improve the selection of invalid IVs by exchanging IVs in $\mathcal{A}$ and $\mathcal{I}$.

Algorithm 1 $f(\theta, \mathbf{b}_X, \mathbf{r}, \mathcal{A}, \mathcal{I}, k_{\max})$

Require: $\theta, \mathbf{b}_X, \mathbf{r}, \mathcal{A}, \mathcal{I}$.

$L_0, L \leftarrow l(\theta, \mathbf{b}_X, \mathbf{r}; \hat{\boldsymbol{\beta}}_X, \hat{\boldsymbol{\beta}}_Y, \boldsymbol{\Sigma}_X, \boldsymbol{\Sigma}_Y)$

$\mathbf{t} \leftarrow (\hat{\boldsymbol{\beta}}_Y - \theta \mathbf{b}_X)^T \boldsymbol{\Sigma}_Y^{-1} / diag(\boldsymbol{\Sigma}_Y^{-1})$

$\xi_i \leftarrow L - l(\theta, \mathbf{b}_X, \mathbf{r}^{\mathcal{A}\setminus\{i\}})$ for $i \in \mathcal{A}$

$\zeta_i \leftarrow l(\theta, \mathbf{b}_X, \mathbf{r}^{\mathcal{A}} + \mathbf{t}^{\{i\}}) - L$ for $i \in \mathcal{I}$

for $k = 1, \ldots, k_{\max}$ **do**

$\mathcal{A}_k \leftarrow \{i \in \mathcal{A} : \sum_{j \in \mathcal{A}} I(\xi_i \geq \xi_j) \leq k\}$

$\mathcal{I}_k \leftarrow \{i \in \mathcal{I} : \sum_{j \in \mathcal{I}} I(\zeta_i \leq \zeta_j) \leq k\}$

$\tilde{\mathcal{A}}_k \leftarrow (\mathcal{A}\setminus\mathcal{A}_k) \cup \mathcal{I}_k$, $\tilde{\mathcal{I}}_k \leftarrow (\mathcal{I}\setminus\mathcal{I}_k) \cup \mathcal{A}_k$

$\{\tilde{\theta}, \tilde{\mathbf{b}}_X, \tilde{\mathbf{r}}\} \leftarrow \arg\max_{\theta, \mathbf{b}_X, \mathbf{r}} l(\theta, \mathbf{b}_X, \mathbf{r})$ under the constraint of $r_i = 0$ if $i \notin \tilde{\mathcal{A}}_k$

if $l(\tilde{\theta}, \tilde{\mathbf{b}}_X, \tilde{\mathbf{r}}) > L$ **then**

$(\hat{\theta}, \hat{\mathbf{b}}_X, \hat{\mathbf{r}}, \hat{\mathcal{A}}, \hat{\mathcal{I}}) \leftarrow (\tilde{\theta}, \tilde{\mathbf{b}}_X, \tilde{\mathbf{r}}, \tilde{\mathcal{A}}_k, \tilde{\mathcal{I}}_k)$

$L \leftarrow l(\tilde{\theta}, \tilde{\mathbf{b}}_X, \tilde{\mathbf{r}})$

end if

end for

if $L_0 > L$ **then**

$(\hat{\theta}, \hat{\mathbf{b}}_X, \hat{\mathbf{r}}, \hat{\mathcal{A}}, \hat{\mathcal{I}}) \leftarrow (\theta, \mathbf{b}_X, \mathbf{r}, \mathcal{A}, \mathcal{I})$

end if

return $(\hat{\theta}, \hat{\mathbf{b}}_X, \hat{\mathbf{r}}, \hat{\mathcal{A}}, \hat{\mathcal{I}})$

Then we can use the above algorithm to update the set of invalid IVs $\mathcal{A}$ iteratively until Eq (4.11) cannot be improved by exchanging of IVs in $\mathcal{A}$ and $\mathcal{I}$.

Algorithm 2

Require: $\hat{\boldsymbol{\beta}}_X, \hat{\boldsymbol{\beta}}_Y, \boldsymbol{\Sigma}_X, \boldsymbol{\Sigma}_Y, K, \theta^{(0)}, \mathbf{b}_X^{(0)}$

$\mathbf{t}^{(0)} \leftarrow \hat{\boldsymbol{\beta}}_Y - \theta^{(0)}\mathbf{b}_X^{(0)}$

$\mathcal{A}^{(0)} \leftarrow \{i : \sum_{j=1}^{m} I(|t_i| \leq |t_j|) \leq K\}$, $\mathcal{I}^{(0)} \leftarrow (\mathcal{A}^{(0)})^C$

$\mathbf{r}^{(0)} \leftarrow (\mathbf{t}^{(0)})^{\mathcal{A}^{(0)}}$

$k_{\max} \leftarrow \min(K, |\mathcal{I}^{(0)}|)$

for $t = 0, 1, \ldots$ **do**

$(\theta^{(t+1)}, \mathbf{b}_X^{(t+1)}, \mathbf{r}^{(t+1)}, \mathcal{A}^{(t+1)}, \mathcal{I}^{(t+1)}) \leftarrow f(\theta^{(t)}, \mathbf{b}_X^{(t)}, \mathbf{r}^{(t)}, \mathcal{A}^{(t)}, \mathcal{I}^{(t)}, k_{\max})$

if $\mathcal{A}^{(t+1)} = \mathcal{A}^{(t)}$ **then** stop

end if

end for

$\hat{\mathcal{A}} \leftarrow \mathcal{A}^{(t+1)}$, $\hat{\mathcal{I}} \leftarrow \mathcal{I}^{(t+1)}$

$\{\hat{\theta}, \hat{\mathbf{b}}_X, \hat{\mathbf{r}}\} \leftarrow \arg\max_{\theta, \mathbf{b}_X, \mathbf{r}} l(\theta, \mathbf{b}_X, \mathbf{r})$ under the constraint of $r_i = 0$ if $i \notin \hat{\mathcal{A}}$

return $(\hat{\theta}, \hat{\mathbf{b}}_X, \hat{\mathbf{r}}, \hat{\mathcal{A}}, \hat{\mathcal{I}})$

In practice, besides the default starting value of $\theta^{(0)} = 0, \mathbf{b}_X^{(0)} = \mathbf{0}$, we can use multiple random starts $\theta^{(0)}, \mathbf{b}_X^{(0)}$ and take the estimate which yields the largest likelihood among the multiple starting points as the final cMLE under the constraint of K invalid IVs.

Denote the estimates for a given K as $\hat{\theta}(K), \hat{\mathbf{b}}_X(K), \hat{\mathbf{r}}(K), \hat{\mathcal{A}}(K), \hat{\mathcal{I}}(K)$. We select K from a candidate set $\mathcal{K} \subseteq \{0, 1, \ldots, m-2\}$ based on the following Bayesian information criterion (BIC):

$$\text{BIC}(K) = -2l(\hat{\theta}(K), \hat{\mathbf{b}}_X(K), \hat{\mathbf{r}}(K); \hat{\boldsymbol{\beta}}_X, \hat{\boldsymbol{\beta}}_Y, \boldsymbol{\Sigma}_X, \boldsymbol{\Sigma}_Y) + log(N) \cdot K,$$

where $N = \min(N_X, N_Y)$. Then $\hat{K} = \arg\min_{K \in \mathcal{K}} \text{BIC}(K)$, $\hat{\mathcal{I}} = \hat{\mathcal{I}}(\hat{K})$, and the final causal estimate of Eq (4.11) is $\hat{\theta} = \hat{\theta}(\hat{K})$. In the proposed algorithm, the

resulting constrained maximum likelihood estimator is the same as the maximum profile likelihood estimator being applied to all IVs in $\hat{\mathcal{I}}$. The standard error of $\hat{\theta}$ can be estimated based on the observed Fisher information from the profile likelihood with IVs in $\hat{\mathcal{I}}$. With $\hat{\theta}$ and its corresponding standard error, the statistical inference is drawn based on the standard normal distribution, the theory of which is to be established in Section 4.2.3.4

The validity of the above inference relies on the selection consistency of invalid IVs (with $r_i \neq 0$), which may not always be realized with finite samples. Instead, to account for the uncertainty/variation in model selection, we will use data perturbation as before for better finite-sample statistical inference (Xue et al., 2021). As shown in Lin et al. (2021), the data perturbation procedure (on a GWAS summary dataset) is equivalent to bootstrapping the corresponding individual-level data. Briefly, for $b = 1, \ldots, B$, we generate perturbed conditional estimates $\hat{\boldsymbol{\beta}}_X^{(b)} \sim \mathcal{MVN}(\hat{\boldsymbol{\beta}}_X, \boldsymbol{\Sigma}_X)$ and $\hat{\boldsymbol{\beta}}_Y^{(b)} \sim \mathcal{MVN}(\hat{\boldsymbol{\beta}}_Y, \boldsymbol{\Sigma}_Y)$, and apply the estimation procedure described above on the perturbed data to obtain $\hat{\theta}^{(b)}$. And we use the sample mean and sample standard deviation of $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}$ as the final causal estimate and its corresponding standard error.

4.2.3.2 Modeling conditional effects versus marginal effects

A possible and seemingly effective alternative is to model marginal effects, instead of modeling joint effects, of SNPs (Eq (4.9)). That is, we have

$$\begin{aligned} \hat{\boldsymbol{\beta}}_X^* &\sim \mathcal{MVN}(\mathbf{b}_X^* = \mathbf{R}\mathbf{b}_X, \mathbf{R} \cdot \boldsymbol{\sigma}_X^* \boldsymbol{\sigma}_X^{*T}), \\ \hat{\boldsymbol{\beta}}_Y^* &\sim \mathcal{MVN}(\mathbf{b}_Y^* = \mathbf{R}\mathbf{b}_Y = \theta \mathbf{b}_X^* + \mathbf{r}^*, \mathbf{R} \cdot \boldsymbol{\sigma}_Y^* \boldsymbol{\sigma}_Y^{*T}), \end{aligned} \tag{4.12}$$

where $\mathbf{r}^* = \mathbf{R}\mathbf{r}$. We can also have a similar relationship $b_{Yi}^* = \theta b_{Xi}^* + r_i^*$ as in Eq (4.4). However, one pitfall is that, IVs without horizontal pleiotropy in the conditional

model (i.e. with $r_i = 0$) may have $r_i^* \neq 0$ in the marginal model. For example, let $\mathbf{r} = (r_1, 0, \ldots, 0)^T$ and $r_1 \neq 0$, then $\mathbf{r}^* = \mathbf{R}\mathbf{r}$ will have non-zero elements for all m SNPs when they are all correlated with the first SNP (i.e., the first column of $\mathbf{R}$ are all non-zeros). Hence, although the plurality condition holds in the conditional model, it is violated in the marginal model. In general, the plurality condition is more likely to hold in the conditional model than in the marginal model. Therefore, in cisMR-cML, we use the joint/conditional effect estimates instead of the marginal effect estimates.

4.2.3.3 Selection of genetic variants as IVs in cisMR-cML

In this section, we discuss which SNPs should be used in the proposed method and how to select them. First, it is crucial to include **all** m SNPs associated with either the exposure or outcome, i.e. those in $\mathcal{I}_X \cup \mathcal{I}_Y$ in Figure 4.1, **not** any of their proper subsets, and calculate their joint estimates with the exposure and the outcome. This is in striking contrast with that in the independent IV case, where only SNPs significantly associated with the exposure (i.e., SNPs in $\mathcal{I}_X$) are used (Xue et al., 2021). This is because, as shown in Figure 4.1, conditional on G_k, G_i in $\mathcal{I}_X \setminus \mathcal{I}_Y$ does not have a direct path to the outcome; but if we do not include SNPs in $\mathcal{I}_Y$ (e.g. G_k), then it will open alternative paths of all other correlated SNPs (with G_k) to the outcome not through the exposure. This will in turn break the plurality condition required by model identifiability since all SNPs will have direct effects on Y. On the other hand, such an issue is unlikely to occur when SNPs are all independent. We also note that, when we include SNPs in $\mathcal{I}_Y \setminus \mathcal{I}_X$, cisMR-cML is expected to select them out as invalid IVs.

In practice, to select these m SNPs, we apply the COJO method (Yang et al., 2012) on the exposure and the outcome respectively to select SNPs in $\mathcal{I}_X$ and $\mathcal{I}_Y$. COJO is suitable in our application since it can identify SNPs that *jointly* are significantly

associated with the phenotype via a stepwise selection procedure. It is applicable to both quantitative traits and case-control studies. Furthermore, it only uses GWAS summary statistics and an estimated LD matrix from a reference panel as in cisMR-cML.

4.2.3.4 Theory

The proposed cisMR-cML enjoys nice asymptotic properties, including selection consistency of the proposed BIC and asymptotic normality of the cMLE. Here we state the assumptions and main conclusions with the proofs relegated to the Supplementary.

Assumption 4.1

(Plurality valid condition.) Suppose that $\mathcal{A}_0 = \{i : r_i \neq 0\}$ is the index set of the true invalid IVs with a non-zero horizontal-pleiotropy effect, and $K_0 = |\mathcal{A}_0|$. For any $\mathcal{A} \subseteq \{1, \ldots, m\}$ and $|\mathcal{A}| = K_0$, if $\mathcal{A} \neq \mathcal{A}_0$, then there does not exist any constant $\tilde{\theta} \neq \theta$ such that $b_{Yi} = \tilde{\theta} b_{Xi}$ for all $i \in \mathcal{A}^C$.

Assumption 4.2

The joint effect estimates $\hat{\boldsymbol{\beta}}_X \sim \mathcal{MVN}(\mathbf{b}_X, \boldsymbol{\Sigma}_X)$ and $\hat{\boldsymbol{\beta}}_Y \sim \mathcal{MVN}(\mathbf{b}_Y, \boldsymbol{\Sigma}_Y)$ with the known covariance matrices $\boldsymbol{\Sigma}_X$ and $\boldsymbol{\Sigma}_Y$.

Assumption 4.3

(Orders of the variances and sample sizes.) Let $N = \min(N_X, N_Y)$, there exist positive constants c_1, c_2 such that $c_1/N \leq (\boldsymbol{\Sigma}_X)_{ij} \leq c_2/N$ and $c_1/N \leq (\boldsymbol{\Sigma}_Y)_{ij} \leq c_2/N$ for $i = 1, \ldots, m$, $j = 1, \ldots, m$, i.e., $\boldsymbol{\Sigma}_X$ and $\boldsymbol{\Sigma}_Y$ are $\Theta(1/N)$.

Assumption 4.1 is the plurality condition, which is equivalent to that in Theorem 1 of Guo et al. (2018), a sufficient and necessary condition for the identifiability of model Eq (4.4). Assumption 4.2 and Assumption 4.3 are reasonable given that

GWAS summary data are usually based on large sample sizes. Then the following theorem gives the selection consistency and asymptotic normality and consistency of the proposed estimator.

Theorem 4.1

With Assumption 4.1 to Assumption 4.3 satisfied, if $K_0 \in \mathcal{K}$, we have $P(\hat{K} = K_0) \to 1$ and $P(\hat{\mathcal{A}}_{\hat{K}} = \mathcal{A}_0) \to 1$ as $N \to \infty$. And the proposed constrained maximum likelihood estimator $\hat{\theta}$, combined with the use of the BIC selection criterion, is consistent for the true causal effect size θ_0, and

$$\sqrt{V}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, 1) \text{ as } N \to \infty,$$

where V is the expected Fisher information for the profile log-likelihood with all IVs in $\mathcal{A}_0^C$ that can be consistently estimated by its sample version.

We note that, as implied by the constraint we use in Eq (4.11), the invalid IVs in the proposed method are referred to as those in $\mathcal{I}_Y$ with a non-zero direct effect on the outcome ($r_i \neq 0$), which can be consistently selected out by the proposed BIC. On the other hand, although an irrelevant IV with $b_{Xi} = r_i = b_{Yi} = 0$ is also considered invalid, cisMR-cML will not select it out but including such an IV will not affect the validity of our inference as long as the conditions of Theorem 4.1 are satisfied. In summary, cisMR-cML is highly robust in the sense of allowing the presence of some invalid IVs violating any of the three valid IV assumptions; these invalid IVs can be more than a half of all the IVs used.

4.3 Simulations

4.3.1 Set-ups

We simulated the GWAS summary statistics largely following the simulation procedure used in the LDA-Egger paper (Barfield et al., 2018):

1. Generated the true joint effect of $|\mathcal{I}_X|$ SNPs on the exposure $b_{Xi} \sim \mathcal{N}(0,1)$, for $i \in \mathcal{I}_X$, and $b_{Xi} = 0$ for $i \notin \mathcal{I}_X$; rescaled the effects according to the proportion of variability in exposure due to SNPs: $\mathbf{b}_X = \sqrt{h_X^2/(\mathbf{b}_X^T\mathbf{R}\mathbf{b}_X)}\mathbf{b}_X$, where $h_X^2 = 0.05$, and $\mathbf{R}$ was the LD matrix generated from an autoregressive model with $\Sigma_{ij} = \rho^{|i-j|}$;

2. Generated the direct effects of $|\mathcal{I}_Y|$ SNPs on the outcome $r_i \sim \mathcal{N}(0,1)$ iid, with K_1 SNPs from $\mathcal{I}_X \cap \mathcal{I}_Y$ and K_2 SNPs from $\mathcal{I}_Y \setminus \mathcal{I}_X$; rescaled the direct effects according to the proportion of variability in outcome due directly to SNPs: $\mathbf{r} = \sqrt{h_Y^2/(\mathbf{r}^T\mathbf{R}\mathbf{r})}\mathbf{r}$, where $h_Y^2 = 0.05$;

3. Generated the true joint effects of SNPs on the outcome $\mathbf{b}_Y = \theta\mathbf{b}_X + \mathbf{r}$ with $\theta = 0$ or 0.2;

4. Generated the observed exposure GWAS estimates $\hat{\boldsymbol{\beta}}_X^* \sim \mathbf{R}\mathbf{b}_X + \mathbf{L}^T\boldsymbol{\epsilon}_X$, $\boldsymbol{\epsilon}_X \sim \mathcal{N}(\mathbf{0}, \frac{1-h_X^2}{N_X}\mathbf{I}_m)$, where $\mathbf{L}$ was the Cholesky decomposition of the LD matrix $\mathbf{R}$, and $N_X = 10000$. Note $\boldsymbol{\sigma}_X^* = \sqrt{\frac{1-h_X^2}{N_X}}\mathbf{1}_m$;

5. Generated the observed outcome GWAS estimates $\hat{\boldsymbol{\beta}}_Y^* \sim \mathbf{R}\mathbf{b}_Y + \mathbf{L}^T\boldsymbol{\epsilon}_Y$, $\boldsymbol{\epsilon}_Y \sim \mathcal{N}(\mathbf{0}, \frac{1-\theta^2h_X^2-h_Y^2}{N_Y}\mathbf{I}_m)$, and $N_Y = 50000$. Note $\boldsymbol{\sigma}_Y^* = \sqrt{\frac{1-\theta^2h_X^2-h_Y^2}{N_Y}}\mathbf{1}_m$.

In total $m = |\mathcal{I}_X \cup \mathcal{I}_Y| = 10$ SNPs were generated. We consider two scenarios: (1) $|\mathcal{I}_X| = 10$; (2) $|\mathcal{I}_X| = 5$, and $K_2 = |\mathcal{I}_Y \setminus \mathcal{I}_X| = 5$. We note that in the second scenario, only 5 SNPs had effects on the exposure, and we will investigate the impact

of not including the other 5 SNPs in $|\mathcal{I}_Y \setminus \mathcal{I}_X|$ in the analysis. In both scenarios, we varied K_1, the number of invalid IVs in $|\mathcal{I}_X \cap \mathcal{I}_Y|$.

Given the simulated GWAS summary statistics $(\hat{\boldsymbol{\beta}}_X^*, \boldsymbol{\sigma}_X^*, \hat{\boldsymbol{\beta}}_Y^*, \boldsymbol{\sigma}_Y^*)$, we transformed them to the conditional estimates by $\hat{\boldsymbol{\beta}}_X = \mathbf{R}^{-1}\hat{\boldsymbol{\beta}}_X^*$, $\hat{\boldsymbol{\beta}}_Y = \mathbf{R}^{-1}\hat{\boldsymbol{\beta}}_Y^*$ and $\boldsymbol{\Sigma}_X = \mathbf{R}^{-1}(\mathbf{R} \cdot \boldsymbol{\sigma}_X^* \boldsymbol{\sigma}_X^{*T})\mathbf{R}^{-1}$, $\boldsymbol{\Sigma}_Y = \mathbf{R}^{-1}(\mathbf{R} \cdot \boldsymbol{\sigma}_Y^* \boldsymbol{\sigma}_Y^{*T})\mathbf{R}^{-1}$. We applied cisMR-cML and LEgger with the conditional estimates calculated based on all 10 SNPs. And we applied GIVW and GEgger with their marginal GWAS estimates.

In scenario (2), we furthermore applied cisMR-cML and LEgger with the conditional estimates *calculated only* based on the 5 SNPs in $\mathcal{I}_X$, which was different from only using the corresponding 5 elements in $\hat{\boldsymbol{\beta}}_X$ and $\hat{\boldsymbol{\beta}}_Y$ calculated based on all 10 SNPs. And we also applied GIVW and GEgger with the GWAS summary data of the 5 SNPs in $\mathcal{I}_X$. We referred these implementations using only SNPs in $\mathcal{I}_X$ as cisMR-cML-X, LEgger-X, GIVW-X and GEgger-X respectively.

Each simulation setup was repeated 500 times. Throughout the simulation, cisMR-cML was implemented with the default starting point, and $B = 100$ data perturbations. GIVW and GEgger were implemented with their default settings in the R package `MendelianRandomization`. LEgger was implemented in the R code provided at https://rbarfield.github.io/Barfield_website/pages/Rcode.html.

4.3.2 Results

In the first scenario, where all 10 IVs had effects on the exposure ($|\mathcal{I}_X| = 10$), Table 4.1 shows the empirical type-I error, power, average of causal effect estimates and root mean squared error (RMSE) from 500 simulation replicates. First, when all 10 IVs were valid (4.1a), all methods yielded well-control type-I error rates. But we note that, even in this ideal scenario with no invalid IV, GEgger had relatively larger RMSE (and less precise estimates) than the other three methods, which may be due to the allele orientation step implemented in the method (see Lin et al. (2022) for

more discussion on this issue). In the presence of 4 invalid IVs (Table 4.1b), only cisMR-cML could control the type-I error and at the same time maintained high power. Furthermore, it had a much lower RMSE than the other three methods. On the other hand, GIVW, GEgger and LEgger had increasingly inflated type-I errors as the correlations among SNPs increased.

Table 4.1: Simulation results in scenario 1, where all 10 IVs have an effect on the exposure. Left: $K_1 = 0$ (no invalid IV). Right: $K_1 = 4$ invalid IVs. In each cell, from top to bottom are empirical type-I error/power, mean($\hat{\theta}$), RMSE.

(a) $K_1 = 0$

ρ	θ	cisMR-cML	GIVW	GEgger	LEgger
0.2	0.0	0.016	0.038	0.034	0.056
		-0.001	0.000	-0.001	-0.001
		0.020	0.019	0.035	0.022
	0.2	0.998	1.000	0.992	1.000
		0.200	0.197	0.197	0.197
		0.023	0.021	0.038	0.024
0.6	0.0	0.010	0.038	0.040	0.052
		-0.001	0.000	-0.001	0.000
		0.020	0.019	0.030	0.025
	0.2	0.998	1.000	0.996	0.994
		0.200	0.197	0.195	0.196
		0.023	0.021	0.033	0.028

(b) $K_1 = 4$

ρ	θ	cisMR-cML	GIVW	GEgger	LEgger
0.2	0.0	0.032	0.102	0.074	0.058
		0.002	0.003	-0.005	0.004
		0.051	0.337	0.558	0.344
	0.2	0.904	0.138	0.080	0.082
		0.205	0.200	0.194	0.202
		0.063	0.337	0.559	0.344
0.6	0.0	0.042	0.214	0.108	0.110
		0.001	0.000	-0.003	0.007
		0.075	0.410	0.478	0.424
	0.2	0.788	0.220	0.128	0.134
		0.199	0.197	0.194	0.203
		0.087	0.411	0.477	0.423

In the second scenario with only 5 SNPs having effects on the exposure, we additionally applied the four methods using only the data of these 5 SNPs, as shown in columns with suffix '-X' in Table 4.2. When $K_1 = 0$ (Table 4.2a), it seemed that all IVs in $\mathcal{I}_X$ were valid. However, due to their correlations with those in $\mathcal{I}_Y$, they absorbed the direct effects of the SNPs in $\mathcal{I}_Y$ on the outcome if we failed to include the SNPs in $\mathcal{I}_Y$. Therefore, all the IVs became invalid and the plurality condition was violated in cisMR-cML-X, which yielded highly inflated type-I error. Similarly, GIVW-X and GEgger-X, only using SNPs conditionally associated with the exposure also yielded inflated type-I errors. On the other hand, cisMR-cML using all 10 SNPs yielded well-control type-I errors, high power and smallest RMSE across all scenarios. Through this example, we can see the importance of including the SNPs in $\mathcal{I}_Y$ besides

those in $\mathcal{I}_X$ when calculating the conditional estimates, because otherwise the plurality condition required by cisMR-cML (or more generally by model identification) may be violated (unless there was no or little LD between the SNPs in $\mathcal{I}_X$ and $\mathcal{I}_Y$).

Table 4.2: Simulation results in scenario 2 with $|\mathcal{I}_X| = |\mathcal{I}_Y \backslash \mathcal{I}_X| = 5$. Top: $|\mathcal{I}_X \cap \mathcal{I}_Y| = 0$. Bottom: $|\mathcal{I}_X \cap \mathcal{I}_Y| = 1$. In each cell, from top to bottom are empirical type-I error/power, mean($\hat{\theta}$), RMSE.

(a) $K_1 = 0$

ρ	θ	cisMR-cML	GIVW	GEgger	LEgger	cisMR-cML-X	GIVW-X	GEgger-X	LEgger-X
0.2	0.0	0.028	0.000	0.014	0.002	0.216	0.116	0.154	0.044
		0.000	0.006	-0.009	-0.020	0.007	0.008	0.011	-0.003
		0.022	0.101	0.279	0.169	0.109	0.091	0.195	0.111
	0.2	0.998	0.000	0.032	0.002	0.716	0.614	0.306	0.288
		0.202	0.204	0.194	0.177	0.218	0.208	0.210	0.196
		0.026	0.101	0.279	0.170	0.116	0.092	0.195	0.111
0.6	0.0	0.016	0.058	0.022	0.014	0.334	0.174	0.156	0.068
		-0.001	0.033	-0.008	-0.021	0.043	0.035	-0.023	-0.015
		0.033	0.294	0.327	0.332	0.350	0.294	0.544	0.365
	0.2	0.972	0.110	0.024	0.046	0.572	0.272	0.174	0.106
		0.201	0.230	0.189	0.175	0.261	0.235	0.175	0.183
		0.036	0.294	0.328	0.333	0.351	0.294	0.544	0.365

(b) $K_1 = 1$

ρ	θ	cisMR-cML	GIVW	GEgger	LEgger	cisMR-cML-X	GIVW-X	GEgger-X	LEgger-X
0.2	0.0	0.032	0.012	0.024	0.010	0.214	0.122	0.154	0.058
		0.000	0.011	0.004	-0.011	0.005	0.011	-0.007	-0.004
		0.042	0.200	0.304	0.228	0.158	0.199	0.429	0.237
	0.2	0.930	0.034	0.044	0.016	0.698	0.330	0.216	0.140
		0.202	0.208	0.206	0.185	0.220	0.210	0.191	0.194
		0.049	0.200	0.304	0.229	0.141	0.199	0.431	0.238
0.6	0.0	0.020	0.096	0.030	0.030	0.370	0.192	0.150	0.062
		0.004	0.026	-0.020	-0.019	0.029	0.026	-0.044	-0.021
		0.125	0.330	0.355	0.332	0.401	0.333	0.597	0.392
	0.2	0.840	0.166	0.050	0.050	0.536	0.244	0.174	0.112
		0.206	0.223	0.177	0.177	0.263	0.225	0.155	0.176
		0.119	0.330	0.357	0.333	0.388	0.333	0.594	0.394

4.4 Real data applications

4.4.1 Reference panel

In the following real data applications, we used the UK Biobank individual-level genotype data (Bycroft et al., 2018) as the reference panel. As the following analysis was based on GWAS datasets of (mostly) European ancestry, 337426 unrelated (field '22020'=1) and self-reported White-British individuals with similar genetic ancestry (field '22006'=1) in UK Biobank were used to calculate the LD matrix among SNPs.

4.4.2 Causal effects of downstream biomarkers on CAD

4.4.2.1 Data sets and methods

In this application, we first applied *cis*-MR in a setup where we used a downstream biomarker as a proxy of protein concentration and activity. Specifically, we assessed the causal relationship of low-density lipoprotein cholesterol (LDL) on coronary artery disease (CAD) using the genetic variants restricted to the *PCSK9* region. Following Gkatzionis et al. (2023), we also assessed the causal relationship of testosterone level on CAD using the genetic variants in the *SHBG* region. The causal effect of LDL on CAD has been extensively studied by randomized trials and MR (Baigent et al., 2010). In particular, Ference et al. (2016) found a protective effect on CAD of lowering LDL using a weighted *PCSK9* genetic score to mimic the effect of PCSK9 inhibitor. On the other hand, while an association between low testosterone level and CAD risk has been reported in some observational studies, its causal relationship is still unclear. Their relationship has also been studied previously using genetic variants in the *SHBG* region (Burgess et al., 2017; Schooling et al., 2018; Gkatzionis et al., 2023). Our analysis here mainly aims to illustrate how to apply cisMR-cML using a downstream biomarker of the target protein to confirm/replicate some well-established results.

GWAS summary data for both LDL cholesterol and testosterone were taken from the Neale Lab UK Biobank GWAS round 2 results (http://www.nealelab.is/uk-biobank/). And the GWAS summary data for CAD was obtained from the CARDIoGRAMplusC4D Consortium (2015). We first extracted genetic variants located 500 kb on both sides of a gene, and retained those present in both the biomarker and CAD GWAS data, and confined our analysis to variants with missing genotypes $< 10\%$, minor allele frequency (MAF) ¿ 0.01, Hardy–Weinberg equilibrium (HWE) $p > 1 \times 10^{-6}$ in the reference panel. Then we performed GCTA-COJO on the exposure (or outcome) GWAS data to select SNPs jointly associated with the exposure (or the outcome) at $p < 5 \times 10^{-6}$, denoted as $\mathcal{I}_X$ (or $\mathcal{I}_Y$) respectively.

We transformed the marginal association estimates $(\hat{\boldsymbol{\beta}}_X^*, \hat{\boldsymbol{\beta}}_Y^*)$ to the conditional estimates $(\hat{\boldsymbol{\beta}}_X, \hat{\boldsymbol{\beta}}_Y)$ for the SNPs in set $\mathcal{I}_X \cup \mathcal{I}_Y$, and calculated the corresponding covariance matrices $\boldsymbol{\Sigma}_X$ and $\boldsymbol{\Sigma}_Y$ according to Section 4.2.1. Then we applied cisMR-cML with $B = 500$ data perturbations with the default starting point, and LDA-Egger. We also applied GIVW-X and GEgger-X only using the marginal association estimates of SNPs in $\mathcal{I}_X$.

4.4.2.2 Results

PCSK9 can bind to and break down LDL receptors, therefore decreasing the clearance of LDL cholesterol. And PCSK9 inhibitors are a new type of drugs that can lower LDL levels by blocking PCSK9 proteins from breaking down LDL receptors. In the analysis of LDL and CAD, GCTA-COJO selected 9 SNPs located in the *PCSK9* region associated with LDL, and 1 SNP associated with CAD, i.e., $|\mathcal{I}_X| = 8$ and $|\mathcal{I}_Y| = 1$. In our *cis*-MR analysis, both cisMR-cML, LEgger, GIVW-X and GEgger-X suggested a significant positive causal effect of LDL on CAD risk, with p-values $1.0 \times 10^{-5}, 9.2 \times 10^{-3}, 8.6 \times 10^{-8}, 0.02$ respectively.

Sex hormone-binding globulin (*SHBG*) can bind to sex hormones in the blood and

help control the amount of sex hormones. Multiple variants in this region have been demonstrated to be associated with testosterone. In the analysis of testosterone level and CAD, GCTA-COJO selected 14 SNPs associated with testosterone, and no SNP associated with CAD. Using the 14 variants in the *SHBG* region, no method identified any significant causal effect of testosterone on CAD risk, which was consistent with previous findings in Burgess et al. (2017); Schooling et al. (2018); Gkatzionis et al. (2023).

4.4.3 Proteome-wide analysis for CAD risk

4.4.3.1 Data sets and methods

In this application, we used protein expression data as the exposure, which was a more direct proxy of drug target, and we assessed their causal effects on the risk of CAD. Specifically, we did a proteome-wide scan using the pQTL summary data derived from ARIC European ancestry (EA) cohort with sample size $N_X = 7213$ (Zhang et al., 2022). We confined our analysis to a list of 1034 proteins with ≥ 3 identified pQTLs in the EA population according to Supplementary Table 6.1 in Zhang et al. (2022). For CAD GWAS, we used the one with a larger sample size of $N_Y = 547261$, which was a meta-analysis result of UK Biobank and CARDIoGRAMplusC4D (van der Harst and Verweij, 2018).

The data preprocessing step was similar to that in Section 4.4.2.1, except that the LDL (or testosterone) GWAS data (i.e., exposure GWAS) was replaced by the pQTL dataset. Again, we ran GCTA-COJO on both the pQTL data and CAD GWAS data with the UK Biobank data as the reference panel to obtain $\mathcal{I}_X$ and $\mathcal{I}_Y$ We retained proteins with ≥ 3 SNPs in $\mathcal{I}_X$, and excluded proteins with highly correlated (using "–cojo-collinear 0.9") SNPs in $\mathcal{I}_X$ and $\mathcal{I}_Y$. After the preprocessing step, 901 proteins remained to be analyzed next. We applied cisMR-cML and LEgger with the

conditional estimates $(\hat{\boldsymbol{\beta}}_X, \hat{\boldsymbol{\beta}}_Y, \boldsymbol{\Sigma}_X, \boldsymbol{\Sigma}_Y)$ calculated on the SNP-set $\mathcal{I}_X \cup \mathcal{I}_Y$, as well as GIVW-X and GEgger-X on the pQTLs that were conditionally associated with the proteins (i.e., those in $\mathcal{I}_X$) (Zheng et al., 2022; Zhao et al., 2022). We also applied the Wald-ratio test using the pQTL with the smallest marginal p-value for each protein.

We further conducted colocalization analysis on the significant proteins with an FDR-adjusted p-value less than 0.05 (using *p.adjust(method='fdr')* in R). Colocalization analysis has been more regularly used and strongly recommended in practice following MR analysis (Zuber et al., 2022). In particular, we used a Bayesian colocalization method called COLOC (Giambartolomei et al., 2014), where a high H4-PP suggested the protein and CAD shared the same causal variant at the locus, while a high H3-PP suggested the protein and CAD had different causal variants at the locus. The former case supported the significant result from MR, however, the latter case suggested the significant MR result may be driven by genetic confounding through LD between pQTLs and CAD-associated SNPs, e.g. SNPs in $\mathcal{I}_Y \setminus \mathcal{I}_X$. COLOC was implemented with *coloc.abf()* in the R package `coloc` with the default setting.

4.4.3.2 Results

We used the Benjamini-Hochberg approach to account for multiple testing in our proteome-wide analysis, and reported significant MR findings with a false discovery rate (FDR) less than 0.05. cisMR-cML identified three proteins with putative causal effect on CAD risk, including *PCSK9*, *COLEC11* and *FGFR1*. Using a threshold of H4-PP ≥ 0.7, there were colocalization evidence for both *PCSK9* and *COLEC11*. As discussed in the previous application, PCSK9 inhibitors can lower LDL level, which is a major risk factor for CAD. Several trials found that evolocumab, a PCSK9 inhibitor, can significantly lower LDL level and cardiovascular disease risk (Robinson et al., 2015; Sabatine et al., 2017; Gaba et al., 2023). *COLEC11* is involved in lectin complement activation pathway and plays an important role in the innate

immune system. The vital role of the complement system in heart diseases has been studied, including promoting inflammation, tissue damage, etc. (Lappegård et al., 2014; Shahini et al., 2017). While complement inhibitors have been suggested as a potential therapeutic target for heart disease, more studies on the relationship between *COLEC11* and CAD are warranted. As for *FGFR1*, colocalization only identified the causal variant for the protein with H1-PP $\approx 96\%$. This was the scenario with insufficient evidence for association with CAD in the CAD GWAS data (Zuber et al., 2022). FGF/FGFR signaling plays an important role in cell proliferation and angiogenesis, and several FGFR1 inhibitors have been used to treat various types of cancer (Katoh, 2016). While overexpression of *FGFR1* may play a role in the development of cardiac hypertrophy (Faul et al., 2011; Freundlich et al., 2014), it is also likely that FGFR expression pattern is altered in response to cardiac stress and injury and facilitate cardiac remolding (Khosravi et al., 2021; Faul, 2017). Further studies are needed to fully understand their complex relationship.

On the other hand, GIVW-X identified 23 proteins, and five of them had colocalization evidence including *BMP1, FN1, COLEC11, PCSK9, ERAP2*. However, there were six proteins with an H3-PP greater than 0.7, suggesting that the protein and CAD have distinct causal variants that were in linkage disequilibrium, and thus the MR assumption may be violated. Similarly, GEgger-X identified eight proteins, 3 of which had colocalization evidence including *HSPC159, PCSK9, TIRAP*, and two had an H3-PP greater than 0.7. We note that this could be the scenario we've seen in our simulation scenario 2, where only using SNPs conditionally associated with the exposure yielded inflate type-I error in GIVW-X and GEgger-X. Such significant MR findings may be attributable to genetic confounding through a variant in linkage disequilibrium as suggested by a high H3-PP. Wald-ratio test using the most significant pQTL identified 40 proteins, 6 of which had colocalization evidence including *HSPC159, PCSK9, TIRAP, FN1, COLEC11, ERAP2*, but 15 of them had evidence

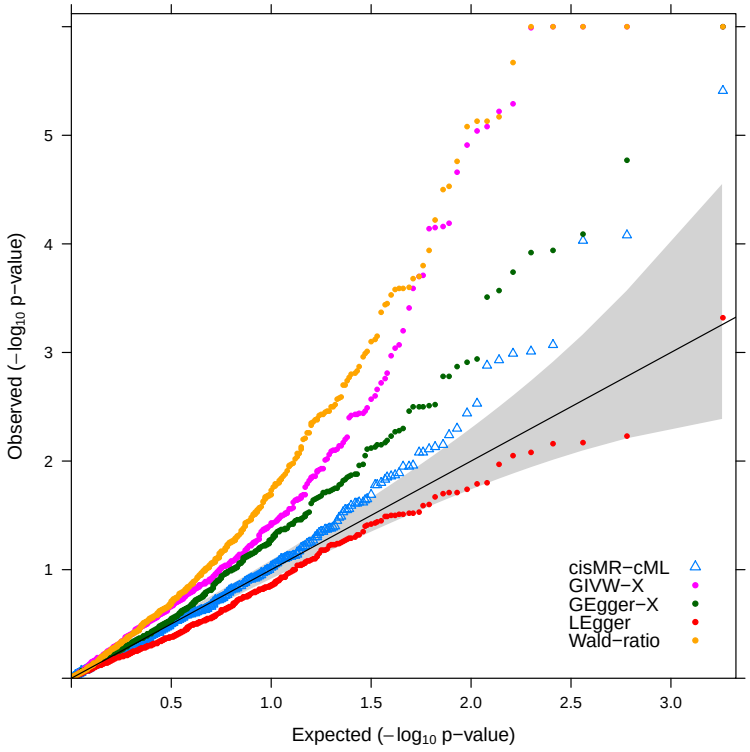


Figure 4.2: Q-Q plots of $-\log_{10}$ p-value for different methods. P-values are truncated at 1×10^{-6}.

of H3-PP greater than 0.7. And lastly, LEgger didn't identify any significant results. We show the Q-Q plots of all methods in Figure 4.2, in which we can see that in the left tail, only cisMR-cML had good alignment with the identity line, while GIVW-X, GEgger-X and Wald-ratio were inflated, and LEgger was deflated. The inflation factor for cisMR-cML was 1.01 (rounded to the second decimal), suggesting that the Type-I error was controlled satisfactorily; while LEgger, GEgger-X, GIVW-X and Wald-ratio test yielded inflation factors 0.64, 1.10, 1.61 and 1.57 respectively.

4.5 Discussion

We have proposed a robust *cis*-MR method called cisMR-cML, which uses correlated SNPs in a genomic region to infer the causal relationship between an exposure (e.g. a protein) and an outcome (e.g. CAD), and is robust to the presence of invalid IVs. It is an important extension of the existing MR-cML method, which has been shown to have good performance in practice but requires the use of independent SNPs (Xue et al., 2021). While such an extension may seem straightforward at the first glance by incorporating LD information in the likelihood, we have pointed out several important implementation details with significant implications to final results. To prevent inducing pleiotropy via LD correlations in the model, we modeled conditional estimates, instead of marginal estimates, by suitably transforming GWAS summary data in cisMR-cML. We also discussed and demonstrated the importance of using SNPs associated with the outcome, in addition to those associated with the exposure, in cisMR-cML, which is stark contrast to the common practice of only using SNPs associated with the exposure in MR, e.g. MR-cML. These caveats would also be applicable to future extensions of other existing robust MR methods only requiring the majority or plurality condition, such as weighted-median and mode-based methods. Furthermore, we have proposed a new algorithm, rather than the standard coordinate-descent algorithm in MR-cML, to solve the constrained maximum likelihood problem in cisMR-cML, which can better account for LD correlations among SNPs in the step of invalid IVs selection. While we have mainly focused on the application of *cis*-MR using one genomic region in this paper, our method can be generalized to multiple independent LD blocks and serve as a useful MR method accounting for both LD and horizontal pleiotropy.

In our simulation studies, we have showcased the better performance of the proposed cisMR-cML over several commonly used *cis*-MR methods including generalized

IVW and Egger, and LDA-Egger. We have also compared different choices of the set of SNPs used in different methods. In particular, we have found that applying generalized IVW and Egger with SNPs conditionally associated with the exposure may yield false positive findings, partly due to that some outcome-associated SNPs not included in the model are in LD with the SNPs in the model, thus leading to the violation of no-pleiotropy assumption. This was confirmed in both simulation studies and real data applications, and we hope to raise the attention to this largely neglected issue in future applications.

There are several limitations in the proposed method. First, while cisMR-cML imposes minimum modeling assumptions, especially no additional assumption on the distribution of pleiotropic effect, it depends critically on the plurality condition, which depends on which SNPs are used in the model due to the correlations among all the SNPs, either selected or not, in the region. We currently use GCTA-COJO to select and include the SNPs that are associated with either the exposure or the outcome. However, COJO is by no means the only method for conditional analysis. Furthermore, SNPs selection in *cis*-MR analysis is an ongoing research topic (see Gkatzionis et al. (2023) for a detailed review, and Schmidt et al. (2020) for another example) and there seems no consensus yet. How to incorporate other robust SNP selection techniques or develop new ones in cisMR-cML is of interest for future work. Second, since individual-level genotypes in the exposure and outcome GWAS data are often unavailable, as in most other applications, we propose using a reference panel of the similar ancestry to approximate an LD matrix. Such an approximation is known to introduce extra variation that is not taken account in our method (and almost all methods). Using a larger reference panel, such as the UK Biobank samples as used in our analysis, is expected to alleviate the problem (Xue et al., 2023). Third, we have considered only the two-sample case with two independent GWAS datasets for the exposure and outcome. To account for overlapping samples between the two GWAS

dataset, we may model the exposure estimates and the outcome estimates jointly with a multivariate normal distribution, instead of treating them as independent. Fourth, the proteome-wide application presented in this work is based on the protein levels measured in plasma samples, while using disease-relevant tissue samples may be preferred in drug-target MR. Since large-scale tissue-specific pQTL data are not accessible thus far, one alternative is to use tissue-specific eQTL data as a proxy (Schmidt et al., 2020). Finally and more importantly, triangulation with evidence from applying different *cis*-MR methods and colocalization analysis to observational data, and direct experimental studies when possible are warranted for more reliable causal inference.

Chapter 5

Conclusions

In this work, we have proposed several robust Mendelian randomization methods based on constrained maximum likelihood, which all enjoy nice statistical properties. We began by extending an existing MR-cML method to allow for overlapping samples, which can accommodate different GWAS data set-ups including two-sample, overlapping-two-sample and even one-sample set-ups, and at the same time, is robust to horizontal pleiotropy. Then we coupled it with a network deconvolution method to infer a direct causal network among multiple traits, which moved beyond the conventional use of MR to infer the total causal effect between a pair of traits. We applied the method to 17 large-scale GWAS summary datasets to infer the causal networks of both total and direct effects among 11 common risk factors, 6 diseases, confirming several well-established results and identifying some interesting causal pathways.

Next, with the similar aim of inferring the direct causal effects and addressing the issue of horizontal pleiotropy, we proposed a robust multivariable MR method, called MVMR-cML. There are only a few MVMR methods in the current literature, and we confirmed the superior performance of MVMR-cML over most of them in the extensive simulations. We applied the proposed method to infer causal relationships between 8 cardiometabolic risk factors and coronary artery disease (CAD), and found that after accounting for possible pleiotropic and mediating effects, triglyceride (TG), low-

density lipoprotein cholesterol (LDL), and systolic blood pressure (SBP) had direct effects on CAD. We also want to point out that, MVMR-cML and Graph-MRcML have their own pros and cons, and they should be considered as complementary to each other. For example, Graph-MRcML relies on the UVMR IV assumptions while MVMR-cML on the MVMR IV assumptions, of which the relevance assumption in the former is weaker while the no-pleiotropy assumption in the latter is weaker.

Lastly, we moved to *cis*-MR analysis, which uses correlated genetic variants in a single genomic region. We proposed cisMR-cML, which can account for linkage disequilibrium and horizontal pleiotropy among local cis-SNPs. We further clarified the significant but largely neglected consequences of modeling conditional effects versus marginal effects in *cis*-MR analysis. While more work is needed to investigate and improve the performance of the current computation algorithm, we demonstrated the robustness of our method (with the current implementation) than competing *cis*-MR methods in some preliminary simulations and real data applications.

Overall, this research has filled several gaps in the current literature. And we expect our proposed methods to be important additions to the toolbox of Mendelian randomization analysis.

References

Abdullah, S. M., Defina, L. F., Leonard, D., Barlow, C. E., Radford, N. B., Willis, B. L., Rohatgi, A., McGuire, D. K., de Lemos, J. A., Grundy, S. M., et al. (2018). Long-term association of low-density lipoprotein cholesterol with cardiovascular mortality in individuals at low 10-year risk of atherosclerotic cardiovascular disease: results from the cooper center longitudinal study. *Circulation*, 138(21):2315–2325.

Ades, P. A. and Savage, P. D. (2017). Obesity in coronary heart disease: An unaddressed behavioral risk factor. *Preventive medicine*, 104:117.

Alshehri, A. M. (2019). Stroke in atrial fibrillation: review of risk stratification and preventive therapy. *Journal of family & community medicine*, 26(2):92.

Andersen, K., Rasmussen, F., Neovius, M., Tynelius, P., and Sundström, J. (2018). Body size and risk of atrial fibrillation: a cohort study of 1.1 million young men. *Journal of Internal Medicine*, 283(4):346–355.

Arboix, A. (2015). Cardiovascular risk factors for acute stroke: Risk profiles in the different subtypes of ischemic stroke. *World Journal of Clinical Cases: WJCC*, 3(5):418.

Arega, Y. and Shao, Y. (2022). Heart failure and late-onset alzheimer's disease: A mendelian randomization study. *Frontiers in Genetics*, 13.

Arvanitis, M., Qi, G., Bhatt, D. L., Post, W. S., Chatterjee, N., Battle, A., and McEvoy, J. W. (2021). Linear and nonlinear mendelian randomization analyses of the association between diastolic blood pressure and cardiovascular events: the j-curve revisited. *Circulation*, 143(9):895–906.

Baigent, C., Blackwell, L., Emberson, J., Holland, L., Reith, C., Bhala, N., Peto, R., Barnes, E., Keech, A., Simes, J., et al. (2010). Efficacy and safety of more intensive lowering of ldl cholesterol: a meta-analysis of data from 170,000 participants in 26 randomised trials. *Lancet (London, England)*, 376(9753):1670–1681.

Barfield, R., Feng, H., Gusev, A., Wu, L., Zheng, W., Pasaniuc, B., and Kraft, P. (2018). Transcriptome-wide association studies accounting for colocalization using egger regression. *Genetic Epidemiology*, 42(5):418–433.

Beddhu, S., Chertow, G. M., Cheung, A. K., Cushman, W. C., Rahman, M., Greene, T., Wei, G., Campbell, R. C., Conroy, M., Freedman, B. I., et al. (2018). Influence of baseline diastolic blood pressure on effects of intensive compared with standard blood pressure control. *Circulation*, 137(2):134–143.

Beeri, M. S., Davidson, M., Silverman, J. M., Noy, S., Schmeidler, J., and Goldbourt, U. (2005). Relationship between body height and dementia. *The American journal of geriatric psychiatry*, 13(2):116–123.

Bentler, P. M. and Freeman, E. H. (1983). Tests for stability in linear structural equation systems. *Psychometrika*, 48(1):143–145.

Berisa, T. and Pickrell, J. K. (2016). Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, 32(2):283.

Boehm, F. J. and Zhou, X. (2022). Statistical methods for mendelian randomiza-

tion in genome-wide association studies: A review. *Computational and Structural Biotechnology Journal*.

Boos, D. D. and Stefanski, L. A. (2013). *Essential statistical inference: theory and methods*, volume 591. Springer.

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(2):512–525.

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(4):304–314.

Brower, M., Hai, Y., Jones, M., Guo, X., Chen, Y.-D., Rotter, J., Krauss, R., Legro, R., Azziz, R., and Goodarzi, M. (2019). Bidirectional mendelian randomization to explore the causal relationships between body mass index and polycystic ovary syndrome. *Human Reproduction*, 34(1):127–136.

Brown, B. C. and Knowles, D. A. (2020). Phenome-scale causal network discovery with bidirectional mediated mendelian randomization. *bioRxiv*.

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., Duncan, L., Perry, J. R., Patterson, N., Robinson, E. B., et al. (2015). An atlas of genetic correlations across human diseases and traits. *Nature genetics*, 47(11):1236–1241.

Burgess, S., Butterworth, A., and Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic epidemiology*, 37(7):658–665.

Burgess, S., Daniel, R. M., Butterworth, A. S., Thompson, S. G., and Consortium, E.-I. (2015). Network mendelian randomization: using genetic variants as instrumental variables to investigate mediation in causal pathways. *International journal of epidemiology*, 44(2):484–495.

Burgess, S., Davies, N. M., and Thompson, S. G. (2016a). Bias due to participant overlap in two-sample mendelian randomization. *Genetic epidemiology*, 40(7):597–608.

Burgess, S., Dudbridge, F., and Thompson, S. G. (2016b). Combining information on multiple instrumental variables in mendelian randomization: comparison of allele score and summarized data methods. *Statistics in medicine*, 35(11):1880–1906.

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):1–11.

Burgess, S. and Thompson, S. G. (2011). Bias in causal estimates from mendelian randomization studies with weak instruments. *Statistics in medicine*, 30(11):1312–1323.

Burgess, S. and Thompson, S. G. (2015). Multivariable mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *American journal of epidemiology*, 181(4):251–260.

Burgess, S. and Thompson, S. G. (2017). Interpreting findings from mendelian randomization using the mr-egger method. *European journal of epidemiology*, 32:377–389.

Burgess, S., Zuber, V., Valdes-Marquez, E., Sun, B. B., and Hopewell, J. C. (2017).

Mendelian randomization with fine-mapped genetic data: choosing from large numbers of correlated instrumental variables. *Genetic epidemiology*, 41(8):714–725.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209.

Carreras-Torres, R., Johansson, M., Haycock, P. C., Relton, C. L., Smith, G. D., Brennan, P., and Martin, R. M. (2018). Role of obesity in smoking behaviour: Mendelian randomisation study in uk biobank. *Bmj*, 361.

Carter, A. R., Sanderson, E., Hammerton, G., Richmond, R. C., Smith, G. D., Heron, J., Taylor, A. E., Davies, N. M., and Howe, L. D. (2021). Mendelian randomisation for mediation analysis: current methods and challenges for implementation. *European journal of epidemiology*, 36(5):465–478.

Chen, J., Spracklen, C. N., Marenne, G., Varshney, A., Corbin, L. J., Luan, J., Willems, S. M., Wu, Y., Zhang, X., Horikoshi, M., et al. (2021). The trans-ancestral genomic architecture of glycemic traits. *Nature genetics*, 53(6):840–860.

Cheng, Q., Qiu, T., Chai, X., Sun, B., Xia, Y., Shi, X., and Liu, J. (2022a). Mr-corr2: a two-sample mendelian randomization method that accounts for correlated horizontal pleiotropy using correlated instrumental variants. *Bioinformatics*, 38(2):303–310.

Cheng, Q., Yang, Y., Shi, X., Yeung, K.-F., Yang, C., Peng, H., and Liu, J. (2020). Mr-ldp: a two-sample mendelian randomization for gwas summary statistics accounting for linkage disequilibrium and horizontal pleiotropy. *NAR genomics and bioinformatics*, 2(2):lqaa028.

Cheng, Q., Zhang, X., Chen, L. S., and Liu, J. (2022b). Mendelian randomization accounting for complex correlated horizontal pleiotropy while elucidating shared genetic etiology. *Nature Communications*, 13(1):6490.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18.

Dai, J. Y., Peters, U., Wang, X., Kocarnik, J., Chang-Claude, J., Slattery, M. L., Chan, A., Lemire, M., Berndt, S. I., Casey, G., et al. (2018). Diagnostics for pleiotropy in mendelian randomization studies: global and individual tests for direct effects. *American journal of epidemiology*, 187(12):2672–2680.

Dare, S., Mackay, D. F., and Pell, J. P. (2015). Relationship between smoking and obesity: a cross-sectional study of 499,504 middle-aged adults in the uk general population. *PloS one*, 10(4):e0123579.

Davey Smith, G. and Ebrahim, S. (2003). 'mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International journal of epidemiology*, 32(1):1–22.

Davies, N. M., Hill, W. D., Anderson, E. L., Sanderson, E., Deary, I. J., and Smith, G. D. (2019). Multivariable two-sample mendelian randomization estimates of the effects of intelligence and education on health. *Elife*, 8:e43990.

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(1):42–53.

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(2):105–116.

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(10):1412–1425.

Faul, C. (2017). Cardiac actions of fibroblast growth factor 23. *Bone*, 100:69–79.

Faul, C., Amaral, A. P., Oskouei, B., Hu, M.-C., Sloan, A., Isakova, T., Gutiérrez, O. M., Aguillon-Prada, R., Lincoln, J., Hare, J. M., et al. (2011). Fgf23 induces left ventricular hypertrophy. *The Journal of clinical investigation*, 121(11).

Feizi, S., Marbach, D., Médard, M., and Kellis, M. (2013). Network deconvolution as a general method to distinguish direct dependencies in networks. *Nature biotechnology*, 31(8):726–733.

Ference, B. A., Robinson, J. G., Brook, R. D., Catapano, A. L., Chapman, M. J., Neff, D. R., Voros, S., Giugliano, R. P., Davey Smith, G., Fazio, S., et al. (2016). Variation in pcsk9 and hmgcr and risk of cardiovascular disease and diabetes. *New England Journal of Medicine*, 375(22):2144–2153.

Freundlich, M., Li, Y. C., Quiroz, Y., Bravo, Y., Seeherunvong, W., Faul, C., Weisinger, J. R., and Rodriguez-Iturbe, B. (2014). Paricalcitol downregulates myocardial renin–angiotensin and fibroblast growth factor expression and attenuates cardiac hypertrophy in uremic rats. *American journal of hypertension*, 27(5):720–726.

Frost, L., Hune, L. J., and Vestergaard, P. (2005). Overweight and obesity as risk

factors for atrial fibrillation or flutter: the danish diet, cancer, and health study. *The American journal of medicine*, 118(5):489–495.

Gaba, P., O'Donoghue, M. L., Park, J.-G., Wiviott, S. D., Atar, D., Kuder, J. F., Im, K., Murphy, S. A., De Ferrari, G. M., Gaciong, Z. A., et al. (2023). Association between achieved low-density lipoprotein cholesterol levels and long-term cardiovascular and safety outcomes: An analysis of fourier-ole. *Circulation*.

Gentle, J. E. (2007). Matrix algebra. *Springer texts in statistics, Springer, New York, NY, doi*, 10:978–0.

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS genetics*, 10(5):e1004383.

Gilbody, J., Borges, M. C., Smith, G. D., and Sanderson, E. (2022). Multivariable mr can mitigate bias in two-sample mr using covariable-adjusted summary associations. *medRxiv*.

Gkatzionis, A., Burgess, S., and Newcombe, P. J. (2023). Statistical methods for cis-mendelian randomization with two-sample summary-level data. *Genetic epidemiology*, 47(1):3–25.

Gleason, K. J., Yang, F., and Chen, L. S. (2021). A robust two-sample transcriptome-wide mendelian randomization method integrating gwas with multi-tissue eqtl summary statistics. *Genetic epidemiology*, 45(4):353–371.

Grant, A. J. and Burgess, S. (2021). Pleiotropy robust methods for multivariable mendelian randomization. *Statistics in medicine*, 40(26):5813–5830.

Guo, Z., Kang, H., Tony Cai, 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(4):793–815.

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(6):1985–1998.

Hartwig, F. P., Tilling, K., Davey Smith, G., Lawlor, D. A., and Borges, M. C. (2021). Bias in two-sample mendelian randomization when using heritable covariable-adjusted summary associations. *International journal of epidemiology*, 50(5):1639–1650.

Hemani, G., Bowden, J., and Davey Smith, G. (2018a). Evaluating the potential role of pleiotropy in mendelian randomization studies. *Human molecular genetics*, 27(R2):R195–R208.

Hemani, G., Tilling, K., and Davey Smith, G. (2017). Orienting the causal relationship between imprecisely measured traits using gwas summary data. *PLoS genetics*, 13(11):e1007081.

Hemani, G., Zheng, J., Elsworth, B., Wade, K., Baird, D., Haberland, V., Laurin, C., Burgess, S., Bowden, J., Langdon, R., Tan, V., Yarmolinsky, J., Shibab, H., Timpson, N., Evans, D., Relton, C., Martin, R., Davey Smith, G., Gaunt, T., Haycock, P., and The MR-Base Collaboration (2018b). The mr-base platform supports systematic causal inference across the human phenome. *eLife*, 7:e34408.

Holmes, M. V., Asselbergs, F. W., Palmer, T. M., Drenos, F., Lanktree, M. B., Nelson, C. P., Dale, C. E., Padmanabhan, S., Finan, C., Swerdlow, D. I., et al. (2015). Mendelian randomization of blood lipids for coronary heart disease. *European heart journal*, 36(9):539–550.

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(7624):248–252.

Hu, X., Zhao, J., Lin, Z., Wang, Y., Peng, H., Zhao, H., Wan, X., and Yang, C. (2022). Mendelian randomization for causal inference accounting for pleiotropy and sample structure using genome-wide summary statistics. *Proceedings of the National Academy of Sciences*, 119(28):e2106858119.

Jansen, I. E., Savage, J. E., Watanabe, K., Bryois, J., Williams, D. M., Steinberg, S., Sealock, J., Karlsson, I. K., Hägg, S., Athanasiu, L., et al. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing alzheimer's disease risk. *Nature genetics*, 51(3):404–413.

Jørgensen, T. S. H., Okholm, G. T., Christensen, K., Sørensen, T. I., and Osler, M. (2020). Body height in young adult men and risk of dementia later in adult life. *Elife*, 9:e51168.

Karpe, F., Dickmann, J. R., and Frayn, K. N. (2011). Fatty acids, obesity, and insulin resistance: time for a reevaluation. *Diabetes*, 60(10):2441–2449.

Katoh, M. (2016). Fgfr inhibitors: Effects on cancer cells, tumor microenvironment and whole-body homeostasis. *International journal of molecular medicine*, 38(1):3–15.

Khan, S. S., Ning, H., Wilkins, J. T., Allen, N., Carnethon, M., Berry, J. D., Sweis, R. N., and Lloyd-Jones, D. M. (2018). Association of body mass index with lifetime risk of cardiovascular disease and compression of morbidity. *JAMA cardiology*, 3(4):280–287.

Khosravi, F., Ahmadvand, N., Bellusci, S., and Sauer, H. (2021). The multifunctional contribution of fgf signaling to cardiac development, homeostasis, disease and repair. *Frontiers in cell and developmental biology*, 9:672935.

Kim, J., Bai, Y., and Pan, W. (2015). An adaptive association test for multiple phenotypes with gwas summary statistics. *Genetic epidemiology*, 39(8):651–663.

Kwak, I.-Y. and Pan, W. (2017). Gene-and pathway-based association tests for multiple traits with gwas summary statistics. *Bioinformatics*, 33(1):64–71.

Lappegård, K. T., Garred, P., Jonasson, L., Espevik, T., Aukrust, P., Yndestad, A., Mollnes, T. E., and Hovland, A. (2014). A vital role for complement in heart disease. *Molecular Immunology*, 61(2):126–134.

Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine*, 27(8):1133–1163.

LeBlanc, M., Zuber, V., Thompson, W. K., Andreassen, O. A., Frigessi, A., and Andreassen, B. K. (2018). A correction for sample overlap in genome-wide association studies in a polygenic pleiotropy-informed framework. *BMC genomics*, 19(1):1–15.

Levin, M. G., Klarin, D., Walker, V. M., Gill, D., Lynch, J., Lee, K. M., Assimes, T. L., Natarajan, P., Hung, A. M., Edwards, T., et al. (2020). Genetic variation in blood pressure and lifetime risk of peripheral artery disease: A mendelian randomization study. *medRxiv*.

Li, C., Shen, X., and Pan, W. (2021a). Inference for a large directed graphical model with interventions. *arXiv preprint arXiv:2110.03805*.

Li, M.-X., Gui, H.-S., Kwan, J. S., and Sham, P. C. (2011). Gates: a rapid and powerful gene-based association test using extended simes procedure. *The American Journal of Human Genetics*, 88(3):283–293.

Li, T., Ning, Z., and Shen, X. (2021b). Improved estimation of phenotypic correlations using summary association statistics. *Frontiers in genetics*, page 1291.

Liang, F. and Wang, Y. (2021). Coronary heart disease and atrial fibrillation: A vicious cycle. *American Journal of Physiology-Heart and Circulatory Physiology*, 320(1):H1–H12.

Lin, Z., Deng, Y., and Pan, W. (2021). Combining the strengths of inverse-variance weighting and egger regression in mendelian randomization using a mixture of regressions model. *PLoS genetics*, 17(11):e1009922.

Lin, Z., Pan, I., and Pan, W. (2022). A practical problem with egger regression in mendelian randomization. *PLoS genetics*, 18(5):e1010166.

Lin, Z., Xue, H., and Pan, W. (2023). Robust multivariable mendelian randomization based on constrained maximum likelihood. *The American Journal of Human Genetics*, 110(4):592–605.

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(2):237–244.

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(7538):197–206.

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(4):524–537.

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(9):981.

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(7):740–747.

Mounier, N. and Kutalik, Z. (2021). Bias correction for inverse variance weighting mendelian randomization. *bioRxiv*, pages 2021–03.

Nielsen, J. B., Thorolfsdottir, R. B., Fritsche, L. G., Zhou, W., Skov, M. W., Graham, S. E., Herron, T. J., McCarthy, S., Schmidt, E. M., Sveinbjornsson, G., et al. (2018). Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nature genetics*, 50(9):1234–1239.

Persson, C. E., Adiels, M., Björck, L., and Rosengren, A. (2018). Young women, body size and risk of atrial fibrillation. *European Journal of Preventive Cardiology*, 25(2):173–180.

Petot, G. J., Vega, U., Traore, F., Fritsch, T., Debanne, S. M., Friedland, R. P., and Lerner, A. J. (2007). Height and alzheimer's disease: findings from a case-control study. *Journal of Alzheimer's Disease*, 11(3):337–341.

Powell-Wiley, T. M., Poirier, P., Burke, L. E., Després, J.-P., Gordon-Larsen, P., Lavie, C. J., Lear, S. A., Ndumele, C. E., Neeland, I. J., Sanders, P., et al. (2021). Obesity and cardiovascular disease: a scientific statement from the american heart association. *Circulation*, 143(21):e984–e1010.

Qi, G. and Chatterjee, N. (2019). Mendelian randomization analysis using mixture models for robust and efficient estimation of causal effects. *Nature communications*, 10(1):1–10.

Rees, J. M., Wood, A. M., and Burgess, S. (2017). Extending the mr-egger method for multivariable mendelian randomization to correct for both measured and unmeasured pleiotropy. *Statistics in medicine*, 36(29):4705–4718.

Relton, C. L. and Davey Smith, G. (2012). Two-step epigenetic mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. *International journal of epidemiology*, 41(1):161–176.

Robinson, J. G., Farnier, M., Krempf, M., Bergeron, J., Luc, G., Averna, M., Stroes, E. S., Langslet, G., Raal, F. J., El Shahawy, M., et al. (2015). Efficacy and safety of alirocumab in reducing lipids and cardiovascular events. *New England Journal of Medicine*, 372(16):1489–1499.

Russ, T. C., Kivimäki, M., Starr, J. M., Stamatakis, E., and Batty, G. D. (2014). Height in relation to dementia death: individual participant meta-analysis of 18 uk prospective cohort studies. *The British Journal of Psychiatry*, 205(5):348–354.

Sabatine, M. S., Giugliano, R. P., Keech, A. C., Honarpour, N., Wiviott, S. D., Murphy, S. A., Kuder, J. F., Wang, H., Liu, T., Wasserman, S. M., et al. (2017). Evolocumab and clinical outcomes in patients with cardiovascular disease. *New England journal of medicine*, 376(18):1713–1722.

Sanderson, E. (2021). Multivariable mendelian randomization and mediation. *Cold Spring Harbor perspectives in medicine*, 11(2):a038984.

Sanderson, E., Davey Smith, G., Windmeijer, F., and Bowden, J. (2019). An examination of multivariable mendelian randomization in the single-sample and two-sample summary data settings. *International journal of epidemiology*, 48(3):713–727.

Sanderson, E., Glymour, M. M., Holmes, M. V., Kang, H., Morrison, J., Munafò, M. R., Palmer, T., Schooling, C. M., Wallace, C., Zhao, Q., et al. (2022). Mendelian randomization. *Nature Reviews Methods Primers*, 2(1):1–21.

Sanderson, E., Spiller, W., and Bowden, J. (2021). Testing and correcting for weak and pleiotropic instruments in two-sample multivariable mendelian randomization. *Statistics in medicine*, 40(25):5434–5452.

Schmidt, A. F., Finan, C., Gordillo-Marañón, M., Asselbergs, F. W., Freitag, D. F., Patel, R. S., Tyl, B., Chopade, S., Faraway, R., Zwierzyna, M., et al. (2020). Genetic drug target validation using mendelian randomisation. *Nature communications*, 11(1):3255.

Schooling, C. M., Luo, S., Yeung, S. L. A., Thompson, D. J., Karthikeyan, S., Bolton, T. R., Mason, A. M., Ingelsson, E., and Burgess, S. (2018). Genetic predictors of testosterone and their associations with cardiovascular disease and risk factors: A mendelian randomization investigation. *International journal of cardiology*, 267:171–176.

Scott, R. A., Lagou, V., Welch, R. P., Wheeler, E., Montasser, M. E., Luan, J., Mägi, R., Strawbridge, R. J., Rehnberg, E., Gustafsson, S., et al. (2012). Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nature genetics*, 44(9):991–1005.

Shahini, N., Michelsen, A. E., Nilsson, P. H., Ekholt, K., Gullestad, L., Broch, K., Dahl, C. P., Aukrust, P., Ueland, T., Mollnes, T. E., et al. (2017). The alternative complement pathway is dysregulated in patients with chronic heart failure. *Scientific reports*, 7(1):1–10.

Sleiman, P. M. and Grant, S. F. (2010). Mendelian randomization in the era of genomewide association studies. *Clinical Chemistry*, 56(5):723–728.

Slob, E. A. and Burgess, S. (2020). A comparison of robust mendelian randomization methods using summary data. *Genetic epidemiology*, 44(4):313–329.

Sohail, H., Hassan, S. M., Yaqoob, U., and Hassan, Z. (2021). The height as an independent risk factor of atrial fibrillation: A review. *Indian Heart Journal*, 73(1):22–25.

Tedrow, U. B., Conen, D., Ridker, P. M., Cook, N. R., Koplan, B. A., Manson, J. E., Buring, J. E., and Albert, C. M. (2010). The long-and short-term impact of elevated body mass index on the risk of new atrial fibrillation: the whs (women's health study). *Journal of the American College of Cardiology*, 55(21):2319–2327.

the CARDIoGRAMplusC4D Consortium (2015). A comprehensive 1000 genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature genetics*, 47(10):1121–1130.

Thorgeirsson, T., Gudbjartsson, D., Sulem, P., Besenbacher, S., Styrkarsdottir, U., Thorleifsson, G., Walters, G., Furberg, H., Sullivan, P., Marchini, J., et al. (2013). A common biological basis of obesity and nicotine addiction. *Translational psychiatry*, 3(10):e308–e308.

Timpson, N. J., Nordestgaard, B. G., Harbord, R. M., Zacho, J., Frayling, T. M., Tybjærg-Hansen, A., and Smith, G. D. (2011). C-reactive protein levels and body

mass index: elucidating direction of causation through reciprocal mendelian randomization. *International journal of obesity*, 35(2):300–308.

Turley, P., Walters, R. K., Maghzian, O., Okbay, A., Lee, J. J., Fontana, M. A., Nguyen-Viet, T. A., Wedow, R., Zacher, M., Furlotte, N. A., et al. (2018). Multi-trait analysis of genome-wide association summary statistics using mtag. *Nature genetics*, 50(2):229–237.

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(3):433–443.

Wang, A., Liu, W., and Liu, Z. (2022a). A two-sample robust bayesian mendelian randomization method accounting for linkage disequilibrium and idiosyncratic pleiotropy with applications to the covid-19 outcomes. *Genetic epidemiology*, 46(3-4):159–169.

Wang, K. (2018). Understanding power anomalies in mediation analysis. *Psychometrika*, 83(2):387–406.

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

Wang, K., Shi, X., Zhu, Z., Hao, X., Chen, L., Cheng, S., Foo, R. S., and Wang, C. (2022b). Mendelian randomization analysis of 37 clinical factors and coronary artery disease in east asian and european populations. *Genome Medicine*, 14(1):1–15.

Waterworth, D. M., Ricketts, S. L., Song, K., Chen, L., Zhao, J. H., Ripatti, S., Aulchenko, Y. S., Zhang, W., Yuan, X., Lim, N., et al. (2010). Genetic variants

influencing circulating lipid levels and risk of coronary artery disease. *Arteriosclerosis, thrombosis, and vascular biology*, 30(11):2264–2276.

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(11):1274.

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(11):1173–1186.

Xue, H. and Pan, W. (2020). Inferring causal direction between two traits in the presence of horizontal pleiotropy with gwas summary data. *PLoS genetics*, 16(11):e1009105.

Xue, H. and Pan, W. (2022). Robust inference of bi-directional causal relationships in presence of correlated pleiotropy with gwas summary data. *PLoS genetics*, 18(5):e1010205.

Xue, H., Shen, X., and Pan, W. (2021). Constrained maximum likelihood-based mendelian randomization robust to both correlated and uncorrelated pleiotropic effects. *The American Journal of Human Genetics*, 108(7):1251–1269.

Xue, H., Shen, X., and Pan, W. (2023). Causal inference in transcriptome-wide association studies with invalid instruments and gwas summary data. *Journal of the American Statistical Association*, 0(0):1–27.

Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., of ANthropometric Traits (GIANT) Consortium, G. I., Replication, D. G., analysis (DIAGRAM) Consortium, M., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., et al. (2012).

Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits. *Nature genetics*, 44(4):369–375.

Yavorska, O. O. and Burgess, S. (2017). Mendelianrandomization: an r package for performing mendelian randomization analyses using summarized data. *International journal of epidemiology*, 46(6):1734–1739.

Young, N. (1981). The rate of convergence of a matrix power series. *Linear Algebra and its Applications*, 35:261–278.

Yuan, Y., Shen, X., Pan, W., and Wang, Z. (2019). Constrained likelihood for reconstructing a directed acyclic gaussian graph. *Biometrika*, 106(1):109–125.

Zhang, J., Dutta, D., Köttgen, A., Tin, A., Schlosser, P., Grams, M. E., Harvey, B., Consortium, C., Yu, B., Boerwinkle, E., et al. (2022). Plasma proteome analyses in individuals of european and african ancestry identify cis-pqtls and models for proteome-wide association studies. *Nature Genetics*, 54(5):593–602.

Zhao, H., Rasheed, H., Nøst, T. H., Cho, Y., Liu, Y., Bhatta, L., Bhattacharya, A., Hemani, G., Smith, G. D., Brumpton, B. M., et al. (2022). Proteome-wide mendelian randomization in global biobank meta-analysis reveals multi-ancestry drug targets for common diseases. *Cell Genomics*, 2(11):100195.

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. *The Annals of Statistics*, 48(3):1742–1769.

Zheng, J., Zhang, Y., Zhao, H., Liu, Y., Baird, D., Karim, M. A., Ghoussaini, M., Schwartzentruber, J., Dunham, I., Elsworth, B., et al. (2022). Multi-ancestry mendelian randomization of omics traits revealing drug targets of covid-19 severity. *EBioMedicine*, 81:104112.

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(52):33117–33123.

Zhu, X. (2021). Mendelian randomization and pleiotropy analysis. *Quantitative Biology*, 9(2):122–132.

Zhu, X., Gerstein, M., and Snyder, M. (2007). Getting connected: analysis and principles of biological networks. *Genes & development*, 21(9):1010–1024.

Zhu, Z., Wang, K., Hao, X., Chen, L., Liu, Z., and Wang, C. (2022). Causal graph among serum lipids and glycemic traits: a mendelian randomization study. *Diabetes*, 71(8):1818–1826.

Zuber, V., Colijn, J. M., Klaver, C., and Burgess, S. (2020). Selecting likely causal risk factors from high-throughput experiments using multivariable mendelian randomization. *Nature communications*, 11(1):1–11.

Zuber, V., Grinberg, N. F., Gill, D., Manipur, I., Slob, E. A., Patel, A., Wallace, C., and Burgess, S. (2022). Combining evidence from mendelian randomization and colocalization: Review and comparison of approaches. *The American Journal of Human Genetics*.

Appendix A

Supplementary material for Chapter 2

A.1 Theory

A.1.1 Large-sample theory

The estimation and selection consistency (of BIC) with the cMLE developed in Xue et al. (2021) can be carried over to the set-up with overlapping samples. Here we first state the main conclusions with the proofs provided in the following subsections.

Denote $\hat{\theta} = \hat{\theta}(\hat{K})$ the cMLE using the set of selected valid IVs $\hat{B}^C_{\hat{K}}$. Now we state three assumptions used to prove the estimation consistency and asymptotic normality of our proposed cMLE $\hat{\theta}$.

Assumption A.1

For every SNP $i = 1, \ldots, m$, $\binom{\hat{\beta}_{Xi}}{\hat{\beta}_{Yi}} \sim \mathcal{N}\Big(\binom{b_{Xi}}{b_{Yi}}, \begin{pmatrix} \sigma^2_{Xi} & \rho\sigma_{Xi}\sigma_{Yi} \\ \rho\sigma_{Xi}\sigma_{Yi} & \sigma^2_{Yi} \end{pmatrix}\Big)$ with the known variances $(\sigma^2_{Xi}, \sigma^2_{Yi})$ and correlation ρ. Furthermore, the m pairs of $(\hat{\beta}_{Xi}, \hat{\beta}_{Yi})_{i=1}^m$ are mutually independent.

Assumption A.2

(Plurality valid condition.) Suppose that B_0 is the index set of the true invalid IVs

with $K_0 = |B_0|$. For any $B \subseteq \{1, \ldots, m\}$ and $|B| = K_0$, if $B \neq B_0$, then the $(m - K_0)$ ratios $\{r_i/b_{Xi}, i \in B^c\}$ are not all equal.

Assumption A.3

(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 $l_X/N_1 \leq \sigma_{Xi}^2 \leq u_X/N_1$, $l_Y/N_2 \leq \sigma_{Yi}^2 \leq u_Y/N_2$, and $l_N \cdot N_2 \leq N_1 \leq u_N \cdot N_2$ for $i = 1, \ldots, m$.

It is noted that Assumption A.1 is reasonable given the usual large sample sizes of GWAS. As usual, for each pair of traits, the m SNPs are selected to be independent, implying that their corresponding SNP-trait association estimates are nearly independent (Zhao et al., 2020). As discussed in Xue et al. (2021), Assumption A.2 is relatively weak as compared to those adopted by many other MR methods; in particular, it is weaker than the majority valid assumption (i.e. at least more than 50% of IVs are valid); as usual, if GWAS summary data are based on the MLEs, Assumption A.3 holds.

Theorem A.1

With Assumption A.1 to Assumption A.3 satisfied, the proposed BIC consistently select invalid IVs, and the proposed constrained maximum likelihood estimator $\hat{\theta}$, combined with the use of the BIC selection criterion, is consistent for the true causal effect size θ_0, and

$$\sqrt{V^*}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, 1) \text{ as } N \to \infty,$$

where

$$V^* = \sum_{i \in \hat{B}_{\hat{K}}^C} \frac{b_{Xi}^2 \sigma_{Yi}^2 + b_{Yi}^2 \sigma_{Xi}^2 - 2\rho b_{Xi} b_{Yi} \sigma_{Xi} \sigma_{Yi}}{(\sigma_{Yi}^2 + \theta_0^2 \sigma_{Xi}^2 - 2\rho\theta_0 \sigma_{Xi} \sigma_{Yi})^2}$$

is consistently estimated by substituting θ, b_{Xi} and b_{Yi} with $\hat{\theta}$, $\hat{\beta}_{Xi}$ and $\hat{\beta}_{Yi}$ respectively.

To prove Theorem A.1, we first show that the selection consistency is achieved by the proposed BIC. Then after correctly selecting out invalid IVs, the resulting cMLE of θ is the same as the maximum profile likelihood estimator being applied to all valid IVs, obtaining its estimation consistency and asymptotic normality similar to Theorems 3.1 and 3.2 in Zhao et al. (2020). The details are given in the Supplementary S1.2. We also show in the Supplementary S1.3 that the Fisher information-based variance estimator of the cMLE used in our method and the one in Theorem A.1 are asymptotically equivalent.

We note that the selection consistency of BIC in MR-cML-BIC-I as proposed in Xue et al. (2021) still holds despite the violation of its assumption due to the presence of sample overlap, so do the estimation consistency and asymptotic normality of MR-cML-BIC-I. However, the corresponding estimation efficiency is lower than that of MR-cML-BIC-C. Furthermore, the usual (naive or model-based) variance estimator in Xue et al. (2021) is not consistent; instead, we propose a robust variance estimator. Here we state the conclusion with the details given in Supplementary Section S1.4.

Theorem A.2

With Assumption A.1 to Assumption A.3 satisfied, under the (mis-specified) working independence model as proposed in Xue et al. (2021), the BIC can still consistently select invalid IVs, and the constrained maximum likelihood estimator $\hat{\theta}_I$, combined with the use of the BIC selection criterion, is consistent for the true causal effect size θ_0, and

$$\sqrt{V_r}(\hat{\theta}_I - \theta_0) \xrightarrow{d} \mathcal{N}(0,1) \text{ as } N \to \infty,$$

where

$$V_r^{-1} = \mathbb{E}[\psi'(\theta_0; \rho = 0)]^{-1}\mathbb{E}[\psi^2(\theta_0; \rho = 0)]\mathbb{E}[\psi'(\theta_0; \rho = 0)]^{-1},$$

and $\psi(\theta; \rho = 0) = \partial \max_{\mathbf{b}_X} l(\theta, \mathbf{b}_X; \rho = 0)/\partial\theta$ is the profile log-likelihood score under

the working independence model (i.e., $\rho = 0$); V_r can be consistently estimated by its sample version (as the sandwich estimator).

Since we apply the proposed MR-cML-BIC-C on every pair of traits to construct a total causal graph, we'd expect that the consistency and asymptotic normality can be obtained for the estimated total causal graph and direct causal graph as stated below. We are now considering the GWAS summary data from all traits together with possibly correlated SNPs as IVs, requiring a joint normal distribution of all the SNP-trait association estimates, which is reasonable based on large-scale GWAS data.

Assumption A.4

The vector of all SNP-trait association estimates $\hat{\beta}_{Xi}$'s and $\hat{\beta}_{Yi}$'s (across all traits X and Y and across all SNPs i) has a multivariate normal distribution.

Assumption A.5

$\mathbf{G}_{tot} = \mathbf{G}_{dir} + \mathbf{G}_{dir}^2 + \mathbf{G}_{dir}^3 + \cdots = \mathbf{G}_{dir}(\mathbf{I} - \mathbf{G}_{dir})^{-1}$.

Corollary A.1

With Assumption A.1 to Assumption A.5 satisfied, if the diagonal elements of $\widehat{\mathbf{G}}_{tot}$ are set consistently, then $\text{vec}(\widehat{\mathbf{G}}_{tot})$ and $\text{vec}(\widehat{\mathbf{G}}_{dir})$ are consistent for $\text{vec}(\mathbf{G}_{tot})$ and $\text{vec}(\mathbf{G}_{dir})$ respectively, and are asymptotically normally distributed.

A.1.2 Proof of Theorem A.1

To prove the estimation consistency and asymptotic normality of the cMLE, we will first prove the selection consistency of BIC following the approach in the Supplementary S1 in Xue et al. (2021), i.e., it will select the correct set of invalid IVs as the sample size N goes to infinity. Then we prove the estimation consistency and asymptotic normality based on the set of (selected) valid IVs.

Lemma A.1

With assumptions 1-3 satisfied, if $K_0 \in \mathcal{K}$, we have $P(\hat{K} = K_0) \to 1$ and $P(\hat{B}_{\hat{K}} = B_0) \to 1$ as $N_1, N_2 \to \infty$.

Proof. First, we show $P(\hat{B}_{K_0} = B_0) \to 1$, which is equivalent to show for any $B_1 \subseteq \{1, \cdots, m\}$ such that $|B_1| = K_0$ and $B_1 \neq B_0$, $P(\hat{B}_{K_0} = B_1) \to 0$ as $N_1, N_2 \to \infty$. We have

$$
\begin{aligned}
&P(\hat{B}_{K_0} = B_1) \\
&\leq P\Big\{ \min_{\tilde{\theta}, \tilde{b}_{Xi}} \sum_{i \in B_1^c} \left(\frac{(\hat{\beta}_{Xi} - \tilde{b}_{Xi})^2}{\sigma_{Xi}^2} + \frac{(\hat{\beta}_{Yi} - \tilde{\theta}\tilde{b}_{Xi})^2}{\sigma_{Yi}^2} - 2\rho \frac{(\hat{\beta}_{Xi} - \tilde{b}_{Xi})(\hat{\beta}_{Yi} - \tilde{\theta}\tilde{b}_{Xi})}{\sigma_{Xi}\sigma_{Yi}} \right) \\
&\qquad \leq \sum_{i \in B_0^c} \left(\frac{(\hat{\beta}_{Xi} - b_{Xi})^2}{\sigma_{Xi}^2} + \frac{(\hat{\beta}_{Yi} - \theta b_{Xi})^2}{\sigma_{Yi}^2} - 2\rho \frac{(\hat{\beta}_{Xi} - b_{Xi})(\hat{\beta}_{Yi} - \theta b_{Xi})}{\sigma_{Xi}\sigma_{Yi}} \right) \Big\}.
\end{aligned}
$$

Note that, for $i \in B_0^c$, $\left(\begin{smallmatrix} \hat{\beta}_{Xi} - b_{Xi} \\ \hat{\beta}_{Yi} - \theta b_{Xi} \end{smallmatrix} \right) \sim \mathcal{N}\Big(\mathbf{0}, \left(\begin{smallmatrix} \sigma_{Xi}^2 & \rho\sigma_{Xi}\sigma_{Yi} \\ \rho\sigma_{Xi}\sigma_{Yi} & \sigma_{Yi}^2 \end{smallmatrix} \right) \Big)$. So for any $\epsilon > 0$, there exists $C > 0$ such that

$$
P\Big\{ \sum_{i \in B_0^c} \left(\frac{(\hat{\beta}_{Xi} - b_{Xi})^2}{\sigma_{Xi}^2} + \frac{(\hat{\beta}_{Yi} - \theta b_{Xi})^2}{\sigma_{Yi}^2} - 2\rho \frac{(\hat{\beta}_{Xi} - b_{Xi})(\hat{\beta}_{Yi} - \theta b_{Xi})}{\sigma_{Xi}\sigma_{Yi}} \right) > C \Big\} < \frac{\epsilon}{2}. \tag{A.1}
$$

And we have

$$P\Big\{ \min_{\tilde{\theta},\tilde{b}_{Xi}} \sum_{i\in B_1^c} \left(\frac{(\hat{\beta}_{Xi}-\tilde{b}_{Xi})^2}{\sigma_{Xi}^2} + \frac{(\hat{\beta}_{Yi}-\tilde{\theta}\tilde{b}_{Xi})^2}{\sigma_{Yi}^2} - 2\rho\frac{(\hat{\beta}_{Xi}-\tilde{b}_{Xi})(\hat{\beta}_{Yi}-\tilde{\theta}\tilde{b}_{Xi})}{\sigma_{Xi}\sigma_{Yi}} \right)$$
$$\leq \sum_{i\in B_0^c} \left(\frac{(\hat{\beta}_{Xi}-b_{Xi})^2}{\sigma_{Xi}^2} + \frac{(\hat{\beta}_{Yi}-\theta b_{Xi})^2}{\sigma_{Yi}^2} - 2\rho\frac{(\hat{\beta}_{Xi}-b_{Xi})(\hat{\beta}_{Yi}-\theta b_{Xi})}{\sigma_{Xi}\sigma_{Yi}} \right) \Big\}$$
$$\leq P\Big\{ \min_{\tilde{\theta},\tilde{b}_{Xi}} \sum_{i\in B_1^c} \left(\frac{(\hat{\beta}_{Xi}-\tilde{b}_{Xi})^2}{\sigma_{Xi}^2} + \frac{(\hat{\beta}_{Yi}-\tilde{\theta}\tilde{b}_{Xi})^2}{\sigma_{Yi}^2} - 2\rho\frac{(\hat{\beta}_{Xi}-b_{Xi})(\hat{\beta}_{Yi}-\theta b_{Xi})}{\sigma_{Xi}\sigma_{Yi}} \right) \leq C \Big\}$$
$$+ P\Big\{ \sum_{i\in B_0^c} \left(\frac{(\hat{\beta}_{Xi}-\beta_{Xi})^2}{\sigma_{Xi}^2} + \frac{(\hat{\beta}_{Yi}-\theta\beta_{Xi})^2}{\sigma_{Yi}^2} - 2\rho\frac{(\hat{\beta}_{Xi}-b_{Xi})(\hat{\beta}_{Yi}-\theta b_{Xi})}{\sigma_{Xi}\sigma_{Yi}} \right) > C \Big\}.$$

After profiling out $\tilde{b}_{Xi}$'s, we get

$$\min_{\tilde{\theta},\tilde{b}_{Xi}} \sum_{i\in B_1^c} \left(\frac{(\hat{\beta}_{Xi}-\tilde{b}_{Xi})^2}{\sigma_{Xi}^2} + \frac{(\hat{\beta}_{Yi}-\tilde{\theta}\tilde{b}_{Xi})^2}{\sigma_{Yi}^2} - 2\rho\frac{(\hat{\beta}_{Xi}-\tilde{b}_{Xi})(\hat{\beta}_{Yi}-\tilde{\theta}\tilde{b}_{Xi})}{\sigma_{Xi}\sigma_{Yi}} \right)$$
$$= \min_{\tilde{\theta}} \sum_{i\in B_1^c} \frac{(\hat{\beta}_{Yi}-\tilde{\theta}\cdot\hat{\beta}_{Xi})^2(1-\rho^2)}{\sigma_{Yi}^2+\tilde{\theta}^2\sigma_{Xi}^2-2\rho\tilde{\theta}\sigma_{Xi}\sigma_{Yi}},$$

so

$$P\Big\{ \min_{\tilde{\theta},\tilde{b}_{Xi}} \sum_{i\in B_1^c} \left(\frac{(\hat{\beta}_{Xi}-\tilde{b}_{Xi})^2}{\sigma_{Xi}^2} + \frac{(\hat{\beta}_{Yi}-\tilde{\theta}\tilde{b}_{Xi})^2}{\sigma_{Yi}^2} - 2\rho\frac{(\hat{\beta}_{Xi}-b_{Xi})(\hat{\beta}_{Yi}-\theta b_{Xi})}{\sigma_{Xi}\sigma_{Yi}} \right) \leq C \Big\}$$
$$= P\Big\{ \min_{\tilde{\theta}} \sum_{i\in B_1^c} \frac{(\hat{\beta}_{Yi}-\tilde{\theta}\cdot\hat{\beta}_{Xi})^2}{\sigma_{Yi}^2+\tilde{\theta}^2\sigma_{Xi}^2-2\rho\tilde{\theta}\sigma_{Xi}\sigma_{Yi}} \leq C/(1-\rho^2) \Big\}.$$

We have $\frac{\hat{\beta}_{Yi}-\tilde{\theta}\cdot\hat{\beta}_{Xi}}{\sqrt{\sigma_{Yi}^2+\tilde{\theta}^2\sigma_{Xi}^2-2\rho\tilde{\theta}\sigma_{Xi}\sigma_{Yi}}} \sim \mathcal{N}(\frac{\theta\cdot b_{Xi}+r_i-\tilde{\theta}\cdot b_{Xi}}{\sqrt{\sigma_{Yi}^2+\tilde{\theta}^2\sigma_{Xi}^2-2\rho\tilde{\theta}\sigma_{Xi}\sigma_{Yi}}}, 1)$, so
$\sum_{i\in B_1^c} \frac{(\hat{\beta}_{Yi}-\tilde{\theta}\cdot\hat{\beta}_{Xi})^2}{\sigma_{Yi}^2+\tilde{\theta}^2\sigma_{Xi}^2-2\rho\tilde{\theta}\sigma_{Xi}\sigma_{Yi}}$ follows non-central χ^2 distribution with degrees of freedom

$(m - K_0)$ and non-centrality parameter $\lambda_{\tilde{\theta}}$ depending on $\tilde{\theta}$

$$\lambda_{\tilde{\theta}} = \sum_{i \in B_1^c} \frac{(\theta \cdot b_{Xi} + r_i - \tilde{\theta} \cdot b_{Xi})^2}{\sigma_{Yi}^2 + \tilde{\theta}^2 \sigma_{Xi}^2 - 2\rho\tilde{\theta}\sigma_{Xi}\sigma_{Yi}}.$$

With Assumption A.3, we get

$$\lambda_{\tilde{\theta}} \geq \sum_{i \in B_1^c} \frac{(\theta \cdot b_{Xi} + r_i - \tilde{\theta} \cdot b_{Xi})^2}{\frac{u_Y}{N_2} + \tilde{\theta}^2 \cdot \frac{u_X}{l_N \cdot N_2} + 2|\tilde{\theta}| \frac{\sqrt{u_X u_Y}}{N_2 \sqrt{l_N}}} = N_2 \cdot \sum_{i \in B_1^c} \frac{(\theta \cdot b_{Xi} + r_i - \tilde{\theta} \cdot b_{Xi})^2}{u_Y + \tilde{\theta}^2 \cdot \frac{u_X}{l_N} + 2|\tilde{\theta}| \sqrt{u_X u_Y / l_N}}.$$

With Assumption A.2, we know

$$\min_{\tilde{\theta}} \sum_{i \in B_1^c} \frac{(\theta \cdot b_{Xi} + r_i - \tilde{\theta} \cdot b_{Xi})^2}{u_Y + \tilde{\theta}^2 \cdot \frac{u_X}{l_N} + 2|\tilde{\theta}| \sqrt{u_X u_Y / l_N}} = v > 0,$$

here v is a constant. This is because, with Assumption A.2, there is no $\tilde{\theta}$ making $\theta \cdot b_{Xi} + r_i - \tilde{\theta} \cdot b_{Xi} = 0$ for all $i \in B_1^c$ simultaneously. So we have $\min_{\tilde{\theta}} \lambda_{\tilde{\theta}} \geq N_2 \cdot v$. Then as N_2 large enough, we have

$$P\Big\{ \min_{\tilde{\theta}} \sum_{i \in B_1^c} \frac{(\hat{\beta}_{Yi} - \tilde{\theta} \cdot \hat{\beta}_{Xi})^2}{\sigma_{Yi}^2 + \tilde{\theta}^2 \sigma_{Xi}^2 - 2\rho\tilde{\theta}\sigma_{Xi}\sigma_{Yi}} \leq C/(1 - \rho^2) \Big\} \leq \frac{\epsilon}{2}. \tag{A.2}$$

Combining (Eq (A.1)) and (Eq (A.2)), we get $P(\hat{B}_{K_0} = B_0) \to 1$ as $N_1, N_2 \to \infty$.

Next, we show $P(\hat{K} = K_0) \to 1$. For any $K_1 < K_0$, we have

$$\begin{aligned} P(\hat{K} = K_1) &\leq P\big\{ \mathrm{BIC}(K_1) \leq \mathrm{BIC}(K_0) \big\} \\ &= P\Big\{ 2l\left(\hat{\theta}(K_0), \hat{b}_{Xi}(K_0), \hat{r}_i(K_0)\right) - 2l\left(\hat{\theta}(K_1), \hat{b}_{Xi}(K_1), \hat{r}_i(K_1)\right) \\ &\qquad \leq \log(N)(K_0 - K_1) \Big\}. \end{aligned}$$

As we have shown $P(\hat{B}_{K_0} = B_0) \to 1$, with probability goes to 1 we have

$$\begin{aligned}
&2l\left(\hat{\theta}(K_0), \hat{b}_{Xi}(K_0), \hat{r}_i(K_0)\right) - 2l\left(\hat{\theta}(K_1), \hat{b}_{Xi}(K_1), \hat{r}_i(K_1)\right) \\
\geq &\min_{\tilde{\theta}, \tilde{b}_{Xi}} \sum_{i \in \hat{B}^c_{K_1}} \left(\frac{(\hat{\beta}_{Xi} - \tilde{b}_{Xi})^2}{\sigma^2_{Xi}} + \frac{(\hat{\beta}_{Yi} - \tilde{\theta}\tilde{b}_{Xi})^2}{\sigma^2_{Yi}} - 2\rho \frac{(\hat{\beta}_{Xi} - \tilde{b}_{Xi})(\hat{\beta}_{Yi} - \tilde{\theta}\tilde{b}_{Xi})}{\sigma_{Xi}\sigma_{Yi}} \right) \\
&- \sum_{i \in B^c_0} \left(\frac{(\hat{\beta}_{Xi} - b_{Xi})^2}{\sigma^2_{Xi}} + \frac{(\hat{\beta}_{Yi} - \theta b_{Xi})^2}{\sigma^2_{Yi}} - 2\rho \frac{(\hat{\beta}_{Xi} - b_{Xi})(\hat{\beta}_{Yi} - \theta b_{Xi})}{\sigma_{Xi}\sigma_{Yi}} \right).
\end{aligned}$$

Then we get

$$\begin{aligned}
P(\hat{K} = K_1) \leq \sum_{|B| = K_1} P\Big\{ &\min_{\tilde{\theta}, \tilde{b}_{Xi}} \sum_{i \in B^c} \left(\frac{(\hat{\beta}_{Xi} - \tilde{b}_{Xi})^2}{\sigma^2_{Xi}} + \frac{(\hat{\beta}_{Yi} - \tilde{\theta}\tilde{b}_{Xi})^2}{\sigma^2_{Yi}} - 2\rho \frac{(\hat{\beta}_{Xi} - \tilde{b}_{Xi})(\hat{\beta}_{Yi} - \tilde{\theta}\tilde{b}_{Xi})}{\sigma_{Xi}\sigma_{Yi}} \right) \\
&\leq \sum_{i \in B^c_0} \left(\frac{(\hat{\beta}_{Xi} - b_{Xi})^2}{\sigma^2_{Xi}} + \frac{(\hat{\beta}_{Yi} - \theta b_{Xi})^2}{\sigma^2_{Yi}} - 2\rho \frac{(\hat{\beta}_{Xi} - b_{Xi})(\hat{\beta}_{Yi} - \theta b_{Xi})}{\sigma_{Xi}\sigma_{Yi}} \right) \\
&\quad + \log(N)(K_0 - K_1)(1 - \rho^2) \Big\}.
\end{aligned}$$

Similar as above, we get

$$\begin{aligned}
&\min_{\tilde{\theta}, \tilde{b}_{Xi}} \sum_{i \in B^c} \left(\frac{(\hat{\beta}_{Xi} - \tilde{b}_{Xi})^2}{\sigma^2_{Xi}} + \frac{(\hat{\beta}_{Yi} - \tilde{\theta}\tilde{b}_{Xi})^2}{\sigma^2_{Yi}} - 2\rho \frac{(\hat{\beta}_{Xi} - \tilde{b}_{Xi})(\hat{\beta}_{Yi} - \tilde{\theta}\tilde{b}_{Xi})}{\sigma_{Xi}\sigma_{Yi}} \right) \\
&= \min_{\tilde{\theta}} \sum_{i \in B^c} \frac{(\hat{\beta}_{Yi} - \tilde{\theta} \cdot \hat{\beta}_{Xi})^2 (1 - \rho^2)}{\sigma^2_{Yi} + \tilde{\theta}^2 \sigma^2_{Xi} - 2\rho\tilde{\theta}\sigma_{Xi}\sigma_{Yi}},
\end{aligned}$$

and $\sum_{i \in B^c} \frac{(\hat{\beta}_{Yi} - \tilde{\theta} \cdot \hat{\beta}_{Xi})^2}{\sigma^2_{Yi} + \tilde{\theta}^2 \sigma^2_{Xi} - 2\rho\tilde{\theta}\sigma_{Xi}\sigma_{Yi}}$ follows non-central χ^2 distribution with degrees of freedom $(m - K_1)$ and non-centrality parameter $\lambda_{\tilde{\theta}}$ depending on $\tilde{\theta}$

$$\lambda_{\tilde{\theta}} = \sum_{i \in B^c} \frac{(\theta \cdot b_{Xi} + r_i - \tilde{\theta} \cdot b_{Xi})^2}{\sigma^2_{Yi} + \tilde{\theta}^2 \sigma^2_{Xi} - 2\rho\tilde{\theta}\sigma_{Xi}\sigma_{Yi}}.$$

Similarly, since $K_1 < K_0$, and with Assumption 3 we have $\lambda_{\tilde{\theta}} \geq N_2 \cdot v$ for some

positive constant v, so for any $|B| = K_1$, we get

$$P\Big\{ \min_{\tilde{\theta},\tilde{b}_{Xi}} \sum_{i\in B^c} \left(\frac{(\hat{\beta}_{Xi} - \tilde{b}_{Xi})^2}{\sigma_{Xi}^2} + \frac{(\hat{\beta}_{Yi} - \tilde{\theta}\tilde{b}_{Xi})^2}{\sigma_{Yi}^2} - 2\rho \frac{(\hat{\beta}_{Xi} - \tilde{b}_{Xi})(\hat{\beta}_{Yi} - \tilde{\theta}\tilde{b}_{Xi})}{\sigma_{Xi}\sigma_{Yi}} \right)$$
$$\leq \sum_{i\in B_0^c} \left(\frac{(\hat{\beta}_{Xi} - \beta_{Xi})^2}{\sigma_{Xi}^2} + \frac{(\hat{\beta}_{Yi} - \theta\beta_{Xi})^2}{\sigma_{Yi}^2} - 2\rho \frac{(\hat{\beta}_{Xi} - b_{Xi})(\hat{\beta}_{Yi} - \theta b_{Xi})}{\sigma_{Xi}\sigma_{Yi}} \right)$$
$$+ \log(N)(K_0 - K_1)(1-\rho^2) \Big\} \to 0$$

This gives us $P(\hat{K} = K_1) \to 0$ for any $K_1 < K_0$. For any $K_1 > K_0$, we have

$$\begin{aligned} & P(\hat{K} = K_1) \\ \leq & P\Big\{ \log(N)(K_1 - K_0) \\ & \leq \sum_{i\in B_0^c} \frac{1}{1-\rho^2} \left(\frac{(\hat{\beta}_{Xi} - \beta_{Xi})^2}{\sigma_{Xi}^2} + \frac{(\hat{\beta}_{Yi} - \theta\beta_{Xi})^2}{\sigma_{Yi}^2} - 2\rho \frac{(\hat{\beta}_{Xi} - b_{Xi})(\hat{\beta}_{Yi} - \theta b_{Xi})}{\sigma_{Xi}\sigma_{Yi}} \right) \Big\} \end{aligned}$$

Since $\sum_{i\in B_0^c} \frac{1}{1-\rho^2} \left(\frac{(\hat{\beta}_{Xi} - \beta_{Xi})^2}{\sigma_{Xi}^2} + \frac{(\hat{\beta}_{Yi} - \theta\beta_{Xi})^2}{\sigma_{Yi}^2} - 2\rho \frac{(\hat{\beta}_{Xi} - b_{Xi})(\hat{\beta}_{Yi} - \theta b_{Xi})}{\sigma_{Xi}\sigma_{Yi}} \right)$ is a weighted sum of χ_1^2, we get $P(\hat{K} = K_1) \to 0$ for any $K_1 > K_0$. So we have $P(\hat{K} = K_0) \to 1$ as $N_1, N_2 \to \infty$. □

After correctly selecting invalid IVs, the resulted cMLE of θ is the same as the maximum profile likelihood estimator (MPLE) (profiling out $\mathbf{b}_{XB}$) being applied to all (selected) valid IVs. And we will prove the consistency and asymptotic normality based on the MPLE. For the simplicity of notation, we assume all m IVs are valid from now on. We follow the similar proofs given in Zhao et al. (2020), but we only consider the 'fixed m large N' scenario here.

Under the theoretical model that m IVs are valid (i.e., $b_{Yi} = \theta b_{Xi}$) and Assumption

1, the log-likelihood function is given by

$$l(\theta, b_{X1}, \ldots, b_{Xm}) = -\frac{1}{2(1-\rho^2)} \sum_{i=1}^{m} \Big(\frac{(\hat{\beta}_{Xi} - b_{Xi})^2}{\sigma_{Xi}^2} + \frac{(\hat{\beta}_{Yi} - \theta b_{Xi})^2}{\sigma_{Yi}^2} - 2\rho \frac{(\hat{\beta}_{Xi} - b_{Xi})}{\sigma_{Xi}} \frac{(\hat{\beta}_{Yi} - \theta b_{Xi})}{\sigma_{Yi}} \Big). \tag{A.3}$$

The true causal effect θ_0 is of interest, and $\mathbf{b}_X = (b_{X1}, \ldots, b_{Xm})$ are treated as nuisance parameters. Then the profile log-likelihood of θ is given by profiling out $\mathbf{b}_X$ in Eq (A.3):

$$l_p(\theta) = \max_{\mathbf{b}_X} l(\theta, \mathbf{b}_X) = -\frac{1}{2} \sum_{i=1}^{m} \frac{(\hat{\beta}_{Yi} - \theta \hat{\beta}_{Xi})^2}{\sigma_{Yi}^2 + \theta^2 \sigma_{Xi}^2 - 2\rho\theta\sigma_{Xi}\sigma_{Yi}}. \tag{A.4}$$

The maximum likelihood estimator of θ is given by $\hat{\theta} = \arg\max_\theta l_p(\theta)$.

It's noted that since we consider a fixed number of IVs, we have $||\mathbf{b}_X||_2^2 = O(1)$ (Assumption 2 in Zhao et al. (2020)). Furthermore, Assumption 3 implies that there exists constants c_σ, c'_σ such that $c_\sigma/N \leq \sigma_{Xi}^2 \leq c'_\sigma/N$ and $c_\sigma/N \leq \sigma_{Yi}^2 \leq c'_\sigma/N$ for $i = 1, \ldots, m$ (Assumption 3 in Zhao et al. (2020)). Now we are ready to state the estimation consistency of $\hat{\theta}$.

Lemma A.2

Under the model that all m IVs are valid and Assumptions 1-3, the maximum likelihood estimator $\hat{\theta}$ is consistent, that is, $\hat{\theta} \xrightarrow{p} \theta_0$ as $N \to \infty$.

Proof. Let $e_i = \hat{\beta}_{Yi} - b_{Yi}$, $\epsilon_i = \hat{\beta}_{Xi} - b_{Xi}$. After some algebra, we have

$$l_p(\theta) = -\frac{1}{2} \sum_{i=1}^{m} \frac{b_{Xi}^2(\theta_0 - \theta)^2 + (e_i - \theta\epsilon_i)^2 + 2b_{Xi}(\theta_0 - \theta)(e_i - \theta\epsilon_i)}{\sigma_{Yi}^2 + \theta^2\sigma_{Xi}^2 - 2\rho\theta\sigma_{Xi}\sigma_{Yi}}.$$

Notice that $e_i - \theta\epsilon_i \sim \mathcal{N}(0, \theta^2\sigma_{Xi}^2 + \sigma_{Yi}^2 - 2\theta\rho\sigma_{Xi}\sigma_{Yi})$. Follow the same argument in

Zhao et al. (2020), we have

$$-2l_p(\theta) \geq \frac{N||\mathbf{b}_X||_2^2}{2c_\sigma} \min\left((\theta_0-\theta)^2, \frac{(\theta_0-\theta)^2}{\theta^2+2|\theta|}\right) + m + O_p(\sqrt{m}+\sqrt{N}||\mathbf{b}_X|| \cdot |\theta_0-\theta|).$$

Similarly, we can show that for any $\epsilon > 0$ there exists constant $C(\theta_0, \epsilon) > 0$ such that $\inf_{|\theta-\theta_0|>\epsilon}(\theta_0-\theta)^2/(\theta^2+2|\theta|) \geq C(\theta_0,\epsilon)$. And the last term $O_p(\sqrt{N}||\mathbf{b}_X|| \cdot |\theta_0-\theta|)$ is negligible compared to the first term when $|\theta-\theta_0| > \epsilon$. Let $C'(\theta_0,\epsilon) = \min(\epsilon^2, C(\theta_0,\epsilon)) > 0$, we have

$$\inf_{|\theta-\theta_0|>\epsilon} -2l_p(\theta) \geq (1+o_p(1))C'(\theta_0,\epsilon)\frac{N||\mathbf{b}_X||_2^2}{2c_\sigma} + m + O_p(\sqrt{m}),$$

and $-2l_p(\theta_0) = m + O_p(\sqrt{m})$. Therefore,

$$P\left(l_p(\theta_0) > \sup_{|\theta-\theta_0|>\epsilon} l_p(\theta)\right) = P\left(O_p(\sqrt{m}) \leq (1+o_p(1))C'(\theta_0,\epsilon)\frac{N||\mathbf{b}_X||_2^2}{2c_\sigma} + O_p(\sqrt{m})\right).$$

When $N \to \infty$ (and thus $m/(N^2||\mathbf{b}_X||_2^4) \to 0$), this probability converges to 1. □

Next, we study the asymptotic normality of $\hat{\theta}$. Define the profile score to be the derivative of the profile log-likelihood:

$$\psi(\theta) := l_p'(\theta) = \sum_{i=1}^{m} \frac{(\hat{\beta}_{Yi}-\theta\hat{\beta}_{Xi})(\hat{\beta}_{Xi}\sigma_{Yi}^2 + \hat{\beta}_{Yi}\sigma_{Xi}^2\theta - \rho\sigma_{Xi}\sigma_{Yi}(\hat{\beta}_{Xi}\theta+\hat{\beta}_{Yi}))}{(\sigma_{Yi}^2+\theta^2\sigma_{Xi}^2-2\rho\theta\sigma_{Xi}\sigma_{Yi})^2}. \tag{A.5}$$

The Taylor expansion of $\psi(\hat{\theta})$ around the truth θ_0 can be expressed as:

$$0 = \psi(\hat{\theta}) = \psi(\theta_0) + \psi'(\theta_0)(\hat{\theta}-\theta_0) + \frac{1}{2}\psi''(\tilde{\theta})(\hat{\theta}-\theta_0)^2,$$

where $\tilde{\theta}$ is between $\hat{\theta}$ and θ_0. Then we have

$$\sqrt{V^*}(\hat{\theta}-\theta_0)=\frac{-\psi(\theta_0)/\sqrt{V^*}}{\psi'(\theta_0)/V^*+(1/2)\psi''(\tilde{\theta})(\hat{\theta}-\theta_0)/V^*}, \tag{A.6}$$

where

$$V^*=\mathbb{E}[-\psi'(\theta_0)]=\sum_{i=1}^{m}\frac{b_{Xi}^2\sigma_{Yi}^2+b_{Yi}^2\sigma_{Xi}^2-2\rho b_{Xi}b_{Yi}\sigma_{Xi}\sigma_{Yi}}{(\sigma_{Yi}^2+\theta_0^2\sigma_{Xi}^2-2\rho\theta_0\sigma_{Xi}\sigma_{Yi})^2}. \tag{A.7}$$

The nominator of Eq (A.6) can be proved to converge in distribution to $\mathcal{N}(0,1)$, the first term in the denominator of Eq (A.6) can be proved to converge in probability to 1 and the second term in the denominator an be proved to be negligible given $\hat{\theta}$ is consistent. We first prove these three statements, and by Slutsky's Theorem, the following asymptotic normality of $\hat{\theta}$ can be established.

Lemma A.3

Under the assumptions in Lemma A.2, we have

$$\sqrt{V^*}(\hat{\theta}-\theta_0)\xrightarrow{d}\mathcal{N}(0,1),$$

where

$$V^*=\sum_{i=1}^{m}\frac{b_{Xi}^2\sigma_{Yi}^2+b_{Yi}^2\sigma_{Xi}^2-2\rho b_{Xi}b_{Yi}\sigma_{Xi}\sigma_{Yi}}{(\sigma_{Yi}^2+\theta_0^2\sigma_{Xi}^2-2\rho\theta_0\sigma_{Xi}\sigma_{Yi})^2}.$$

Proof. We first show $(1/\sqrt{V^*})\psi(\theta_0)\xrightarrow{d}\mathcal{N}(0,1)$.

We can rewrite Eq (A.5) to obtain

$$\begin{aligned}\psi(\theta_0) &= \sum_{i=1}^{m} \frac{(e_i - \theta_0\epsilon_i)b_{Xi}}{\sigma_{Yi}^2 + \theta_0^2\sigma_{Xi}^2 - 2\rho\theta_0\sigma_{Xi}\sigma_{Yi}} + \\ &\sum_{i=1}^{m} \frac{(e_i - \theta_0\epsilon_i)(\epsilon_i\sigma_{Yi}^2 + e_i\sigma_{Xi}^2\theta_0 - \rho\sigma_{Xi}\sigma_{Yi}(\epsilon_i\theta_0 + e_i))}{(\sigma_{Yi}^2 + \theta_0^2\sigma_{Xi}^2 - 2\rho\theta_0\sigma_{Xi}\sigma_{Yi})^2} \\ &:= \sum_{i=1}^{m} \psi_{1i} + \sum_{i=1}^{m} \psi_{2i}. \end{aligned} \tag{A.8}$$

Since $\mathbb{E}[e_i - \theta_0\epsilon_i] = 0$ and $\mathrm{Var}(e_i - \theta_0\epsilon_i) = \sigma_{Yi}^2 + \theta_0^2\sigma_{Xi}^2 - 2\rho\theta_0\sigma_{Xi}\sigma_{Yi}$, the first term on the right hand side ($\sum_{i=1}^{m}\psi_{1i}$) is distributed as $\mathcal{N}(0, V^*)$ and $V^* = \Theta(N||\mathbf{b}_X||_2^2)$. The second term is negligible compared to the first term since it has variance $O(m)$. Therefore, $(1/\sqrt{V^*})\psi(\theta_0) \to \mathcal{N}(0,1)$.

We next show $(-1/V^*)\psi'(\theta_0) \xrightarrow{p} 1$.

Since $\mathbb{E}[-\psi'(\theta_0)] = V^*$ by definition, it suffices to show $\mathrm{Var}(\phi'(\theta_0)/V^*) \to 0$. With the observation that the $i-$th summand in Eq (A.5), $\psi_i(\theta)$, is a homogeneous quadratic polynomial of $(\tilde{b}_{Xi}, \tilde{e}_i, \tilde{\epsilon}_i) = (\sqrt{N}b_{Xi}, \sqrt{N}e_i, \sqrt{N}\epsilon_i)$:

$$\psi_i(\theta) = \frac{A_i}{B_i},$$

where

$$\begin{aligned} A_i = \Big\{ &((\theta_0 - \theta)\tilde{b}_{Xi} + \tilde{e}_i - \theta\tilde{\epsilon}_i) \cdot \big[(N\sigma_{Yi}^2 + N\theta_0\theta\sigma_{Xi}^2 - N\rho(\theta + \theta_0)\sigma_{Xi}\sigma_{Yi})\tilde{b}_{Xi} \\ &+ (N\sigma_{Yi}^2 - N\rho\sigma_{Xi}\sigma_{Yi}\theta)\tilde{\epsilon}_i + (N\sigma_{Xi}^2\theta - N\rho\sigma_{Xi}\sigma_{Yi})\tilde{e}_i\big]\Big\}, \\ B_i = &(N\sigma_{Yi}^2 + N\theta^2\sigma_{Xi}^2 - 2N\rho\theta\sigma_{Xi}\sigma_{Yi})^2, \end{aligned}$$

it is easy to see that $\psi_i'(\theta)$ is also a homogeneous quadratic polynomial of $(\tilde{b}_{Xi}, \tilde{e}_i, \tilde{\epsilon}_i)$. Also note that $\tilde{b}_{Xi}$ is treated as fixed and other terms such as $\mathrm{Var}(\tilde{e}_i)$, $\mathrm{Var}(\tilde{\epsilon}_i)$, $\mathrm{Var}(\tilde{e}_i\tilde{\epsilon}_i)$, $\mathrm{Cov}(\tilde{e}_i, \tilde{\epsilon}_i)$ are all $O(1)$, we have $\mathrm{Var}(\psi'(\theta_0)) = O(N||\mathbf{b}_X||_2^2) \ll V^{*2} =$

$\Theta(N^2||\mathbf{b}_X||^4)$.

Since $\psi_i'(\theta)$ is a homogeneous quadratic polynomial of $(\tilde{b}_{Xi}, \tilde{e}_i, \tilde{\epsilon}_i)$, so is $\psi_i''(\theta)$, and we can apply the same argument for $\psi'(\theta_0)$ above to $\psi''(\theta)$ and obtain that for a neighborhood $\mathcal{N}$ of θ_0, $\sup_{\theta\in\mathcal{N}}(1/V^*)\psi''(\theta) = O_p(1)$.

Lastly by Slutsky's Theorem, we have $\sqrt{V^*}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, 1)$. □

With Lemmas 1-3, we can conclude Theorem 1.

So far, coupled with the selection consistency, we have established the estimation consistency and asymptotic normality of the cMLE $\hat{\theta}(\hat{K})$ for the new extended MR-cML-BIC-C method that accounts for sample overlap.

For the Graph-MRcML with T traits of interest, we first apply MR-cML-BIC-C on every pair of traits to estimate the total causal effect graph $\widehat{\mathbf{G}}_{tot}$, then estimate the direct causal effect graph as $\widehat{\mathbf{G}}_{dir} = \widehat{\mathbf{G}}_{tot}(\mathbf{I} + \widehat{\mathbf{G}}_{tot})^{-1}$. Now, we are going to prove Corollary 1.

Corollary A.2

With Assumptions 1-4 satisfied, and if the spectral radius (i.e. the largest absolute value of the eigenvalues) of $\mathbf{G}_{dir}$ is smaller than 1, then vec($\widehat{\mathbf{G}}_{tot}$) and vec($\widehat{\mathbf{G}}_{dir}$) are consistent for vec($\mathbf{G}_{tot}$) and vec($\mathbf{G}_{dir}$) respectively, and are asymptotically normally distributed.

Proof. First, each element in vec($\widehat{\mathbf{G}}_{tot}$) is consistent and asymptotic normal by Theorem 1. Next, since matrix inversion and matrix multiplication are continuous transformations (Gentle, 2007), vec($\widehat{\mathbf{G}}_{dir}$) is consistent by the Continuous Mapping Theorem (Boos and Stefanski, 2013). For the asymptotic normality, we consider a $T \times (T-1)$ vector of the profile scores $\boldsymbol{\psi}$, where each element corresponds to the cMLE for an MR analysis on a pair of traits. Using the representation of Eq (A.8), we can write $\boldsymbol{\psi} = \boldsymbol{\psi}_1 + \boldsymbol{\psi}_2$, where the second term of is ignorable compared to the first term.

Further denote $\mathbf{E} = (e_{it})$, where $e_{it} = \hat{\beta}_{it} - b_{it}$, $\hat{\beta}_{it}$ is the GWAS estimated effect of the i-th SNP on the t-th trait, and $\mathbb{E}[\hat{\beta}_{it}] = b_{it}$. Now, with Assumption 4, $\boldsymbol{\psi}_1$ is normally distributed. And we can follow the same proofs for Lemmas 1 to 3 to conclude that $\text{vec}(\widehat{\mathbf{G}}_{tot})$ is jointly asymptotic normal. Finally, since matrix inversion and matrix multiplication are continuous and differentiable (Gentle, 2007), $\text{vec}(\widehat{\mathbf{G}}_{dir})$ is asymptotic normal by the Delta method (Boos and Stefanski, 2013). □

A.1.3 Standard error of the causal parameter estimate

Following Section S2 in the Supplementary in Xue et al. (2021), we can similarly show that the proposed cMLE and the MPLE in Lemma A.3 asymptotically share the same variance. Suppose we've already selected the set of invalid IVs and for the simplicity of notation, we assume that the m IVs are valid and used in the estimation and inference for $\hat{\theta}$. Then the $(m+1) \times (m+1)$ Fisher information matrix is

$$\mathcal{I} = -\begin{pmatrix} \frac{\partial^2 l}{\partial \theta^2} & \frac{\partial^2 l}{\partial \theta \partial \mathbf{b}_X'} \\ \frac{\partial^2 l}{\partial \theta \partial \mathbf{b}_X} & \frac{\partial^2 l}{\partial \mathbf{b}_X \mathbf{b}_X'} \end{pmatrix}, \tag{A.9}$$

where $\mathbf{b}_X = (b_{X1}, \ldots, b_{Xm})$. The second derivatives of log-likelihood are given as follows:

$$\begin{aligned} \frac{\partial^2 l}{\partial \theta^2} &= -\frac{1}{1-\rho^2} \sum_{i=1}^{m} \frac{b_{Xi}^2}{\sigma_{Yi}^2}, \\ \frac{\partial^2 l}{\partial \theta \partial b_{Xi}} &= -\frac{1}{1-\rho^2} \Big(\frac{\rho \hat{\beta}_{Xi} - 2\rho b_{Xi}}{\sigma_{Xi}\sigma_{Yi}} - \frac{\hat{\beta}_{Yi} - 2\theta b_{Xi}}{\sigma_{Yi}^2} \Big), \\ \frac{\partial^2 l}{\partial b_{Xi}^2} &= -\frac{1}{1-\rho^2} \Big(\frac{1}{\sigma_{Xi}^2} - \frac{2\rho\theta}{\sigma_{Xi}\sigma_{Yi}} + \frac{\theta^2}{\sigma_{Yi}^2} \Big). \end{aligned}$$

And we plug the cMLE $\hat{\theta}$ and $\hat{b}_{Xi}$ into above formulas and the variance estimator of the cMLE $\hat{\theta}$ is given as $\widehat{\text{Var}(\hat{\theta})} = (\mathcal{I}^{-1})_{11}$. Using the formula for block matrix

inversion, we have $\widehat{\mathrm{Var}(\hat{\theta})} = 1/\hat{V}$ and

$$\hat{V} = \frac{1}{1-\rho^2}\sum_i \left\{ \frac{\hat{b}_{Xi}^2}{\sigma_{Yi}^2} - \left(\frac{\rho\hat{\beta}_{Xi} - 2\rho\hat{b}_{Xi}}{\sigma_{Xi}\sigma_{Yi}} - \frac{\hat{\beta}_{Yi} - 2\hat{\theta}\hat{b}_{Xi}}{\sigma_{Yi}^2} \right)^2 \cdot \frac{1}{\frac{1}{\sigma_{Xi}^2} - \frac{2\rho\hat{\theta}}{\sigma_{Xi}\sigma_{Yi}} + \frac{\hat{\theta}^2}{\sigma_{Yi}^2}} \right\},$$

$$\hat{b}_{Xi} = (\hat{\beta}_{Xi}\sigma_{Yi}^2 - \rho\sigma_{Xi}\sigma_{Yi}(\hat{\beta}_{Yi} + \hat{\theta}\hat{\beta}_{Xi}) + \hat{\theta}\hat{\beta}_{Yi}\sigma_{Xi}^2)/(\sigma_{Yi}^2 - 2\rho\sigma_{Xi}\sigma_{Yi}\hat{\theta} + \hat{\theta}^2\sigma_{Xi}^2).$$

Similar to Section S2 in Xue et al. (2021), we have asymptotic variance of the MPLE $\hat{\theta}$ in Lemma A.3 as $1/\widehat{V^*}$ and

$$\widehat{V^*} = \sum_{i=1}^m \frac{\hat{\beta}_{Xi}^2\sigma_{Yi}^2 + \hat{\beta}_{Yi}^2\sigma_{Xi}^2 - 2\rho\hat{\beta}_{Xi}\hat{\beta}_{Yi}\sigma_{Xi}\sigma_{Yi}}{(\sigma_{Yi}^2 + \hat{\theta}^2\sigma_{Xi}^2 - 2\rho\hat{\theta}\sigma_{Xi}\sigma_{Yi})^2}.$$

Denote $c_i = \rho\sigma_{Xi}\sigma_{Yi}$, $d_i = \hat{\beta}_{Yi} - \hat{\theta}\hat{\beta}_{Xi}$ and $D = \sigma_{Yi}^2 + \hat{\theta}^2\sigma_{Xi}^2 - 2\rho\hat{\theta}\sigma_{Xi}\sigma_{Yi}$. Then we have

$$\widehat{V^*} = \frac{1}{1-\rho^2}\sum_{i=1}^m \frac{(\hat{\beta}_{Xi}^2\sigma_{Yi}^2 + \hat{\beta}_{Yi}^2\sigma_{Xi}^2 - 2c_i\hat{\beta}_{Xi}\hat{\beta}_{Yi}) \cdot (\sigma_{Xi}^2\sigma_{Yi}^2 - c_i^2) \cdot D}{D^3(\sigma_{Xi}^2\sigma_{Yi}^2)}, \tag{A.10}$$

and

$$\begin{aligned}
\hat{V} - \widehat{V^*} = \frac{1}{1-\rho^2}\sum_i \Big\{ & \frac{\sigma_{Xi}^2\sigma_{Yi}^2(2\sigma_{Xi}^2\hat{\beta}_{Yi}(\hat{\theta}^2\sigma_{Xi}^2 + \sigma_{Yi}^2)d_i - 4\sigma_{Xi}^2\sigma_{Yi}^2 d_i^2)}{D^3(\sigma_{Xi}^2\sigma_{Yi}^2)} && \text{(A.11)} \\
& - \frac{2c_i\sigma_{Xi}^2\sigma_{Yi}^2(\hat{\beta}_{Xi}\sigma_{Yi}^2 d_i + 2\hat{\theta}\sigma_{Xi}^2 d_i^2 + 3\hat{\theta}^2\hat{\beta}_{Xi}\sigma_{Xi}^2 d_i)}{D^3(\sigma_{Xi}^2\sigma_{Yi}^2)} && \text{(A.12)} \\
& + \frac{2c_i^2(2\sigma_{Xi}^2\sigma_{Yi}^2 d_i^2 + \hat{\beta}_{Yi}\sigma_{Xi}^2(\sigma_{Yi}^2 - \hat{\theta}^2\sigma_{Xi}^2)d_i)}{D^3(\sigma_{Xi}^2\sigma_{Yi}^2)} && \text{(A.13)} \\
& + \frac{2c_i^3(\hat{\beta}_{Xi}\sigma_{Yi}^2 d_i + 2\hat{\theta}\sigma_{Xi}^2 d_i^2 + 3\hat{\theta}^2\hat{\beta}_{Xi}\sigma_{Xi}^2 d_i)}{D^3(\sigma_{Xi}^2\sigma_{Yi}^2)} && \text{(A.14)} \\
& - \frac{4c_i^4\hat{\beta}_{Yi}d_i}{D^3(\sigma_{Xi}^2\sigma_{Yi}^2)} \Big\}. && \text{(A.15)}
\end{aligned}$$

Divided each part in Eq (A.11) to Eq (A.15) by the corresponding part in Eq (A.10),

we obtain $\hat{V} - \widehat{V^*} = o_p(\widehat{V^*})$, since c_i, σ^2_{Xi} and σ^2_{Yi} are all $O_p(1/N)$, and $d_i = o_p(1)$ by the consistency of $\hat{\theta}$. Thus, our cMLE and the MPLE asymptotically share the same variance. This is also confirmed in numerical results as shown in Appendix A.3.2 and Appendix A.3.3.

A.1.4 Asymptotic properties of MR-cML-BIC-I in the presence of sample overlap

A.1.4.1 Selection consistency

It is noted that the selection consistency of MR-cML-BIC-I provided in Xue et al. (2021) would still hold despite the correlation between the two GWAS summary data. This can be seen from the proof in Xue et al. (2021) that most of the arguments still hold except for some distributional arguments. For example, $\frac{\hat{\beta}_{Yi} - \tilde{\theta}\cdot\hat{\beta}_{Xi}}{\sqrt{\sigma^2_{Yi} + \tilde{\theta}^2\sigma^2_{Xi}}}$ is no longer $\mathcal{N}(\frac{\theta\cdot b_{Xi} + r_i - \tilde{\theta}\cdot b_{Xi}}{\sqrt{\sigma^2_{Yi} + \tilde{\theta}^2\sigma^2_{Xi}}}, 1)$, but we can multiply it by a strictly positive scalar $s_i = \frac{\sqrt{\sigma^2_{Yi} + \tilde{\theta}^2\sigma^2_{Xi}}}{\sqrt{\sigma^2_{Yi} + \tilde{\theta}^2\sigma^2_{Xi} - 2\rho\tilde{\theta}\sigma_{Xi}\sigma_{Yi}}}$. Notice that $s_i = O_p(1)$ and all the inequalities still hold asymptotically. Hence MR-cML-BIC-I, which use the mis-specified likelihood with $\rho = 0$ in the presence of sample overlap, is still able to select the correct set of invalid IVs asymptotically. We perform simulations to confirm this.

To avoid generating a large number of individual data, we simulate the GWAS summary statistics directly as follows:

$$b_{Xi} = \gamma_i + \beta_{XU}\phi_i, \tag{A.16}$$

$$b_{Yi} = \theta b_{Xi} + \beta_{YU}\phi_i + \alpha_i, \tag{A.17}$$

$$\begin{pmatrix} \hat{\beta}_{Xi} \\ \hat{\beta}_{Yi} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} b_{Xi} \\ b_{Yi} \end{pmatrix}, \begin{pmatrix} \sigma^2_{Xi} & \rho\sigma_{Xi}\sigma_{Yi} \\ \rho\sigma_{Xi}\sigma_{Yi} & \sigma^2_{Yi} \end{pmatrix} \right) iid, \quad i = 1, \ldots, m. \tag{A.18}$$

We set $\gamma_i = 0.08$, $\beta_{XU} = \beta_{YU} = 1$, $\theta = 0.2$, $m = 100$. We considered 30% invalid IVs with the direct effect α_i iid from $\mathcal{N}(0.04, 0.05^2)$, and $\phi_i = 0$ (uncorrelated pleiotropy),

or ϕ_i iid from Unif(0,0.01) (correlated pleiotropy). σ_{Xi} and σ_{Yi} were set to be $1/\sqrt{N}$, where the sample size $N_1 = N_2 = N$ varied from $\{25\,000, 100\,000, 500\,000, 1\,000\,000, 5\,000\,000\}$. ρ was set to be 0 or 0.8, and we used the true ρ in MR-cML-BIC-C. We ran 10 000 replications for each scenario and calculated the average true positive rate of correctly identifying invalid IVs, and the average accuracy of classifying IVs. True positive rate is calculated as #correctly identified invalid IVs/#invalid IVs. Accuracy is calculated as (#correctly identified invalid IVs + #correctly identified valid IVs)/#IVs.

First, when $\rho = 0$, MR-cML-BIC-C and MR-cML-BIC-I were the same and gave consistent estimate and selection as shown in the left column in Figure A.1 to Figure A.3. On the other hand, when $\rho = 0.8$, MR-cML-BIC-I gave biased estimate when the sample size was not large enough, while MR-cML-BIC-C was almost unbiased. It is also noted that MR-cML-BIC-C performed better than MR-cML-BIC-I in selecting invalid IVs, though the probabilities of selecting the correct set of invalid IVs for both methods were going to 1 as the sample size increased as shown in the right column in Figure A.2 and Figure A.3.

A.1.4.2 Estimation consistency and a robust variance estimator

Using the representation in Eq (A.8), we can see that even though we ignore the correlation between $\hat{\beta}_{Xi}$ and $\hat{\beta}_{Yi}$ (i.e. $\rho = 0$) and mis-specify the (profile) likelihood, we still have $\psi(\theta_0) = o_p(\sqrt{V^*})$. Thus, the MLE under the mis-specified profile log-likelihood in Zhao et al. (2020) is still consistent. But the usual (naive or model-based) variance estimator used in Xue et al. (2021); Zhao et al. (2020) is not. Instead, we could use the sandwich variance estimator that is robust to model mis-specification. The asymptotic normality of $\hat{\theta}_I$ in Theorem A.2 can be similarly derived as in Lemma A.3

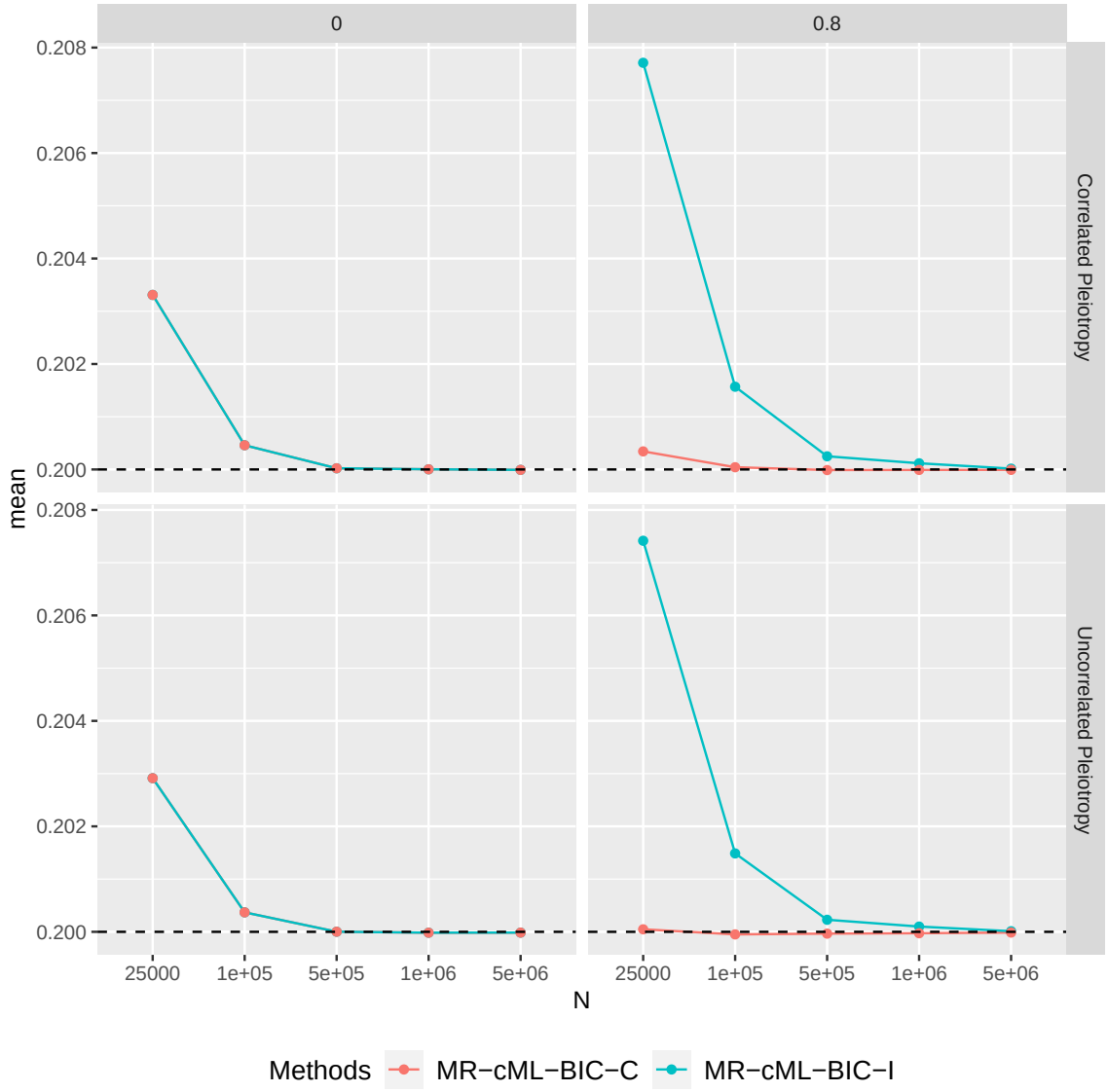


Figure A.1: The y-axis is the mean of $\hat{\theta}$ among 10 000 replications, and the x-axis represents different sample sizes. Dashed line is the true $\theta = 0.2$. Top: correlated pleiotropy. Bottom: uncorrelated pleiotropy. Left: $\rho = 0$. Right: $\rho = 0.8$

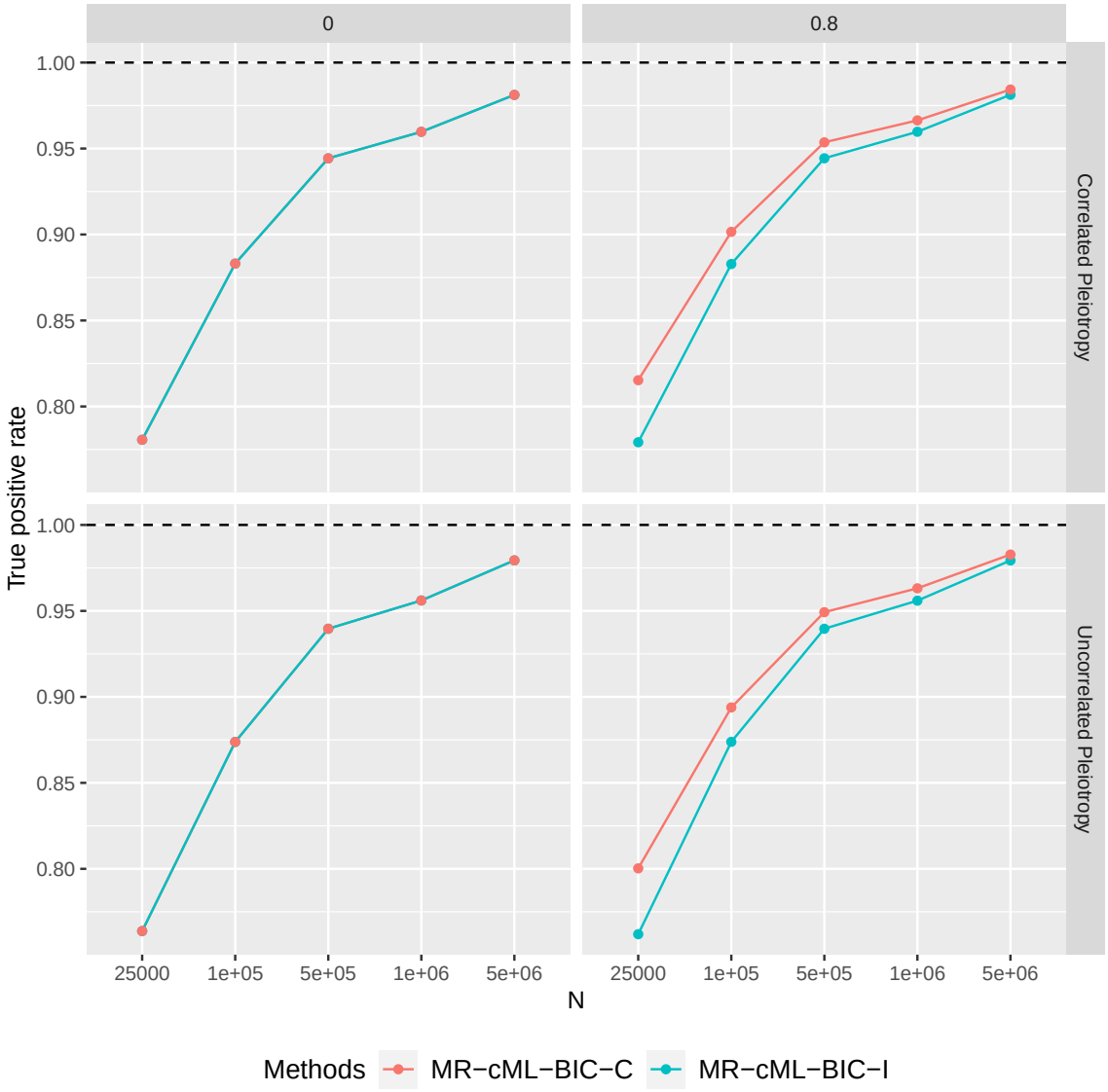


Figure A.2: The y-axis is the mean of true positive rates among 10 000 replications, and the x-axis represents different sample sizes. Top: correlated pleiotropy. Bottom: uncorrelated pleiotropy. Left: $\rho = 0$. Right: $\rho = 0.8$.

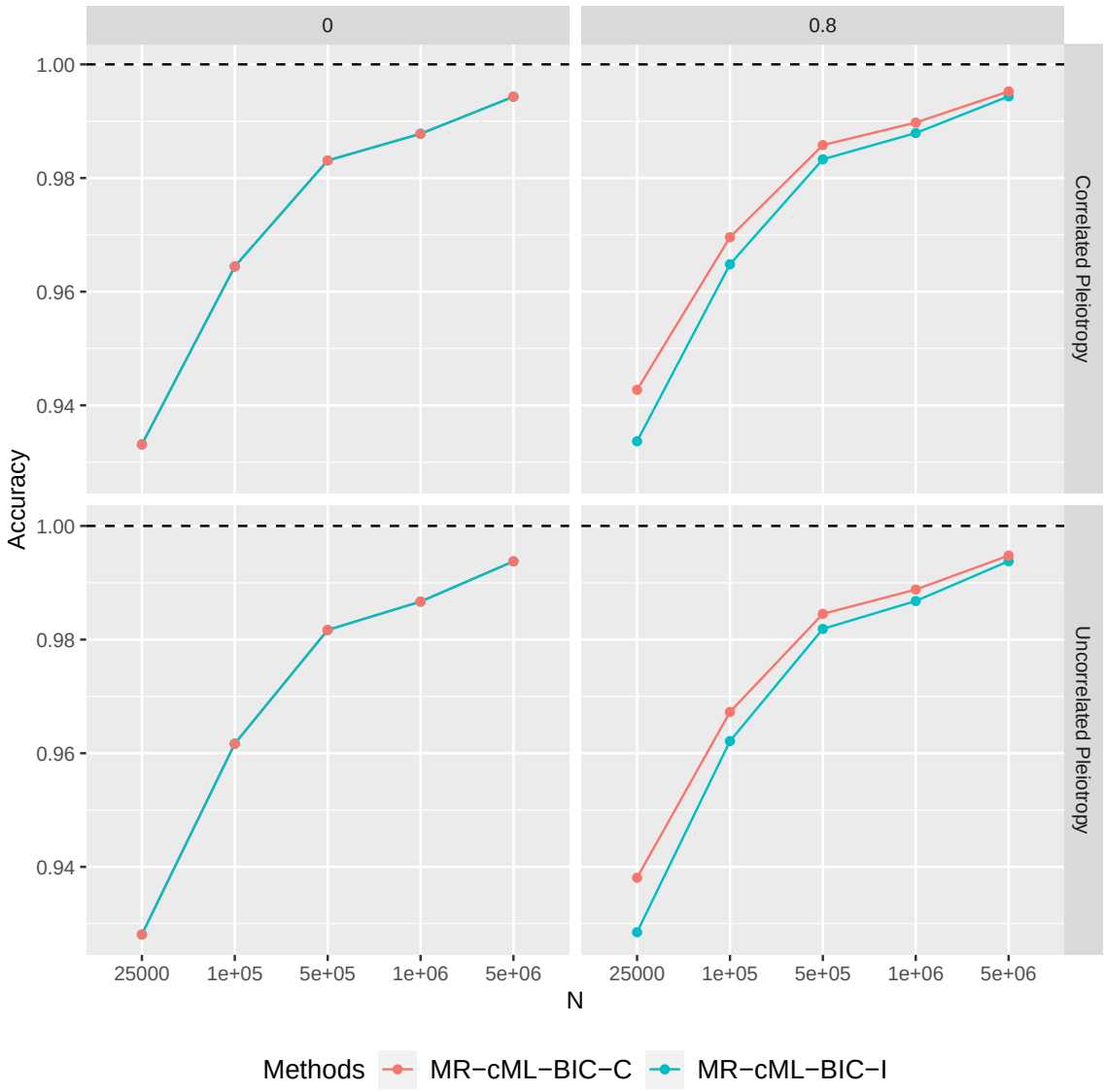


Figure A.3: The y-axis is the mean of accuracy among 10 000 replications, and the x-axis represents different sample sizes. Top: correlated pleiotropy. Bottom: uncorrelated pleiotropy. Left: $\rho = 0$. Right: $\rho = 0.8$.

with the following representation from Taylor expansion:

$$\sqrt{V_r}(\hat{\theta}_I - \theta_0) = \frac{-\psi(\theta_0; \rho = 0)/\sqrt{\mathbf{B}}}{\psi'(\theta_0; \rho = 0)/\mathbf{A} + (1/2)\psi''(\tilde{\theta}; \rho = 0)(\hat{\theta}_I - \theta_0)/\mathbf{A}}, \tag{A.19}$$

where $V_r = \mathbf{A}\mathbf{B}^{-1}\mathbf{A}, \mathbf{B} = \mathbb{E}[\psi^2(\theta_0; \rho = 0)], \mathbf{A} = \mathbb{E}[\psi'(\theta_0; \rho = 0)]$.

The 'bread' can be estimated empirically as

$$\begin{aligned}\widehat{\mathbf{A}} &= \widehat{\mathbb{E}[\psi'(\theta_0)]} = \sum_{i=1}^{m} \psi_i'(\hat{\theta}; \rho = 0) \\ &= \Big\{(\hat{\beta}_{Yi}^2\sigma_{Xi}^2 - \hat{\beta}_{Xi}^2\sigma_{Yi}^2 - 2\hat{\theta}\hat{\beta}_{Xi}\hat{\beta}_{Yi}\sigma_{Xi}^2)(\sigma_{Yi}^2 + \hat{\theta}^2\sigma_{Xi}^2)^2 - \\ &\quad 2(\sigma_{Yi}^2 + \hat{\theta}^2\sigma_{Xi}^2)(2\hat{\theta}\sigma_{Xi}^2)(\hat{\beta}_{Yi} - \hat{\theta}\hat{\beta}_{Xi})(\hat{\beta}_{Xi}\sigma_{Yi}^2 + \hat{\beta}_{Yi}\sigma_{Xi}^2)\Big\}/(\sigma_{Yi}^2 + \hat{\theta}^2\sigma_{Xi}^2)^4.\end{aligned} \tag{A.20}$$

And the 'meat' can be estimated empirically as

$$\begin{aligned}\widehat{\mathbf{B}} &= \widehat{\mathbb{E}[\psi^2(\theta_0)]} = \sum_{i=1}^{m} \psi_i^2(\hat{\theta}; \rho = 0) \\ &= \sum_{i=1}^{m} \left(\frac{(\hat{\beta}_{Yi} - \hat{\theta}\hat{\beta}_{Xi})(\hat{\beta}_{Xi}\sigma_{Yi}^2 + \hat{\beta}_{Yi}\sigma_{Xi}^2\hat{\theta})}{(\sigma_{Yi}^2 + \hat{\theta}^2\sigma_{Xi}^2)^2}\right)^2.\end{aligned} \tag{A.21}$$

Then we have the robust variance estimator $\hat{V}_{robust} = \widehat{\mathbf{A}}^{-1}\widehat{\mathbf{B}}(\widehat{\mathbf{A}}^{-1})^T$.

We further perform simulations to confirm this. We followed the similar simulation set-up as described in Appendix A.1.4.1 but under the ideal scenario with no invalid IV. We applied MR-cML-BIC-I and MR-cML-BIC-C with the oracle $K = 0$ (i.e. there is no model selection process and all IVs were valid), called **cML-BIC-I-K0** and **cML-BIC-C-K0** respectively. For MR-cML-BIC-I, we also calculated the robust variance estimate, referred to **cML-BIC-I-K0-robust**. We ran 10 000 replications for each scenario and compared the mean of the estimated standard errors (mean($SE(\hat{\theta})$)) and the empirical standard deviation of $\hat{\theta}$ ($SD(\hat{\theta})$) from the 10 000

replications.

Figure A.4 shows the simulation results under different scenarios for different methods. First, as shown in the first column in Figure A.4, when there was no sample overlap ($\rho = 0$), cML-BIC-I-K0 and cML-BIC-C-K0 performed similarly with unbiased estimates. When $\rho = 0.8$, cML-BIC-C-K0 (top) yielded unbiased estimates, while cML-BIC-I-K0 (middle) yielded biased estimates, but the bias was getting smaller when the sample size increased. This confirms that cML-BIC-I-K0 is still consistent. As for the variance estimate, we can see that when there was no sample overlap (first column), the two error bars representing $SD(\hat{\theta})$ and mean$(SE(\hat{\theta}))$ aligned with each other as the sample size increased for all methods. However, when $\rho = 0.8$, the two error bars aligned for cML-BIC-C-K0 and cML-BIC-I-K0-robust, but not for cML-BIC-I-K0. We can see that in the middle row, there were discrepancies between the two error bars even when the sample size kept increasing. On the other hand, cML-BIC-I-K0-robust (bottom) still gave correct variance estimates when we used the robust sandwich variance estimator.

Figure A.5 shows the empirical coverage rates for different methods. In the first column with $\rho = 0$, the red line (cML-BIC-C-K0) and the green line (cML-BIC-I-K0) coincided with each other, yielding correct coverage rates. When $\rho = 0.8$, only cML-BIC-C-K0 gave correct coverage rates close to 95%. When the sample size was small ($N = 25\,000$), cML-BIC-I-K0 seemed to yield correct coverage rates close to 95%. But this is because it yielded biased estimates and at the same time inflated variances as shown in Figure A.4. As the sample size increased, cML-BIC-I-K0 yielded more conservative coverage rates. This is because it became less biased but the naive variance estimator over-estimated the true variance as shown in Figure A.4. It is also noted that using the robust variance estimator, cML-BIC-I-K0-robust yielded anti-conservative empirical coverage rates (and always more anti-conservative than cML-BIC-I-K0), but it was getting closer to the nominal 95% as the sample size

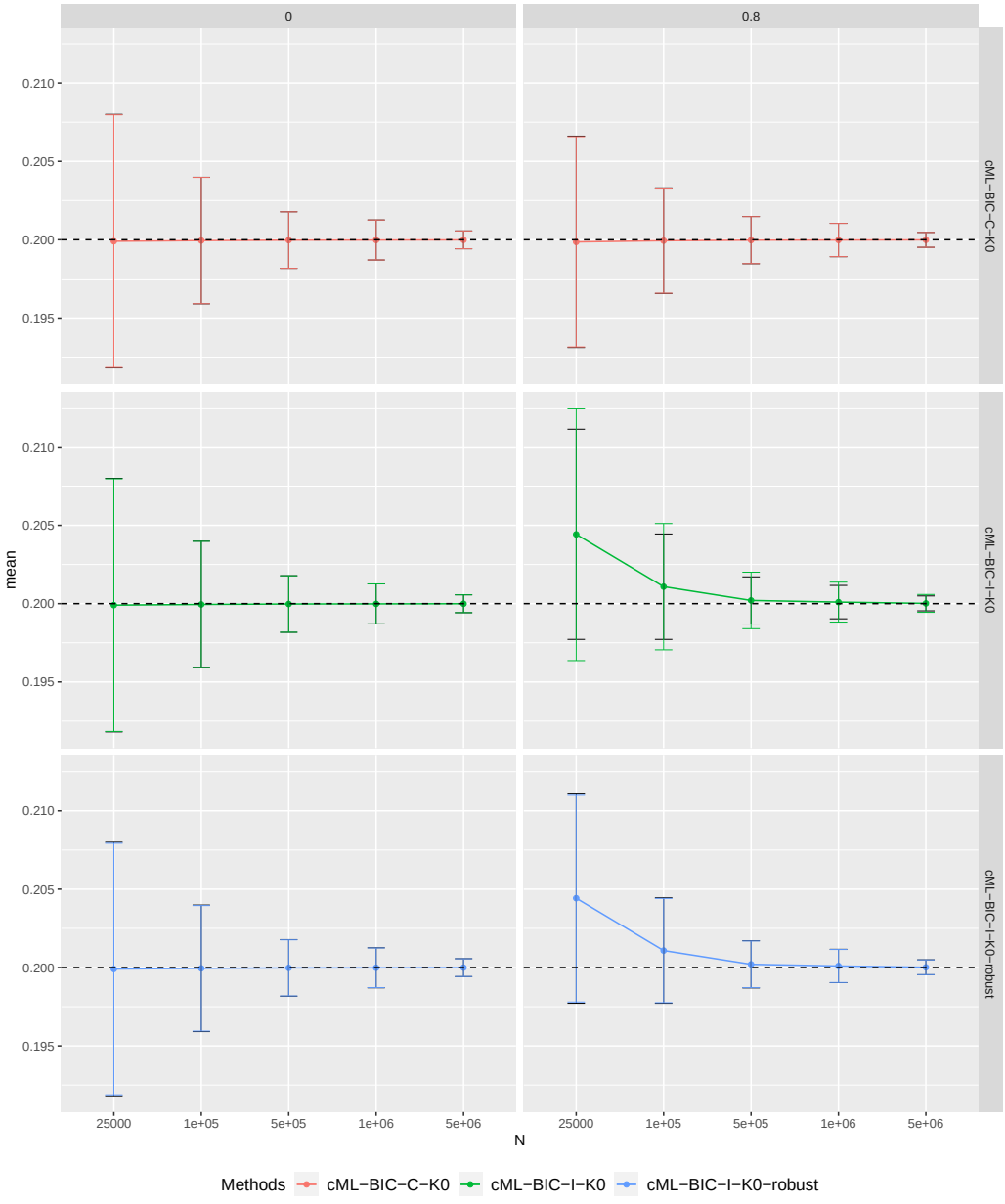


Figure A.4: The y-axis is the mean of $\hat{\theta}$ among 10 000 replications, and the x-axis represents different sample sizes. Dashed line is the true $\theta = 0.2$. The black error bar is $SD(\hat{\theta})$ and the colored error bar is mean$(SE(\hat{\theta}))$. Top row: cML-BIC-C-K0. Middle row: cML-BIC-I-K0. Bottom row: cML-BIC-I-K0-robust. Left: $\rho = 0$. Right: $\rho = 0.8$

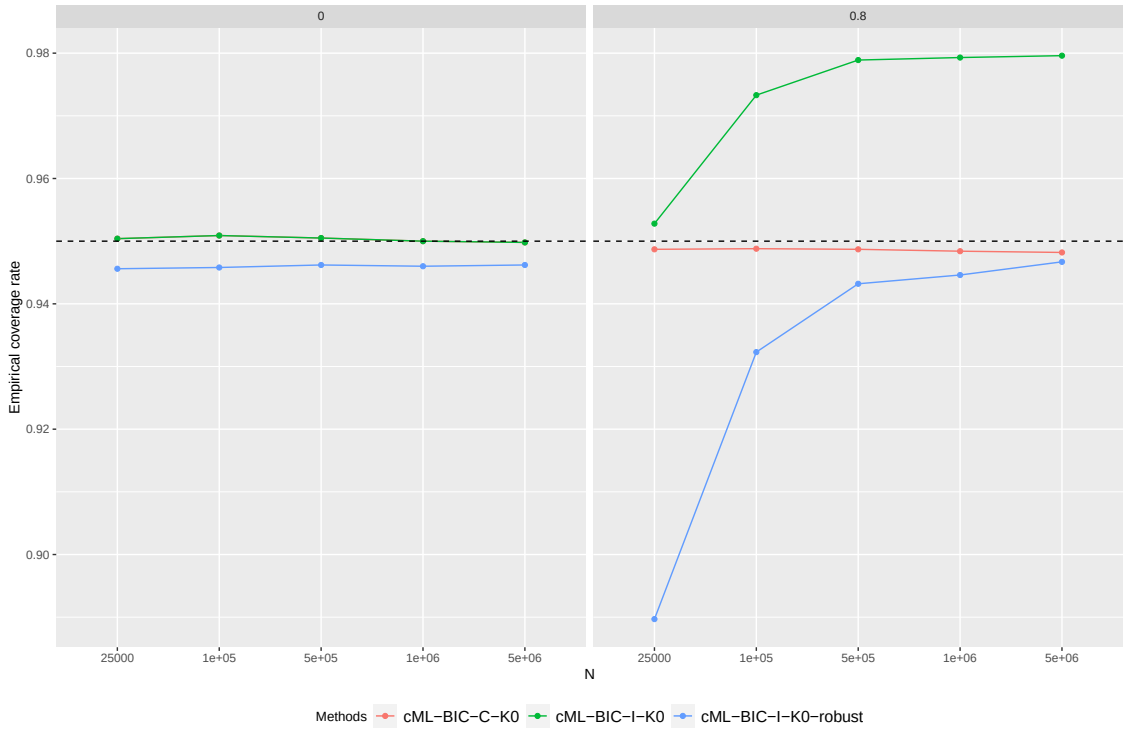


Figure A.5: The y-axis is the empirical coverage rate among 10 000 replications, and the x-axis represents different sample sizes. Dashed line is the nominal level 95%. Left: $\rho = 0$. Right: $\rho = 0.8$.

increased. This is because although it gave correct variance estimates, the estimates of θ was still biased unless the sample size was large enough (Figure A.4).

A.2 More simulation results for MR methods with sample overlap

Figure A.15A and Figure A.15B show the distributions of the causal estimates in the case of 100% sample overlap and 30% invalid IVs with uncorrelated and correlated pleiotropy respectively. We note that although all estimates were biased upward, MR-cML-C yielded much less biased estimates than MR-cML-I in both scenarios.

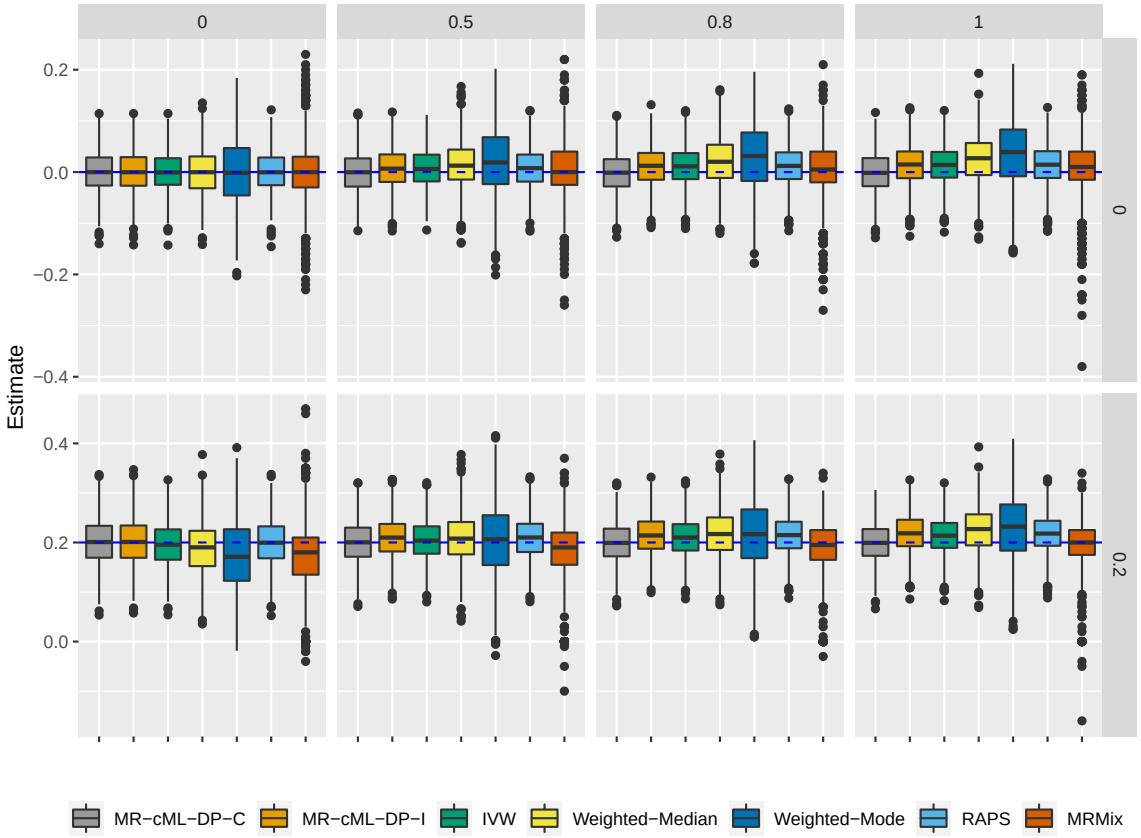


Figure A.6: Estimates of the causal effect θ with 0% invalid IVs across 1000 replicates. From left to right correspond to 0%, 50%, 80% and 100% overlapping samples. Top panel:$\theta = 0$ and bottom panel: $\theta = 0.2$.

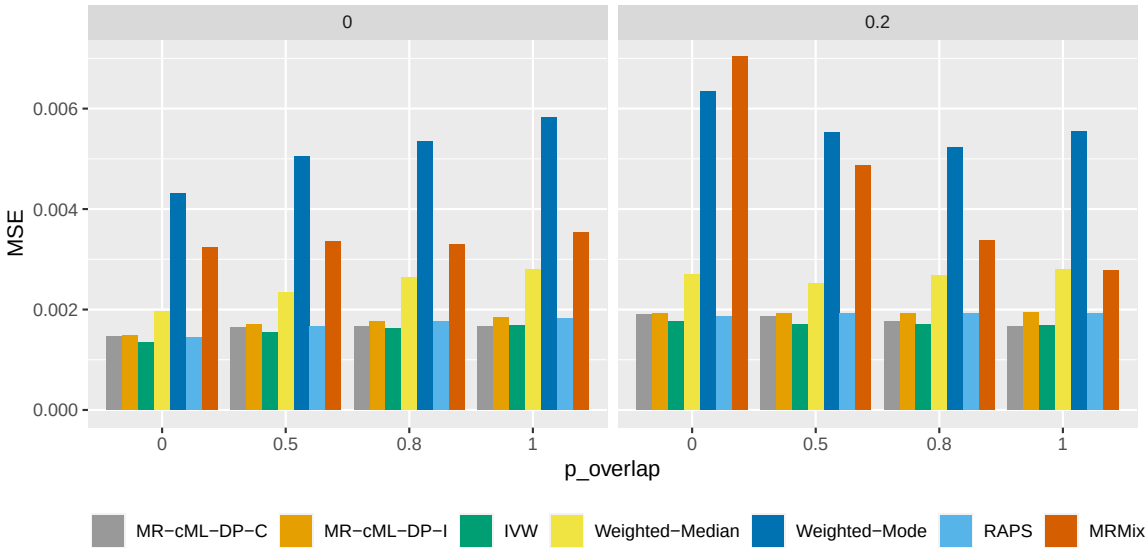


Figure A.7: Mean squared error (MSE) in the presence of 0% invalid IVs . X-axis represents different proportions of sample overlap (0%, 50%, 80% and 100%). Left panel:$\theta = 0$ and right panel: $\theta = 0.2$.

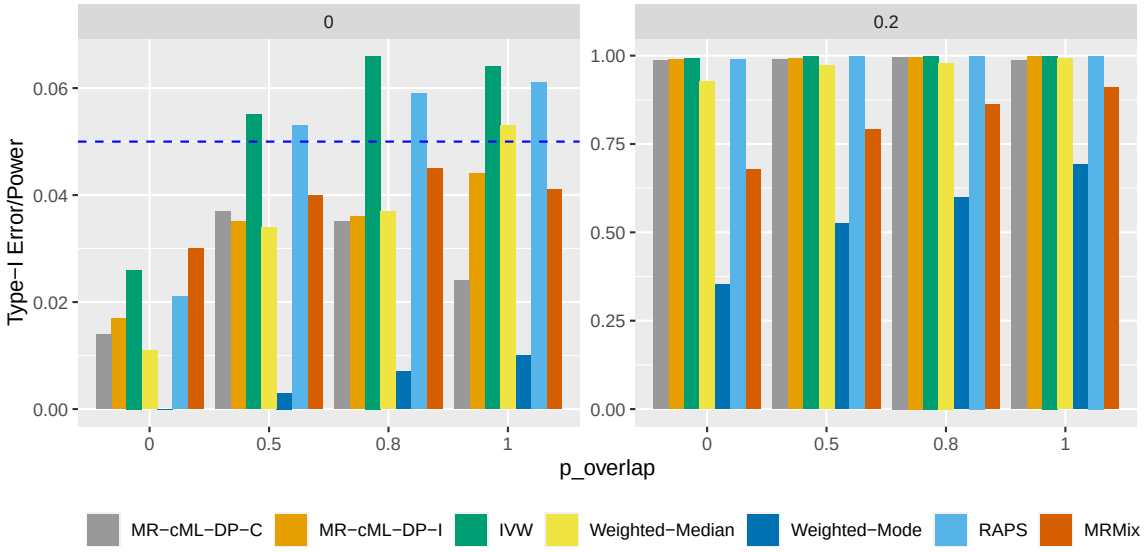


Figure A.8: Empirical type-I error and power in the presence of 0% invalid IVs . X-axis represents different proportions of sample overlap (0%, 50%, 80% and 100%). Left: $\theta = 0$ (type-I error) and right: $\theta = 0.2$ (power).

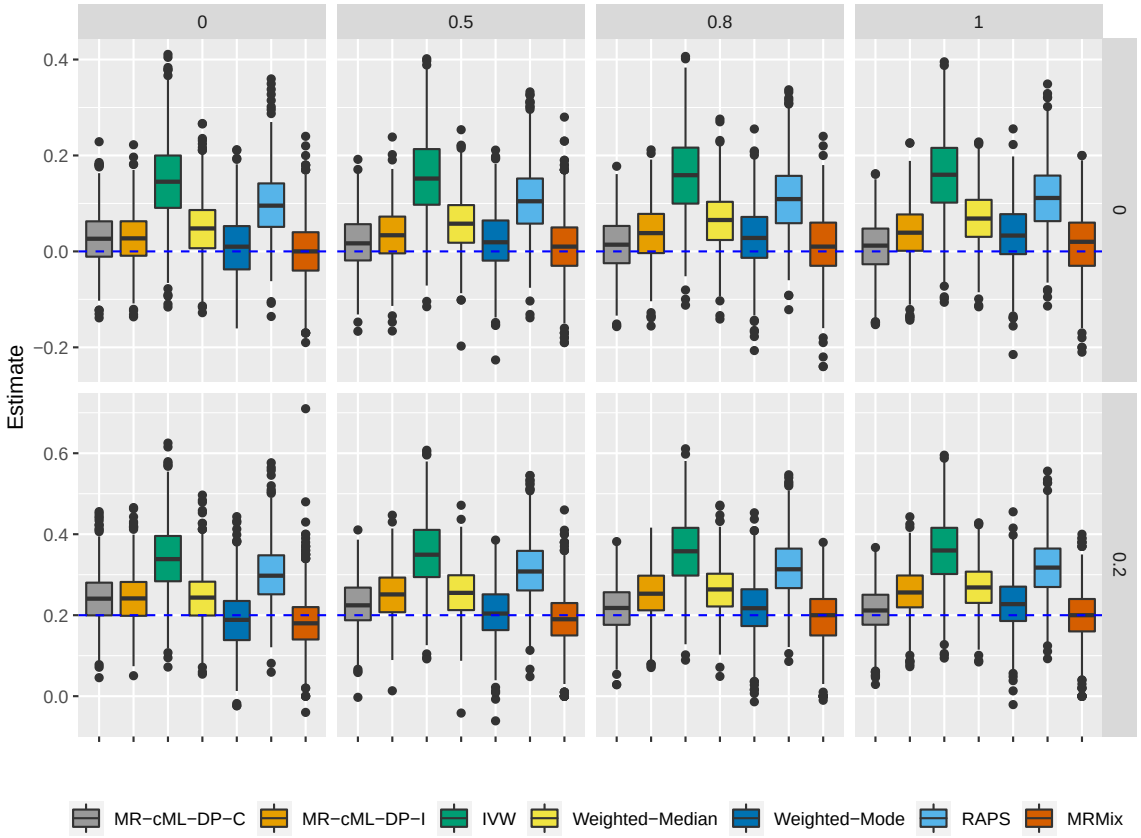


Figure A.9: Estimates of the causal effect θ with 30% invalid IVs (uncorrelated pleiotropy) across 1000 replicates. From left to right correspond to 0%, 50%, 80% and 100% overlapping samples. Top panel:$\theta = 0$ and bottom panel: $\theta = 0.2$.

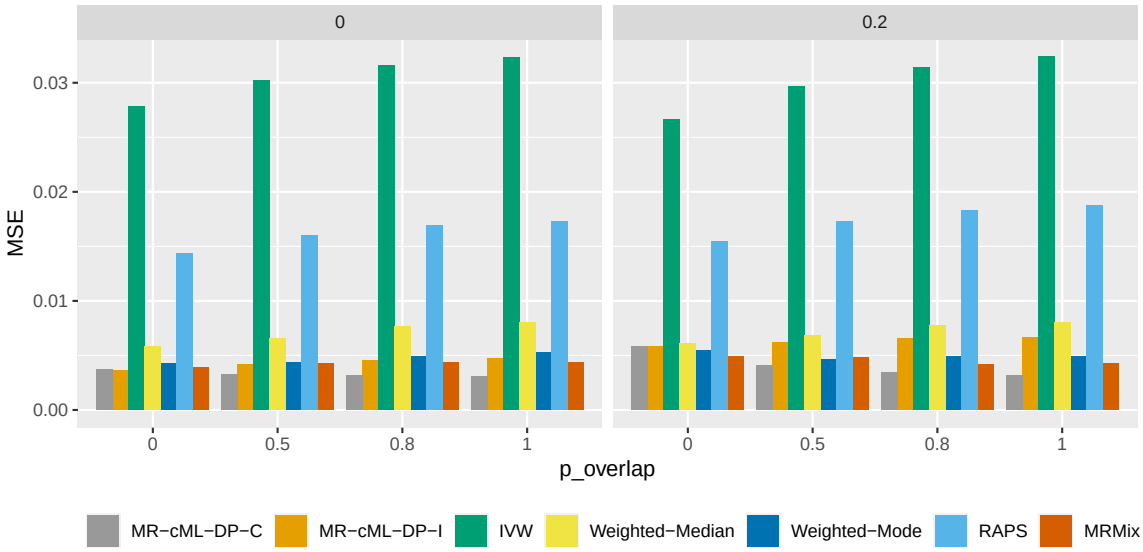


Figure A.10: Mean squared error (MSE) in the presence of 30% invalid IVs (uncorrelated pleiotropy). X-axis represents different proportions of sample overlap (0%, 50%, 80% and 100%). Left panel:$\theta = 0$ and right panel: $\theta = 0.2$.

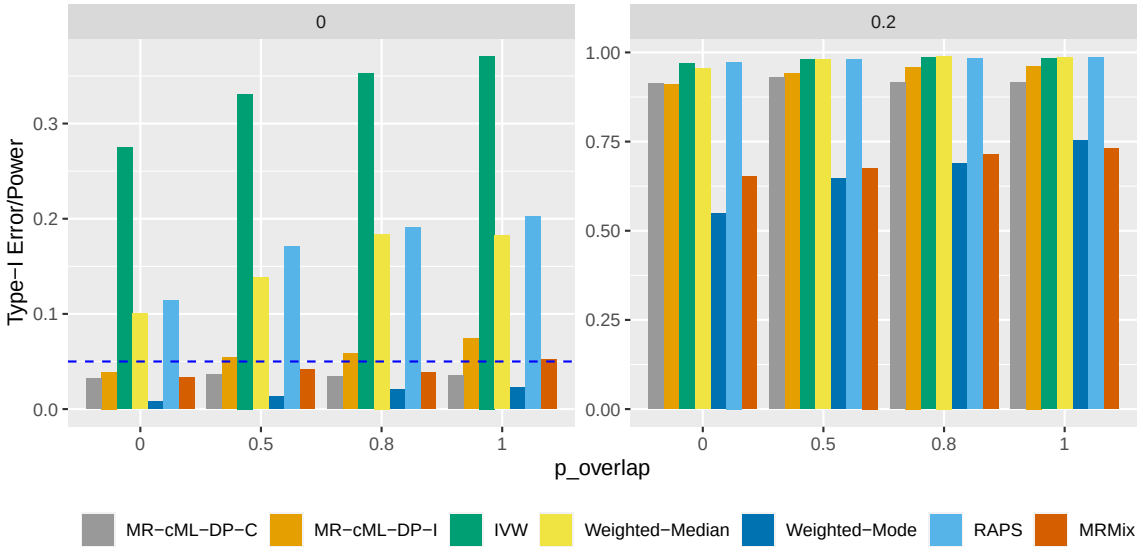


Figure A.11: Empirical type-I error and power in the presence of 30% invalid IVs with uncorrelated pleiotropy. X-axis represents different proportions of sample overlap (0%, 50%, 80% and 100%). Left: $\theta = 0$ (type-I error) and right: $\theta = 0.2$ (power).

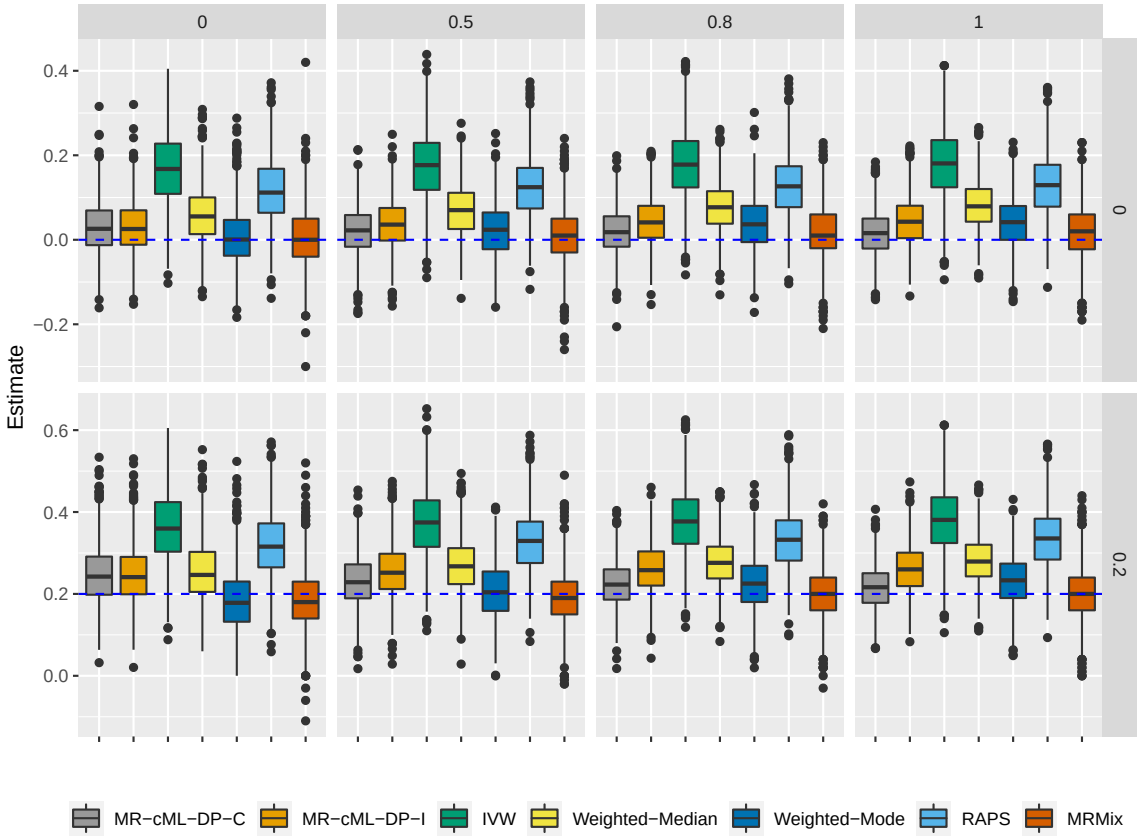


Figure A.12: Estimates of the causal effect θ with 30% invalid IVs (correlated pleiotropy) across 1000 replicates. From left to right correspond to 0%, 50%, 80% and 100% overlapping samples. Top panel:$\theta = 0$ and bottom panel: $\theta = 0.2$.

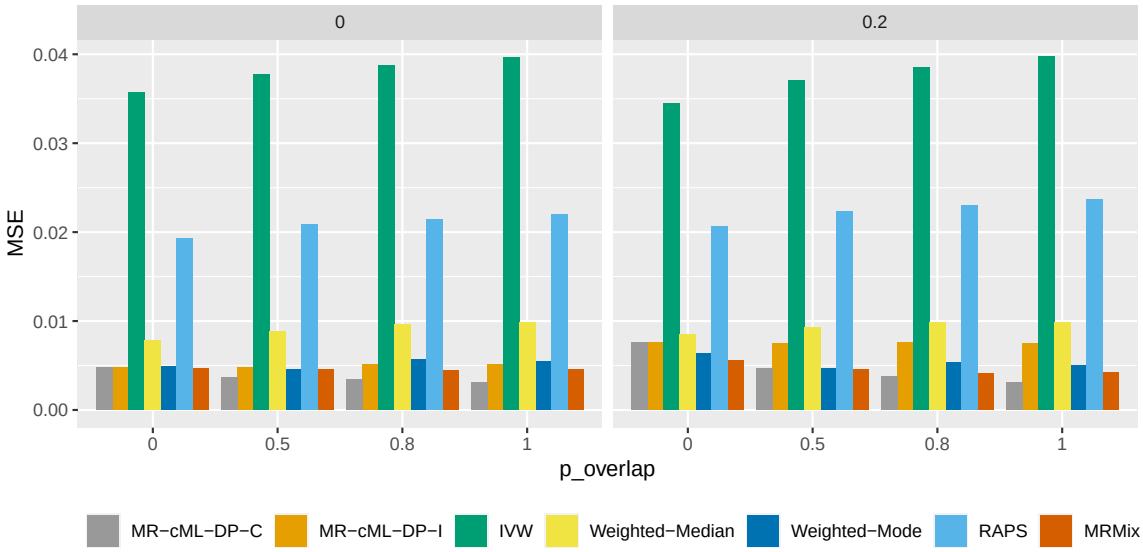


Figure A.13: Mean squared error (MSE) in the presence of 30% invalid IVs (correlated pleiotropy). X-axis represents different proportions of sample overlap (0%, 50%, 80% and 100%). Left panel:$\theta = 0$ and right panel: $\theta = 0.2$.

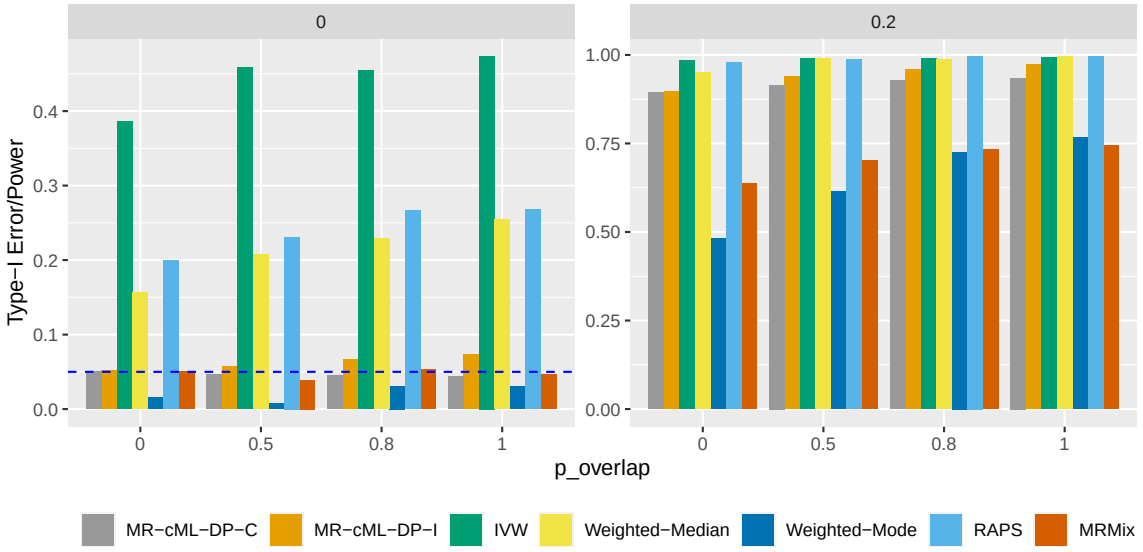


Figure A.14: Empirical type-I error and power in the presence of 30% invalid IVs with correlated pleiotropy. X-axis represents different proportions of sample overlap (0%, 50%, 80% and 100%). Left: $\theta = 0$ (type-I error) and right: $\theta = 0.2$ (power).

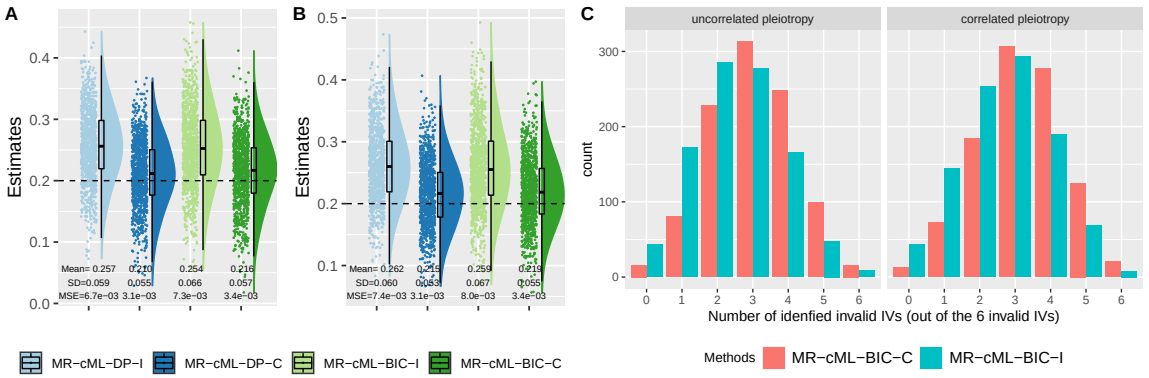


Figure A.15: Simulation results with 30% invalid IVs and 100% sample overlap. Panel A: empirical distributions of the estimates of the causal effect $\theta = 0.2$ with uncorrelated pleiotropy. Panel B: empirical distributions of the estimates of the causal effect $\theta = 0.2$ with correlated pleiotropy.Panel C: empirical frequency distributions of the numbers of identified invalid IVs (out of a total of 6) by BIC from 1000 replications (left: with uncorrelated pleiotropy; right: with correlated pleiotropy).

One possible reasons is that MR-cML-C performed better in identifying invalid IVs than MR-cML-I in the presence of sample overlap, while both methods sometimes might fail to identify all invalid IVs (perhaps due to the small sample size and/or small effects of some invalid IVs). Figure A.15C confirmed this point by showing the frequencies (from 1000 independent replications) of the numbers of the correctly identified invalid IVs (out of $20 \times 0.3 = 6$) using MR-cML-BIC-C and MR-cML-BIC-I respectively. We can clearly see that MR-cML-BIC-C tended to identify more of the true invalid IVs than MR-cML-BIC-I.

A.3 More simulation results on the different versions of MR-cML

In this section, we omit the prefix 'MR-' in all methods for a clearer presentation. For example, MR-cML-BIC-I becomes cML-BIC-I, and MR-cML-DP-I becomes cML-

DP-I.

A.3.1 Detailed results of simulations in the main text

Table A.1: In each cell, from top to bottom are empirical type-I error/power, mean($\hat{\theta}$), SD($\hat{\theta}$), mean(SE($\hat{\theta}$)), coverage rate, MSE, when $N = 25000$ and all valid IVs.

p_overlap	θ	$\hat{\rho}$	cML-BIC-I	cML-DP-I	cML-MA-I	cML-MA-DP-I	cML-BIC-C	cML-DP-C	cML-MA-C	cML-MA-DP-C
0.0	0.0	-0.001	0.034	0.017	0.033	0.020	0.036	0.014	0.036	0.015
			0.001	0.000	0.000	0.000	0.001	0.000	0.000	0.000
			0.038	0.039	0.038	0.039	0.038	0.038	0.038	0.038
			0.039	0.047	0.039	0.046	0.039	0.047	0.039	0.046
			0.966	0.983	0.967	0.980	0.964	0.986	0.964	0.985
			0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	0.2	0.000	0.995	0.989	0.994	0.991	0.995	0.986	0.994	0.988
			0.201	0.201	0.201	0.201	0.201	0.201	0.201	0.201
			0.043	0.044	0.043	0.044	0.043	0.044	0.043	0.043
			0.044	0.053	0.044	0.052	0.044	0.053	0.044	0.052
			0.966	0.985	0.969	0.981	0.967	0.982	0.970	0.981
			0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
0.5	0.0	0.247	0.065	0.035	0.062	0.041	0.063	0.037	0.062	0.040
			0.008	0.008	0.008	0.008	0.001	0.000	0.001	0.000
			0.040	0.040	0.040	0.040	0.040	0.041	0.040	0.041
			0.039	0.047	0.039	0.046	0.039	0.047	0.039	0.046
			0.935	0.965	0.938	0.959	0.937	0.963	0.938	0.960
			0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
	0.2	0.312	0.999	0.992	0.999	0.996	0.997	0.989	0.998	0.990
			0.210	0.210	0.210	0.210	0.201	0.201	0.201	0.201
			0.042	0.043	0.042	0.043	0.043	0.043	0.043	0.043
			0.044	0.052	0.044	0.051	0.041	0.050	0.042	0.049
			0.948	0.970	0.948	0.969	0.939	0.972	0.943	0.967
			0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
0.8	0.0	0.395	0.081	0.036	0.080	0.042	0.068	0.035	0.061	0.040
			0.011	0.011	0.011	0.011	0.000	-0.001	0.000	-0.001
			0.040	0.040	0.040	0.040	0.040	0.041	0.040	0.041
			0.039	0.047	0.039	0.046	0.039	0.047	0.039	0.046
			0.919	0.964	0.920	0.958	0.932	0.965	0.939	0.960
			0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
	0.2	0.500	0.999	0.996	0.999	0.996	0.999	0.996	0.997	0.996
			0.215	0.215	0.215	0.215	0.200	0.200	0.200	0.199
			0.041	0.041	0.041	0.041	0.041	0.042	0.041	0.042
			0.044	0.051	0.044	0.050	0.040	0.048	0.040	0.047
			0.945	0.972	0.947	0.968	0.934	0.966	0.941	0.967
			0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
1.0	0.0	0.493	0.082	0.044	0.077	0.049	0.064	0.024	0.059	0.031
			0.014	0.014	0.014	0.014	0.000	-0.001	0.000	-0.001
			0.040	0.040	0.040	0.040	0.040	0.041	0.040	0.041
			0.039	0.047	0.039	0.046	0.039	0.047	0.039	0.045
			0.918	0.956	0.923	0.951	0.936	0.976	0.941	0.969
			0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
	0.2	0.625	0.999	0.998	0.999	0.999	0.999	0.986	0.999	0.989
			0.219	0.219	0.219	0.219	0.200	0.199	0.200	0.199
			0.040	0.040	0.039	0.040	0.040	0.041	0.040	0.041
			0.044	0.050	0.044	0.049	0.039	0.047	0.039	0.045
			0.935	0.966	0.940	0.964	0.935	0.977	0.941	0.972
			0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002

Table A.2: In each cell, from top to bottom are empirical type-I error/power, mean($\hat{\theta}$), SD($\hat{\theta}$), mean(SE($\hat{\theta}$)), coverage rate, MSE, when $N = 25000$ and 30% invalid IVs with uncorrelated pleiotropy.

p_overlap	θ	$\hat{\rho}$	cML-BIC-I	cML-DP-I	cML-MA-I	cML-MA-DP-I	cML-BIC-C	cML-DP-C	cML-MA-C	cML-MA-DP-C
0.0	0.0	0.002	0.167	0.039	0.131	0.045	0.165	0.032	0.130	0.039
			0.023	0.027	0.024	0.028	0.023	0.027	0.024	0.027
			0.057	0.054	0.055	0.054	0.056	0.055	0.055	0.055
			0.042	0.066	0.046	0.063	0.042	0.066	0.046	0.063
			0.833	0.961	0.870	0.955	0.835	0.968	0.870	0.961
			0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
	0.2	0.002	0.989	0.911	0.991	0.929	0.986	0.912	0.990	0.928
			0.236	0.243	0.237	0.243	0.234	0.242	0.237	0.243
			0.069	0.063	0.065	0.063	0.068	0.063	0.066	0.063
			0.048	0.080	0.052	0.076	0.048	0.079	0.052	0.075
			0.796	0.961	0.835	0.948	0.800	0.963	0.836	0.955
			0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006
0.5	0.0	0.248	0.205	0.054	0.169	0.070	0.169	0.036	0.138	0.047
			0.029	0.033	0.030	0.033	0.018	0.018	0.017	0.018
			0.058	0.056	0.056	0.056	0.056	0.054	0.055	0.054
			0.042	0.066	0.046	0.062	0.042	0.063	0.045	0.060
			0.795	0.946	0.831	0.930	0.831	0.964	0.862	0.953
			0.004	0.004	0.004	0.004	0.003	0.003	0.003	0.003
	0.2	0.314	0.992	0.942	0.993	0.964	0.983	0.930	0.980	0.944
			0.246	0.250	0.246	0.250	0.224	0.226	0.223	0.226
			0.068	0.061	0.064	0.061	0.062	0.058	0.060	0.058
			0.048	0.076	0.052	0.072	0.045	0.069	0.048	0.065
			0.737	0.956	0.801	0.938	0.799	0.970	0.848	0.961
			0.007	0.006	0.006	0.006	0.004	0.004	0.004	0.004
0.8	0.0	0.397	0.241	0.058	0.200	0.080	0.161	0.034	0.120	0.039
			0.035	0.036	0.034	0.037	0.016	0.013	0.013	0.013
			0.060	0.057	0.057	0.057	0.056	0.055	0.055	0.054
			0.042	0.066	0.046	0.062	0.042	0.062	0.045	0.059
			0.759	0.942	0.800	0.920	0.839	0.966	0.880	0.961
			0.005	0.005	0.004	0.005	0.003	0.003	0.003	0.003
	0.2	0.501	0.994	0.959	0.991	0.970	0.986	0.916	0.979	0.932
			0.252	0.254	0.252	0.255	0.218	0.216	0.216	0.217
			0.067	0.060	0.063	0.060	0.058	0.056	0.057	0.056
			0.048	0.074	0.052	0.071	0.043	0.064	0.046	0.061
			0.705	0.938	0.764	0.919	0.838	0.964	0.879	0.954
			0.007	0.007	0.007	0.007	0.004	0.003	0.003	0.003
1.0	0.0	0.496	0.252	0.074	0.207	0.091	0.161	0.035	0.125	0.043
			0.036	0.039	0.037	0.039	0.016	0.010	0.011	0.010
			0.060	0.057	0.058	0.057	0.056	0.055	0.056	0.055
			0.042	0.065	0.046	0.062	0.042	0.061	0.045	0.058
			0.748	0.926	0.793	0.909	0.839	0.965	0.875	0.957
			0.005	0.005	0.005	0.005	0.003	0.003	0.003	0.003
	0.2	0.626	0.992	0.962	0.994	0.970	0.984	0.916	0.967	0.932
			0.254	0.257	0.255	0.257	0.216	0.210	0.211	0.211
			0.066	0.059	0.062	0.059	0.057	0.055	0.056	0.055
			0.048	0.073	0.052	0.069	0.042	0.061	0.045	0.058
			0.708	0.921	0.761	0.897	0.834	0.959	0.873	0.952
			0.007	0.007	0.007	0.007	0.003	0.003	0.003	0.003

Table A.3: In each cell, from top to bottom are empirical type-I error/power, mean($\hat{\theta}$), SD($\hat{\theta}$), mean(SE($\hat{\theta}$)), coverage rate, MSE, when $N = 25000$ and 30% invalid IVs with correlated pleiotropy.

p_overlap	θ	$\hat{\rho}$	cML-BIC-I	cML-DP-I	cML-MA-I	cML-MA-DP-I	cML-BIC-C	cML-DP-C	cML-MA-C	cML-MA-DP-C
0.0	0.0	0.001	0.221	0.052	0.177	0.065	0.218	0.050	0.176	0.062
			0.026	0.030	0.027	0.031	0.026	0.030	0.027	0.031
			0.064	0.062	0.063	0.062	0.063	0.063	0.063	0.063
			0.042	0.068	0.046	0.065	0.042	0.069	0.046	0.065
			0.779	0.948	0.823	0.935	0.782	0.950	0.824	0.938
			0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005
	0.2	0.001	0.977	0.897	0.977	0.927	0.978	0.895	0.980	0.920
			0.241	0.247	0.242	0.248	0.239	0.248	0.241	0.249
			0.078	0.073	0.074	0.073	0.075	0.073	0.074	0.073
			0.047	0.082	0.052	0.077	0.047	0.082	0.052	0.077
			0.735	0.935	0.795	0.924	0.745	0.934	0.794	0.919
			0.008	0.008	0.007	0.008	0.007	0.008	0.007	0.008
0.5	0.0	0.250	0.228	0.057	0.188	0.065	0.184	0.046	0.147	0.055
			0.033	0.037	0.034	0.038	0.022	0.021	0.020	0.022
			0.063	0.059	0.060	0.059	0.060	0.057	0.058	0.057
			0.042	0.068	0.046	0.064	0.042	0.064	0.045	0.061
			0.772	0.943	0.812	0.935	0.816	0.954	0.853	0.945
			0.005	0.005	0.005	0.005	0.004	0.004	0.004	0.004
	0.2	0.315	0.988	0.939	0.986	0.962	0.982	0.915	0.977	0.937
			0.249	0.255	0.250	0.256	0.228	0.230	0.227	0.231
			0.072	0.066	0.068	0.066	0.065	0.062	0.063	0.062
			0.048	0.079	0.053	0.075	0.045	0.071	0.048	0.067
			0.728	0.928	0.783	0.916	0.785	0.960	0.828	0.949
			0.008	0.007	0.007	0.008	0.005	0.005	0.005	0.005
0.8	0.0	0.397	0.256	0.067	0.206	0.084	0.176	0.045	0.139	0.055
			0.039	0.042	0.040	0.043	0.020	0.018	0.019	0.019
			0.061	0.058	0.059	0.058	0.057	0.055	0.055	0.055
			0.042	0.067	0.046	0.063	0.042	0.062	0.045	0.059
			0.744	0.933	0.794	0.916	0.824	0.955	0.861	0.945
			0.005	0.005	0.005	0.005	0.004	0.003	0.003	0.003
	0.2	0.502	0.993	0.958	0.992	0.972	0.984	0.928	0.985	0.945
			0.257	0.261	0.258	0.261	0.223	0.222	0.221	0.223
			0.069	0.063	0.066	0.063	0.059	0.057	0.058	0.057
			0.048	0.076	0.052	0.072	0.043	0.064	0.046	0.061
			0.697	0.920	0.753	0.899	0.819	0.958	0.855	0.948
			0.008	0.008	0.008	0.008	0.004	0.004	0.004	0.004
1.0	0.0	0.496	0.259	0.073	0.222	0.090	0.151	0.044	0.111	0.052
			0.041	0.044	0.041	0.044	0.019	0.015	0.015	0.015
			0.061	0.057	0.058	0.056	0.055	0.054	0.054	0.054
			0.042	0.066	0.046	0.063	0.042	0.061	0.045	0.058
			0.741	0.927	0.778	0.910	0.849	0.956	0.889	0.948
			0.005	0.005	0.005	0.005	0.003	0.003	0.003	0.003
	0.2	0.627	0.994	0.972	0.994	0.982	0.988	0.935	0.985	0.949
			0.259	0.262	0.260	0.263	0.219	0.215	0.215	0.216
			0.067	0.060	0.063	0.059	0.055	0.053	0.054	0.053
			0.048	0.075	0.052	0.071	0.042	0.061	0.045	0.058
			0.709	0.916	0.754	0.896	0.849	0.953	0.890	0.943
			0.008	0.007	0.008	0.007	0.003	0.003	0.003	0.003

A.3.2 Detailed results of simulations in Appendix A.1.4.1

Table A.4: In each cell, from top to bottom are empirical power, mean($\hat{\theta}$), SD($\hat{\theta}$), mean(SE($\hat{\theta}$)), coverage rate, MSE, when $\theta = 0.2$, $m = 100$ and 30% invalid IVs with uncorrelated pleiotropy.

(a) $\rho = 0$

N	cML-BIC-C	cML-BIC-C-MPLE	cML-BIC-I
25,000	1.000e+00	1.000e+00	1.000e+00
	2.029e-01	2.029e-01	2.029e-01
	1.136e-02	1.136e-02	1.136e-02
	9.202e-03	9.137e-03	9.202e-03
	8.799e-01	8.771e-01	8.799e-01
	1.374e-04	1.374e-04	1.374e-04
100,000	1.000e+00	1.000e+00	1.000e+00
	2.004e-01	2.004e-01	2.004e-01
	5.337e-03	5.337e-03	5.337e-03
	4.696e-03	4.688e-03	4.696e-03
	9.130e-01	9.122e-01	9.130e-01
	2.862e-05	2.862e-05	2.862e-05
500,000	1.000e+00	1.000e+00	1.000e+00
	2.000e-01	2.000e-01	2.000e-01
	2.283e-03	2.283e-03	2.283e-03
	2.128e-03	2.127e-03	2.128e-03
	9.327e-01	9.327e-01	9.327e-01
	5.211e-06	5.211e-06	5.211e-06
1,000,000	1.000e+00	1.000e+00	1.000e+00
	2.000e-01	2.000e-01	2.000e-01
	1.600e-03	1.600e-03	1.600e-03
	1.510e-03	1.510e-03	1.510e-03
	9.327e-01	9.327e-01	9.327e-01
	2.560e-06	2.560e-06	2.560e-06
5,000,000	1.000e+00	1.000e+00	1.000e+00
	2.000e-01	2.000e-01	2.000e-01
	7.043e-04	7.043e-04	7.043e-04
	6.784e-04	6.784e-04	6.784e-04
	9.400e-01	9.400e-01	9.401e-01
	4.962e-07	4.962e-07	4.962e-07

(b) $\rho = 0.8$

N	cML-BIC-C	cML-BIC-C-MPLE	cML-BIC-I
25,000	1.000e+00	1.000e+00	1.000e+00
	2.000e-01	2.000e-01	2.074e-01
	9.390e-03	9.390e-03	9.732e-03
	7.718e-03	7.690e-03	9.201e-03
	8.949e-01	8.938e-01	8.550e-01
	8.817e-05	8.817e-05	1.497e-04
100,000	1.000e+00	1.000e+00	1.000e+00
	2.000e-01	2.000e-01	2.015e-01
	4.416e-03	4.416e-03	4.559e-03
	3.925e-03	3.922e-03	4.696e-03
	9.204e-01	9.202e-01	9.449e-01
	1.950e-05	1.950e-05	2.300e-05
500,000	1.000e+00	1.000e+00	1.000e+00
	2.000e-01	2.000e-01	2.002e-01
	1.887e-03	1.890e-03	1.931e-03
	1.774e-03	1.774e-03	2.128e-03
	9.346e-01	9.344e-01	9.670e-01
	3.560e-06	3.571e-06	3.783e-06
1,000,000	1.000e+00	1.000e+00	1.000e+00
	2.000e-01	2.000e-01	2.001e-01
	1.325e-03	1.325e-03	1.349e-03
	1.258e-03	1.258e-03	1.510e-03
	9.372e-01	9.371e-01	9.700e-01
	1.756e-06	1.756e-06	1.831e-06
5,000,000	1.000e+00	1.000e+00	1.000e+00
	2.000e-01	2.000e-01	2.000e-01
	5.828e-04	5.828e-04	5.906e-04
	5.649e-04	5.649e-04	6.784e-04
	9.400e-01	9.400e-01	9.743e-01
	3.398e-07	3.398e-07	3.489e-07

Table A.5: In each cell, from top to bottom are empirical power, mean($\hat{\theta}$), SD($\hat{\theta}$), mean(SE($\hat{\theta}$)), coverage rate, MSE, when $\theta = 0.2$, $m = 100$ and 30% invalid IVs with correlated pleiotropy.

(a) $\rho = 0$

N	cML-BIC-C	cML-BIC-C-MPLE	cML-BIC-I
25,000	1.000e+00	1.000e+00	1.000e+00
	2.033e-01	2.033e-01	2.033e-01
	1.141e-02	1.141e-02	1.141e-02
	9.183e-03	9.120e-03	9.183e-03
	8.722e-01	8.698e-01	8.722e-01
	1.411e-04	1.411e-04	1.411e-04
100,000	1.000e+00	1.000e+00	1.000e+00
	2.005e-01	2.005e-01	2.005e-01
	5.355e-03	5.355e-03	5.355e-03
	4.691e-03	4.684e-03	4.691e-03
	9.129e-01	9.125e-01	9.129e-01
	2.889e-05	2.889e-05	2.889e-05
500,000	1.000e+00	1.000e+00	1.000e+00
	2.000e-01	2.000e-01	2.000e-01
	2.294e-03	2.294e-03	2.294e-03
	2.127e-03	2.126e-03	2.127e-03
	9.299e-01	9.297e-01	9.299e-01
	5.261e-06	5.261e-06	5.261e-06
1,000,000	1.000e+00	1.000e+00	1.000e+00
	2.000e-01	2.000e-01	2.000e-01
	1.607e-03	1.607e-03	1.607e-03
	1.509e-03	1.509e-03	1.509e-03
	9.358e-01	9.358e-01	9.358e-01
	2.581e-06	2.581e-06	2.581e-06
5,000,000	1.000e+00	1.000e+00	1.000e+00
	2.000e-01	2.000e-01	2.000e-01
	7.029e-04	7.029e-04	7.029e-04
	6.784e-04	6.783e-04	6.784e-04
	9.416e-01	9.416e-01	9.416e-01
	4.941e-07	4.941e-07	4.941e-07

(b) $\rho = 0.8$

N	cML-BIC-C	cML-BIC-C-MPLE	cML-BIC-I
25,000	1.000e+00	1.000e+00	1.000e+00
	2.003e-01	2.003e-01	2.077e-01
	9.412e-03	9.412e-03	9.793e-03
	7.702e-03	7.675e-03	9.182e-03
	8.906e-01	8.896e-01	8.453e-01
	8.869e-05	8.869e-05	1.554e-04
100,000	1.000e+00	1.000e+00	1.000e+00
	2.000e-01	2.000e-01	2.016e-01
	4.432e-03	4.432e-03	4.582e-03
	3.921e-03	3.918e-03	4.690e-03
	9.183e-01	9.178e-01	9.423e-01
	1.964e-05	1.964e-05	2.346e-05
500,000	1.000e+00	1.000e+00	1.000e+00
	2.000e-01	2.000e-01	2.002e-01
	1.900e-03	1.900e-03	1.945e-03
	1.774e-03	1.773e-03	2.127e-03
	9.323e-01	9.322e-01	9.672e-01
	3.610e-06	3.610e-06	3.845e-06
1,000,000	1.000e+00	1.000e+00	1.000e+00
	2.000e-01	2.000e-01	2.001e-01
	1.319e-03	1.319e-03	1.356e-03
	1.258e-03	1.258e-03	1.509e-03
	9.372e-01	9.372e-01	9.700e-01
	1.740e-06	1.740e-06	1.853e-06
5,000,000	1.000e+00	1.000e+00	1.000e+00
	2.000e-01	2.000e-01	2.000e-01
	5.805e-04	5.805e-04	5.876e-04
	5.649e-04	5.649e-04	6.784e-04
	9.431e-01	9.431e-01	9.754e-01
	3.370e-07	3.370e-07	3.455e-07

Table A.6: In each cell, from top to bottom are empirical accuracy, true positive rate, true negative rate, when $\theta = 0.2$, $m = 100$ and 30% invalid IVs with uncorrelated pleiotropy.

(a) $\rho = 0$

n	cML-BIC-I	cML-BIC-C
25,000	9.281e-01 7.638e-01 9.985e-01	9.281e-01 7.638e-01 9.985e-01
100,000	9.617e-01 8.738e-01 9.993e-01	9.617e-01 8.738e-01 9.993e-01
500,000	9.817e-01 9.396e-01 9.997e-01	9.817e-01 9.396e-01 9.997e-01
1,000,000	9.867e-01 9.561e-01 9.998e-01	9.867e-01 9.561e-01 9.998e-01
5,000,000	9.938e-01 9.794e-01 9.999e-01	9.938e-01 9.794e-01 9.999e-01

(b) $\rho = 0.8$

n	cML-BIC-I	cML-BIC-C
25,000	9.285e-01 7.620e-01 9.999e-01	9.381e-01 8.003e-01 9.972e-01
100,000	9.621e-01 8.739e-01 1.000e+00	9.673e-01 8.939e-01 9.987e-01
500,000	9.819e-01 9.396e-01 1.000e+00	9.845e-01 9.493e-01 9.996e-01
1,000,000	9.868e-01 9.560e-01 1.000e+00	9.888e-01 9.632e-01 9.998e-01
5,000,000	9.938e-01 9.793e-01 1.000e+00	9.948e-01 9.828e-01 9.999e-01

Table A.7: In each cell, from top to bottom are empirical accuracy, true positive rate, true negative rate, when $\theta = 0.2$, $m = 100$ and 30% invalid IVs with correlated pleiotropy.

(a) $\rho = 0$

n	cML-BIC-I	cML-BIC-C
25,000	9.331e-01 7.806e-01 9.985e-01	9.331e-01 7.806e-01 9.985e-01
100,000	9.644e-01 8.831e-01 9.993e-01	9.644e-01 8.831e-01 9.993e-01
500,000	9.831e-01 9.443e-01 9.997e-01	9.831e-01 9.443e-01 9.997e-01
1,000,000	9.878e-01 9.597e-01 9.998e-01	9.878e-01 9.597e-01 9.998e-01
5,000,000	9.943e-01 9.812e-01 9.999e-01	9.943e-01 9.812e-01 9.999e-01

(b) $\rho = 0.8$

n	cML-BIC-I	cML-BIC-C
25,000	9.337e-01 7.792e-01 9.999e-01	9.427e-01 8.153e-01 9.974e-01
100,000	9.648e-01 8.828e-01 1.000e+00	9.696e-01 9.016e-01 9.987e-01
500,000	9.833e-01 9.443e-01 1.000e+00	9.858e-01 9.536e-01 9.996e-01
1,000,000	9.879e-01 9.598e-01 1.000e+00	9.898e-01 9.664e-01 9.998e-01
5,000,000	9.944e-01 9.812e-01 1.000e+00	9.952e-01 9.843e-01 9.999e-01

A.3.3 Detailed results of simulations in Appendix A.1.4.2

Table A.8: In each cell, from top to bottom are empirical power, mean($\hat{\theta}$), SD($\hat{\theta}$), mean(SE($\hat{\theta}$)), coverage rate, MSE, when $\theta = 0.2$, $m = 100$, ρ=0.

N	cML-BIC-C-K0	cML-BIC-C-K0-MPLE	cML-BIC-I-K0	cML-BIC-I-K0-robust
25,000	1.000e+00	1.000e+00	1.000e+00	1.000e+00
	1.999e-01	1.999e-01	1.999e-01	1.999e-01
	8.092e-03	8.092e-03	8.092e-03	8.092e-03
	8.062e-03	8.015e-03	8.062e-03	8.028e-03
	9.504e-01	9.487e-01	9.504e-01	9.456e-01
	6.548e-05	6.548e-05	6.548e-05	6.548e-05
100,000	1.000e+00	1.000e+00	1.000e+00	1.000e+00
	1.999e-01	1.999e-01	1.999e-01	1.999e-01
	4.039e-03	4.039e-03	4.039e-03	4.039e-03
	4.031e-03	4.025e-03	4.031e-03	4.006e-03
	9.509e-01	9.505e-01	9.509e-01	9.458e-01
	1.631e-05	1.631e-05	1.631e-05	1.631e-05
500,000	1.000e+00	1.000e+00	1.000e+00	1.000e+00
	2.000e-01	2.000e-01	2.000e-01	2.000e-01
	1.806e-03	1.806e-03	1.806e-03	1.806e-03
	1.803e-03	1.802e-03	1.803e-03	1.791e-03
	9.505e-01	9.504e-01	9.505e-01	9.462e-01
	3.260e-06	3.260e-06	3.260e-06	3.260e-06
1,000,000	1.000e+00	1.000e+00	1.000e+00	1.000e+00
	2.000e-01	2.000e-01	2.000e-01	2.000e-01
	1.277e-03	1.277e-03	1.277e-03	1.277e-03
	1.275e-03	1.275e-03	1.275e-03	1.266e-03
	9.500e-01	9.499e-01	9.500e-01	9.460e-01
	1.630e-06	1.630e-06	1.630e-06	1.630e-06
5,000,000	1.000e+00	1.000e+00	1.000e+00	1.000e+00
	2.000e-01	2.000e-01	2.000e-01	2.000e-01
	5.710e-04	5.710e-04	5.710e-04	5.710e-04
	5.701e-04	5.701e-04	5.701e-04	5.662e-04
	9.498e-01	9.498e-01	9.498e-01	9.462e-01
	3.261e-07	3.261e-07	3.261e-07	3.261e-07

Table A.9: In each cell, from top to bottom are empirical power, mean($\hat{\theta}$), SD($\hat{\theta}$), mean(SE($\hat{\theta}$)), coverage rate, MSE, when $\theta = 0.2$, $m = 100$, ρ=0.8.

N	cML-BIC-C-K0	cML-BIC-C-K0-MPLE	cML-BIC-I-K0	cML-BIC-I-K0-robust
25,000	1.000e+00	1.000e+00	1.000e+00	1.000e+00
	1.999e-01	1.999e-01	2.044e-01	2.044e-01
	6.747e-03	6.747e-03	6.706e-03	6.706e-03
	6.710e-03	6.689e-03	8.063e-03	6.626e-03
	9.487e-01	9.479e-01	9.528e-01	8.897e-01
	4.554e-05	4.554e-05	6.455e-05	6.455e-05
100,000	1.000e+00	1.000e+00	1.000e+00	1.000e+00
	1.999e-01	1.999e-01	2.011e-01	2.011e-01
	3.369e-03	3.369e-03	3.365e-03	3.365e-03
	3.354e-03	3.352e-03	4.031e-03	3.326e-03
	9.488e-01	9.488e-01	9.733e-01	9.323e-01
	1.136e-05	1.136e-05	1.249e-05	1.249e-05
500,000	1.000e+00	1.000e+00	1.000e+00	1.000e+00
	2.000e-01	2.000e-01	2.002e-01	2.002e-01
	1.506e-03	1.506e-03	1.506e-03	1.506e-03
	1.500e-03	1.500e-03	1.803e-03	1.489e-03
	9.487e-01	9.487e-01	9.789e-01	9.432e-01
	2.269e-06	2.269e-06	2.310e-06	2.310e-06
1,000,000	1.000e+00	1.000e+00	1.000e+00	1.000e+00
	2.000e-01	2.000e-01	2.001e-01	2.001e-01
	1.065e-03	1.065e-03	1.065e-03	1.065e-03
	1.061e-03	1.061e-03	1.275e-03	1.053e-03
	9.484e-01	9.484e-01	9.793e-01	9.446e-01
	1.135e-06	1.135e-06	1.144e-06	1.144e-06
5,000,000	1.000e+00	1.000e+00	1.000e+00	1.000e+00
	2.000e-01	2.000e-01	2.000e-01	2.000e-01
	4.763e-04	4.763e-04	4.763e-04	4.763e-04
	4.743e-04	4.743e-04	5.701e-04	4.711e-04
	9.482e-01	9.482e-01	9.796e-01	9.467e-01
	2.269e-07	2.269e-07	2.271e-07	2.271e-07

A.4 Graph-MRcML

A.4.1 Bidirectional MR-cML-C

To infer a possibly bi-directional causal relationship between a pair of traits, say X and Y, we apply bidirectional MR-cML-C Xue and Pan (2022). In practice, we first select (genome-wide) significant (and nearly independent) SNPs for trait X, then do so for Y. Next for each SNP selected for both X and Y, we only use it as IV for the trait that has a larger absolute value of correlation with the SNP; this screening rule combines a simple application of Steiger's method Hemani et al. (2017) with the fact that an SNP cannot be a valid IV for both traits at the same time. Then we use the significant SNPs for X after excluding those (if any) based on the above screening rule as the IVs and apply MR-cML-C for causal direction $X \to Y$; similarly, we use the significant SNPs for Y after excluding those (if any) based on the screening rule as the IVs in MR-cML-C for $Y \to X$. In this way, we do not need to specify the causal direction(s) a priori and thus infer a possibly bi-directional relationship between the two traits.

A.4.2 Effective number of tests

For T traits, $T \times (T-1)$ tests are performed in total. To account for multiple testing, one can use a conservative Bonferroni adjustment with the total number of tests. However, we expect that these $T \times (T-1)$ tests are not independent with each other as we reuse each GWAS dataset multiple times and some of them might have overlapping samples. We take the approach described in Li et al. (2011) to calculate the effective number of independent p-values (M_e). Specifically, in each data perturbation, we could obtain p-values from MR-cML-BIC-C for the $T \times (T-1)$ tests, and thus through hundreds of times of data perturbation, we could obtain a

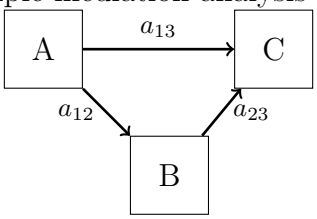
$T(T-1) \times T(T-1)$ sample correlation matrix for the $T \times (T-1)$ tests. Then the effective number of independent tests is estimated as

$$M_e = T(T-1) - \sum_{i=1}^{T(T-1)} I(\omega_i > 1) \times (\omega_i - 1),$$

where ω_i's are the eigenvalues of the $T(T-1) \times T(T-1)$ sample correlation matrix of the p-values and $I(\cdot)$ is the indicator function. We apply the Bonferroni adjustment with M_e and claim an edge in the graph is statistically significant if its p-value was smaller than $0.05/M_e$.

A.4.3 A simple example of network deconvolution

Here, we illustrate the idea of network deconvolution (Eq.(7) in the main text) via a simple mediation analysis with three traits as follows:



Then the corresponding $\mathbf{G}_{dir}$, $\mathbf{G}_{dir}^2$ and $\mathbf{G}_{dir}^3$ are given as: $\mathbf{G}_{dir} = \begin{bmatrix} 0 & a_{12} & a_{13} \\ 0 & 0 & a_{23} \\ 0 & 0 & 0 \end{bmatrix}$, $\mathbf{G}_{dir}^2 = \begin{bmatrix} 0 & 0 & a_{12}a_{23} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$, $\mathbf{G}_{dir}^3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$. And the total network is $\mathbf{G}_{tot} = \begin{bmatrix} 0 & a_{12} & a_{13} + a_{12}a_{23} \\ 0 & 0 & a_{23} \\ 0 & 0 & 0 \end{bmatrix}$. We can see that, the (1,3)th element in $\mathbf{G}_{dir}^2$ corresponds to the indirect effect $a_{12}a_{23}$ from A to C (mediated through B). Since there is no causal pathway of length 3, $\mathbf{G}_{dir}^3$ will be a zero matrix. And the total effect from A to C is $a_{13} + a_{12}a_{23}$. Finally, it is easy to verify that $\mathbf{G}_{tot} = \mathbf{G}_{dir} + \mathbf{G}_{dir}^2 + \mathbf{G}_{dir}^3 + \cdots =$

$\mathbf{G}_{dir} + \mathbf{G}_{dir}^2 = \mathbf{G}_{dir}(\mathbf{I} - \mathbf{G}_{dir})^{-1}$.

A.4.4 The diagonal elements of a total graph

In the main text, we proposed **Graph-MRcML-d0** and **Graph-MRcML-d1** with two ways to specify the diagonal elements of a total graph. In Graph-MRcML-d0, we set the diagonal elements to zeros, following the practice in (Feizi et al., 2013). However, this may be problematic if there are cycles in the underlying direct graph, thus we proposed the following iterative algorithm to update the diagonal elements of $\mathbf{G}_{tot}$ in Graph-MRcML-d1. Denote $\mathbf{G}_{tot} = (T_{ij})$, $\mathbf{G}_{dir} = (D_{ij})$ and the bidirectional MR-cML estimates $\hat{T}_{ij}$ $(i \neq j)$.

Algorithm 3 Estimation of $\mathbf{G}_{dir}$ in Graph-MRcML-d1

$\hat{T}_{ii} \leftarrow \sum_{j \neq i} \hat{T}_{ji}\hat{T}_{ij}$
$\widehat{\mathbf{G}}^0_{dir}, \widehat{\mathbf{G}}_{dir} \leftarrow \widehat{\mathbf{G}}_{tot}(\mathbf{I} + \widehat{\mathbf{G}}_{tot})^{-1}$
$t \leftarrow 0$
while $|\hat{D}_{ii}| > \epsilon$ and $t <$ maxit **do** ▷ ϵ is a small value, e.g. 10^{-4}
  $\hat{T}_{ii} \leftarrow \sum_{j \neq i} \hat{D}_{ji}\hat{T}_{ij}$ ▷ Eq.(8) in the main text
  $\widehat{\mathbf{G}}_{dir} \leftarrow \widehat{\mathbf{G}}_{tot}(\mathbf{I} + \widehat{\mathbf{G}}_{tot})^{-1}$
  $t \leftarrow t + 1$
end while
if $t =$ maxit **then**
  $\widehat{\mathbf{G}}_{dir} \leftarrow \widehat{\mathbf{G}}^0_{dir}$
end if
return $\widehat{\mathbf{G}}_{dir}$

The proposed iterative algorithm is motivated and illustrated by the following examples.

A.4.4.1 Example 1

Consider the true direct graph ($\mathbf{G}_{dir}$) with edges $A \to B, B \to C$ and $C \to A$, and all with effect size of 0.6 (Table A.10a). The corresponding true total graph ($\mathbf{G}_{tot}$) is shown in Table A.10b (round to 7 decimal places), which has non-zero diagonal elements (0.2755102, 0.2755102, 0.2755102). Incorrectly specifying the diagonal elements of $\mathbf{G}_{tot}$ to zeros (as in Graph-MRcML-d0) would lead to an incorrect direct graph as shown in Table A.10c.

Table A.10: Example 1: (a) true direct graph, (b) true total graph and (c) incorrect direct graph by setting diag($\mathbf{G}_{tot}$) to zeros, among three nodes.

	A	B	C
A	0	0.6	0
B	0	0	0.6
C	0.6	0	0

(a) True direct graph

	A	B	C
A	0.2755102	0.7653061	0.4591837
B	0.4591837	0.2755102	0.7653061
C	0.7653061	0.4591837	0.2755102

(b) True total graph

	A	B	C
A	-0.3214689	1.1296873	-0.2577596
B	-0.2577596	-0.3214689	1.1296873
C	1.1296873	-0.2577596	-0.3214689

(c) Incorrect direct graph

We applied Algorithm 3 to iteratively estimate $\mathbf{G}_{dir}$, and plotted the values of $\hat{D}_{11}$, $\hat{D}_{12}$ and $\hat{D}_{13}$ in each iteration as shown in the black line in Figure A.16. We can see that they converged to the true values (in red) in 8 iterations (so did other entries not plotted here), while the blue dashed line was the estimate in Table A.10c.

A.4.4.2 Example 2

Here we give an example that Algorithm 3 failed to converge, but using the result in the first iteration (i.e., $\widehat{\mathbf{G}}^0_{dir}$) was able to improve the result over that by simply

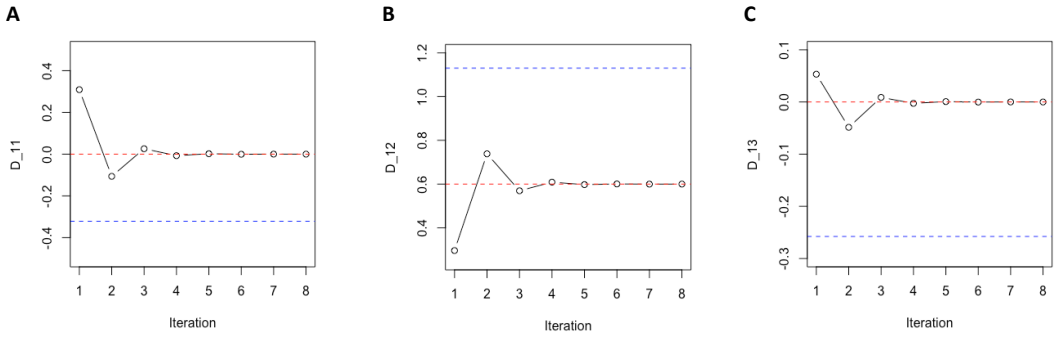


Figure A.16: Example 1: values of D_{11} (panel A), D_{12} (panel B) and D_{13} (panel C) estimates in each iteration (black line). Dashed line in red is the true value and dashed line in blue is the estimate by setting diagonal elements of the total graph to zeros.

setting the diagonal elements to zeros. The true direct graph, true total graph and the incorrect direct graph are given in Table A.11a, Table A.11b, Table A.11c respectively. As shown in the panel A in Figure A.17, Algorithm 3 failed to converge (as $\hat{D}_{11}$ failed to converge to 0), but the initial estimates (as shown in iteration 1) were much closer to the truth (in red) than to the blue lines. The initial estimate $\widehat{\mathbf{G}}^0_{dir}$ is given in Table A.11d, which was closer to the truth (Table A.11a) than that given in Table A.11c. In fact, the spectral radius of Table A.11c was also larger than 1, casting doubt on its reliability.

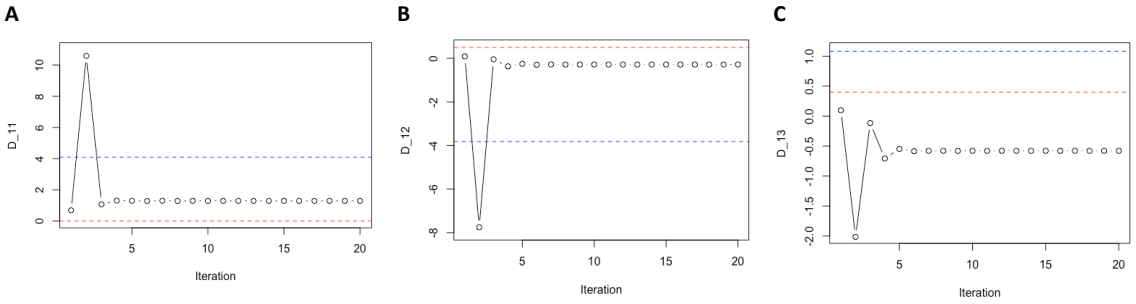


Figure A.17: Example 2: values of D_{11} (panel A), D_{12} (panel B) and D_{13} (panel C) estimates in each iteration (black line). Dashed line in red is the true value and dashed line in blue is the estimate by setting diagonal elements of the total graph to zeros.

Table A.11: Example 2: (a) true direct graph, (b) true total graph, (c) incorrect direct graph by setting diag($\mathbf{G}_{tot}$) to zeros, and (d) initial estimate $\widehat{\mathbf{G}}^0_{dir}$ among three nodes.

	A	B	C
A	0	0.5	0.4
B	0.4	0	0.5
C	0.5	0	0

(a) True direct graph

	A	B	C
A	1.105263	1.052632	1.368421
B	1.368421	0.684211	1.389474
C	1.052632	0.526316	0.684211

(b) True total graph

	A	B	C
A	4.087102	-3.819095	1.082077
B	1.082077	-4.060302	5.550419
C	-3.819095	6.683417	-4.060302

(c) Incorrect direct graph

	A	B	C
A	0.684749	0.088488	0.097248
B	0.097248	0.632704	0.118948
C	0.088488	0.031582	0.632704

(d) $\widehat{\mathbf{G}}^0_{dir}$

A.4.5 Simulation for direct graph inference in the main text

We simulated data based on the estimated direct graph by Graph-MRcML-d0 among 6 traits as shown in Table A.12. We point out that in Set-up (a), there was no loop, and thus the diagonal elements of the total graph were all zeros. But in Set-up (b),

there was a cycle between CAD and AF, and their corresponding diagonal elements in the total graph were not zeros (Table A.13a). Table A.13b shows the resulting (incorrect) direct graph if we set the diagonal elements of the total graph to zeros in Set-up (b).

Table A.12: True direct graphs in the simulation for (a) Set-up (a), and (b) Set-up (b).

	BMI	LDL	FG	TG	CAD	Stroke
BMI	0	0	0.07	0.23	0.23	0
LDL	0	0	0	0	0.41	0
FG	0	0	0	0	0.31	0
TG	0	0	0	0	0	0
CAD	0	0	0	0	0	0.23
Stroke	0	0	0	0	0	0

(a) Set-up (a)

	BMI	LDL	FG	TG	CAD	AF
BMI	0	0	0.08	0.23	0.22	0.32
LDL	0	0	0	0	0.42	0
FG	0	0	0	0	0.33	0
TG	0	0	0	0	0	0
CAD	0	0	0	0	0	0.17
AF	0	0	0	0	0.10	0

(b) Set-up (b)

Table A.13: In Set-up (b): (a) the true total graph $\mathbf{G}_{tot}$, and (b) resulting direct graph when specifying diag($\mathbf{G}_{tot}$) to zeros. Numbers are rounded to 3 decimal places.

	BMI	LDL	FG	TG	CAD	AF
BMI	0	0	0.08	0.23	0.283	0.368
LDL	0	0	0	0	0.427	0.073
FG	0	0	0	0	0.336	0.057
TG	0	0	0	0	0	0
CAD	0	0	0	0	0.017	0.173
AF	0	0	0	0	0.102	0.017

(a)

	BMI	LDL	FG	TG	CAD	AF
BMI	0	0	0.08	0.23	0.223	0.325
LDL	0	0	0	0	0.427	-0.001
FG	0	0	0	0	0.336	-0.001
TG	0	0	0	0	0	0
CAD	0	0	0	0	-0.018	0.176
AF	0	0	0	0	0.104	-0.018

(b)

Let BMI, LDL, FG, TG, CAD and Stroke (or AF) be the first, second, ..., and the 6-th trait respecitvely. Then we generated the *true* association effect sizes $\mathbf{B} = (b_{ij})$, where b_{ij} was the association effect size between the i-th SNP and the j-th trait, as follows. First we initialized $\mathbf{B}$ as a zero matrix. Then for the j-th trait, $j = 1, 2, \ldots, 6$,

1. for $i \in \mathcal{S}_j$, where $\mathcal{S}_j$ is the set of GWAS significant SNPs (i.e. p-value<5e-8) for the j-th trait, let $b_{ij} = \hat{\beta}_{ij}$, where $\hat{\beta}_{ij}$ is the GWAS estimate from the corresponding GWAS dataset used in the real data analysis;

2. for the k-th trait, $k \neq j$, let $b_{ik} = \theta_{jk} b_{ij}$, $i \in \mathcal{S}_j$, where θ_{jk} is the true total effect size from the j-th trait to the k-th trait.

With this data-generating procedure, some invalid IVs (with horizontal pleiotropy) were also generated for some pairs of traits, as some traits (e.g. LDL and TG) shared common IVs in nature. Then with the generated $\mathbf{B}$, in each simulation replicate, we simulated the *estimated* GWAS summary statistics as $\widehat{\mathbf{B}} = \mathbf{B} + \mathbf{S} * \mathbf{E}$, where $\mathbf{S}$ was the standard errors of GWAS estimates obtained from the real GWAS summary statistics. Alternatively we considered different GWAS sample sizes by setting $\mathbf{S} = (S_{ij})$, and $S_{ij} = 1/\sqrt{N}$, $N = 300\,000$, $500\,000$ or $1\,000\,000$ for all 6 traits. $\mathbf{E}$ was generated from a matrix normal distribution as in data perturbation, $\mathbf{E} \sim \mathcal{MN}(\mathbf{0}, \mathbf{R}, \mathbf{P})$, and $\mathbf{R}$ and $\mathbf{P}$ were the LD matrix of the SNPs and the correlation matrix among the 6 GWAS traits respectively, estimated from the real GWAS data as well.

A.4.5.1 Results with S from real GWAS summary statistics

Table A.14: Empirical type-I error and power by Graph-MRcML-d0 for (a) Set-up (a) and (b) Set-up (b). Numbers underlined correspond to power.

	BMI	LDL	FG	TG	CAD	Stroke
BMI		0	0.6	1	0.9	0
LDL	0		0	0.01	1	0
FG	0	0		0	0.89	0
TG	0	0.01	0		0.06	0.01
CAD	0	0	0	0		1
Stroke	0	0.01	0	0	0	

(a) Set-up (a)

	BMI	LDL	FG	TG	CAD	AF
BMI		0	0.81	1	0.87	1
LDL	0.02		0	0	1	0
FG	0	0		0	0.97	0
TG	0	0	0		0.01	0
CAD	0	0	0	0		1
AF	0	0	0	0	1	

(b) Set-up (b)

Table A.15: Set-up (a): mean estimated direct graph by (a) Graph-MRcML-d0, and (b) Graph-MRcML-d1 across 100 replicates. Numbers are rounded to 3 decimal places.

	BMI	LDL	FG	TG	CAD	Stroke	
BMI	0	-0.001	0.069	0.231	0.233	-0.006	BMI
LDL	0	0	0.001	0	0.409	0.005	LDL
FG	0.001	0.002	0	0.002	0.311	0	FG
TG	0	0	0	0	-0.004	0.019	TG
CAD	0	0	0	0	0	0.23	CAD
Stroke	0.001	0	-0.002	0	0	0	Stroke

(a) Graph-MRcML-d0

	BMI	LDL	FG	TG	CAD	Stroke
BMI	0	-0.001	0.069	0.23	0.233	-0.006
LDL	0	0	0.001	0	0.409	0.005
FG	0.001	0.002	0	0.002	0.311	0
TG	0	0	0	0	-0.004	0.019
CAD	0	0	0	0	0	0.23
Stroke	0.001	0	-0.002	0	0	0

(b) Graph-MRcML-d1

Table A.16: Set-up (b): mean estimated direct graph by (a) Graph-MRcML-d0, and (b) Graph-MRcML-d1 across 100 replicates. Numbers are rounded to 3 decimal places.

	BMI	LDL	FG	TG	CAD	AF	
BMI	0	-0.001	0.083	0.232	0.19	0.309	BMI
LDL	0	0	-0.001	-0.004	0.45	0.004	LDL
FG	0.001	0.002	0	0.004	0.339	0	FG
TG	-0.001	0.001	0	0	0.08	0.018	TG
CAD	0.001	0	0.001	0	-0.018	0.174	CAD
AF	-0.001	0	0	0	0.104	-0.018	AF

(a) Graph-MRcML-d0

	BMI	LDL	FG	TG	CAD	AF
BMI	0	-0.001	0.083	0.232	0.188	0.305
LDL	0	0	-0.001	-0.004	0.442	0.006
FG	0	0.002	0	0.004	0.332	0.001
TG	-0.001	0.001	0	0	0.079	0.018
CAD	0.001	0	0.001	0	0	0.168
AF	-0.001	0	0	0	0.1	0

(b) Graph-MRcML-d1

A.4.5.2 Results with varying sample sizes

A.4.5.2.1 Set-up (a) In the absence of cycles in the direct graph, the diagonal elements of the true total graph were zeros, and Graph-MRcML-d0 correctly specified them. When the diagonal elements were consistently specified, the resulting estimate of the direct graph was consistent. Table A.17 to Table A.19 show the mean of the estimated direct graphs across 100 simulation replicates by Graph-MRcML-d0 (panel (a)) and Graph-MRcML-d1 (panel (b)) for different sample sizes; Table A.20 to Table A.22 show the empirical type-I error and power. As expected, the estimate from Graph-MRcML-d0 had smaller bias as sample size increased. Also, Graph-MRcML-d1 performed almost identically as Graph-MRcML-d0, suggesting that the

iterative algorithm was able to correctly specify the diagonal elements (of the total graph) when there was no cycle in the direct graph. Type-I error was also controlled as the sample size increased.

Table A.17: Set-up (a): mean estimated direct graph by (a) Graph-MRcML-d0, and (b) Graph-MRcML-d1 across 100 replicates when $N = 300\,000$. Numbers are rounded to 3 decimal places.

	BMI	LDL	FG	TG	CAD	Stroke
BMI	0	-0.001	0.069	0.231	0.233	-0.006
LDL	0	0	0.001	0	0.409	0.005
FG	0.001	0.002	0	0.002	0.311	0
TG	0	0	0	0	-0.004	0.019
CAD	0	0	0	0	0	0.23
Stroke	0.001	0	-0.002	0	0	0

(a) Graph-MRcML-d0

	BMI	LDL	FG	TG	CAD	Stroke
BMI	0	-0.001	0.069	0.23	0.233	-0.006
LDL	0	0	0.001	0	0.409	0.005
FG	0.001	0.002	0	0.002	0.311	0
TG	0	0	0	0	-0.004	0.019
CAD	0	0	0	0	0	0.23
Stroke	0.001	0	-0.002	0	0	0

(b) Graph-MRcML-d1

Table A.18: Set-up (a): mean estimated direct graph by (a) Graph-MRcML-d0, and (b) Graph-MRcML-d1 across 100 replicates when $N = 500\,000$. Numbers are rounded to 3 decimal places.

	BMI	LDL	FG	TG	CAD	Stroke
BMI	0	-0.001	0.069	0.23	0.232	-0.004
LDL	0	0	0	0	0.409	0.003
FG	0.001	0.002	0	0.001	0.31	0
TG	0	0	0	0	-0.002	0.013
CAD	0	0	0	0	0	0.23
Stroke	0.001	0	-0.002	0	0	0

(a) Graph-MRcML-d0

	BMI	LDL	FG	TG	CAD	Stroke
BMI	0	-0.001	0.069	0.23	0.232	-0.004
LDL	0	0	0	0	0.41	0.003
FG	0.001	0.002	0	0.001	0.311	0
TG	0	0	0	0	-0.002	0.013
CAD	0	0	0	0	0	0.23
Stroke	0.001	0	-0.002	0	0	0

(b) Graph-MRcML-d1

Table A.19: Set-up (a): mean estimated direct graph by (a) Graph-MRcML-d0, and (b) Graph-MRcML-d1 across 100 replicates when $N = 1\,000\,000$. Numbers are rounded to 3 decimal places.

	BMI	LDL	FG	TG	CAD	Stroke
BMI	0	-0.001	0.07	0.23	0.231	-0.001
LDL	0	0	0	0	0.41	0.001
FG	0.001	0.001	0	0.001	0.31	0
TG	0	0	0	0	0	0.005
CAD	0	0	0	0	0	0.23
Stroke	0	0	-0.001	0	0	0

(a) Graph-MRcML-d0

	BMI	LDL	FG	TG	CAD	Stroke
BMI	0	-0.001	0.07	0.23	0.231	-0.001
LDL	0	0	0	0	0.41	0.001
FG	0.001	0.001	0	0.001	0.31	0
TG	0	0	0	0	0	0.005
CAD	0	0	0	0	0	0.23
Stroke	0	0	-0.001	0	0	0

(b) Graph-MRcML-d1

Table A.20: Set-up (a): empirical type-I error and power by (a) Graph-MRcML-d0, and (b) Graph-MRcML-d1 when $N = 300\,000$. Numbers underlined correspond to power.

	BMI	LDL	FG	TG	CAD	Stroke
BMI		0	$\underline{1}$	$\underline{1}$	$\underline{1}$	0.01
LDL	0		0	0.01	$\underline{1}$	0.05
FG	0	0		0	$\underline{1}$	0
TG	0	0.01	0.01		0	0.09
CAD	0	0	0	0		$\underline{1}$
Stroke	0	0	0	0	0	

(a) Graph-MRcML-d0

	BMI	LDL	FG	TG	CAD	Stroke
BMI		0	$\underline{1}$	$\underline{1}$	$\underline{1}$	0.01
LDL	0		0	0.01	$\underline{1}$	0.06
FG	0	0		0	$\underline{1}$	0
TG	0	0.01	0.01		0	0.09
CAD	0	0	0	0		$\underline{1}$
Stroke	0	0	0	0	0	

(b) Graph-MRcML-d1

Table A.21: Set-up (a): empirical type-I error and power by (a) Graph-MRcML-d0, and (b) Graph-MRcML-d1 when $N = 500\,000$. Numbers underlined correspond to power.

	BMI	LDL	FG	TG	CAD	Stroke
BMI		0	$\underline{1}$	$\underline{1}$	$\underline{1}$	0
LDL	0		0	0.01	$\underline{1}$	0.02
FG	0	0		0	$\underline{1}$	0
TG	0	0.01	0.01		0	0.05
CAD	0	0	0	0		$\underline{1}$
Stroke	0	0.01	0	0	0	

(a) Graph-MRcML-d0

	BMI	LDL	FG	TG	CAD	Stroke
BMI		0	$\underline{1}$	$\underline{1}$	$\underline{1}$	0
LDL	0		0	0.01	$\underline{1}$	0.01
FG	0	0		0	$\underline{1}$	0
TG	0	0.01	0.01		0	0.05
CAD	0	0	0	0		$\underline{1}$
Stroke	0	0.01	0	0	0	

(b) Graph-MRcML-d1

Table A.22: Set-up (a): empirical type-I error and power by (a) Graph-MRcML-d0, and (b) Graph-MRcML-d1 when $N = 1\,000\,000$. Numbers underlined correspond to power.

	BMI	LDL	FG	TG	CAD	Stroke
BMI		0	1	1	1	0
LDL	0.01		0	0.01	1	0
FG	0	0		0	1	0
TG	0	0.01	0.01		0	0
CAD	0	0	0	0		1
Stroke	0	0.01	0	0	0	

(a) Graph-MRcML-d0

	BMI	LDL	FG	TG	CAD	Stroke
BMI		0	1	1	1	0
LDL	0.01		0	0.01	1	0
FG	0	0		0	1	0
TG	0	0.01	0.01		0	0
CAD	0	0	0	0		1
Stroke	0	0.01	0	0	0	

(b) Graph-MRcML-d1

A.4.5.2.2 Set-up (b) In the presence of a cycle in the direct graph, the diagonal elements of the true total graph were not all zeros, and incorrectly specifying them to zeros led to an incorrect direct graph given in Table A.13b. We can see that, as the sample size increased, the estimates from Graph-MRcML-d0 (Table A.23a to Table A.25a) approached the incorrect direct graph Table A.13b. But the type-I error was controlled in the simulation as the sample size increased, probably because the incorrectly inferred direct effects were small (e.g. LDL $\rightarrow$ AF and FG $\rightarrow$ AF both had -0.001 in Table A.13b). On the other hand, the iterative algorithm in Graph-MRcML-d1 converged successfully, and its estimates approached the true direct graph as the sample size increased. Type-I error was also well-controlled as the sample size increased.

Table A.23: Set-up (b): mean estimated direct graph by (a) Graph-MRcML-d0, and (b) Graph-MRcML-d1 across 100 replicates when $N = 300\,000$. Numbers are rounded to 3 decimal places.

	BMI	LDL	FG	TG	CAD	AF
BMI	0	0	0.081	0.23	0.224	0.32
LDL	0	0	0	-0.001	0.427	0.005
FG	0.001	0.001	0	0.002	0.336	-0.001
TG	-0.001	0	0	0	-0.006	0.02
CAD	0	0	0	0	-0.018	0.176
AF	0	0	0	0	0.104	-0.018

(a) Graph-MRcML-d0

	BMI	LDL	FG	TG	CAD	AF
BMI	0	0	0.081	0.23	0.22	0.315
LDL	0	0	0	-0.001	0.42	0.006
FG	0.001	0.001	0	0.002	0.33	0
TG	-0.001	0	0	0	-0.006	0.02
CAD	0	0	0	0	0	0.17
AF	0	0	0	0	0.1	0

(b) Graph-MRcML-d1

Table A.24: Set-up (b): mean estimated direct graph by (a) Graph-MRcML-d0, and (b) Graph-MRcML-d1 across 100 replicates when $N = 500\,000$. Numbers are rounded to 3 decimal places.

	BMI	LDL	FG	TG	CAD	AF
BMI	0	0	0.081	0.23	0.223	0.321
LDL	0	0	0	0	0.427	0.004
FG	0.001	0.001	0	0.001	0.336	-0.001
TG	-0.001	0	0	0	-0.003	0.015
CAD	0	0	0	0	-0.018	0.176
AF	0	0	0	0	0.104	-0.018

(a) Graph-MRcML-d0

	BMI	LDL	FG	TG	CAD	AF
BMI	0	0	0.081	0.23	0.22	0.316
LDL	0	0	0	0	0.42	0.005
FG	0.001	0.001	0	0.001	0.33	0
TG	-0.001	0	0	0	-0.003	0.015
CAD	0	0	0	0	0	0.17
AF	0	0	0	0	0.1	0

(b) Graph-MRcML-d1

Table A.25: Set-up (b): mean estimated direct graph by (a) Graph-MRcML-d0, and (b) Graph-MRcML-d1 across 100 replicates when $N = 1\,000\,000$. Numbers are rounded to 3 decimal places.

	BMI	LDL	FG	TG	CAD	AF
BMI	0	0	0.081	0.23	0.223	0.322
LDL	0	0	0	0	0.427	0.002
FG	0.001	0.001	0	0.001	0.336	-0.001
TG	0	0	0	0	-0.001	0.01
CAD	0	0	0	0	-0.018	0.176
AF	0	0	0	0	0.104	-0.018

(a) Graph-MRcML-d0

	BMI	LDL	FG	TG	CAD	AF
BMI	0	0	0.081	0.23	0.22	0.318
LDL	0	0	0	0	0.42	0.003
FG	0.001	0.001	0	0.001	0.33	0
TG	0	0	0	0	-0.001	0.009
CAD	0	0	0	0	0	0.17
AF	0	0	0	0	0.1	0

(b) Graph-MRcML-d1

Table A.26: Set-up (b): empirical type-I error and power by (a) Graph-MRcML-d0, and (b) Graph-MRcML-d1 when $N = 300\,000$. Numbers underlined correspond to power.

	BMI	LDL	FG	TG	CAD	AF
BMI		0	1	1	1	1
LDL	0		0	0	1	0.02
FG	0	0		0	1	0
TG	0	0	0		0	0.17
CAD	0	0	0	0		1
AF	0	0	0	0	1	

(a) Graph-MRcML-d0

	BMI	LDL	FG	TG	CAD	AF
BMI		0	1	1	1	1
LDL	0		0	0	1	0.12
FG	0	0		0	1	0
TG	0	0	0		0	0.17
CAD	0	0	0	0		1
AF	0	0	0	0	1	

(b) Graph-MRcML-d1

Table A.27: Set-up (b): empirical type-I error and power by (a) Graph-MRcML-d0, and (b) Graph-MRcML-d1 when $N = 500\,000$. Numbers underlined correspond to power.

	BMI	LDL	FG	TG	CAD	AF
BMI		0	1	1	1	1
LDL	0.01		0	0	1	0.01
FG	0	0		0.01	1	0
TG	0	0	0		0	0.09
CAD	0	0	0	0		1
AF	0	0	0	0	1	

(a) Graph-MRcML-d0

	BMI	LDL	FG	TG	CAD	AF
BMI		0	1	1	1	1
LDL	0.01		0	0	1	0.11
FG	0	0		0.01	1	0
TG	0	0	0		0	0.09
CAD	0	0	0	0		1
AF	0	0	0	0	1	

(b) Graph-MRcML-d1

Table A.28: Set-up (b): empirical type-I error and power by (a) Graph-MRcML-d0, and (b) Graph-MRcML-d1 when $N = 1\,000\,000$. Numbers underlined correspond to power.

	BMI	LDL	FG	TG	CAD	AF
BMI		0	1	1	1	1
LDL	0.01		0	0.01	1	0.01
FG	0	0		0	1	0
TG	0	0	0		0	0.06
CAD	0	0	0	0		1
AF	0	0.01	0	0	1	

(a) Graph-MRcML-d0

	BMI	LDL	FG	TG	CAD	AF
BMI		0	1	1	1	1
LDL	0.01		0	0.01	1	0.03
FG	0	0		0	1	0
TG	0	0	0		0	0.06
CAD	0	0	0	0		1
AF	0	0.01	0	0	1	

(b) Graph-MRcML-d1

A.4.5.3 Results with more IVs for FG

In our real data analysis, the number of IVs used in each MR analysis ranged from 6 to 374, with fasting glucose (FG) having the least IVs. As shown in Xue et al. (2021), using more (valid) IVs in the MR-cML analysis will yield a more precise causal effect estimate (in the first step of Graph-MRcML). And this in turn will affect the direct effect estimates as well. In this section, we performed additional simulations based on Set-up (a) to study the influence of increasing number of IVs for FG in Graph-MRcML, especially in the second step. In Set-up (a), the number of IVs for FG ranged from 7 to 12 in the MR analyses. We added 50 extra IVs for FG (i.e., SNPs only significantly associated with FG), and performed the simulation 100 times with $N = 1\,000\,000$. We compared the standard deviations of the causal network estimates with and without the extra IVs for FG. For the estimated total network in step one (Table A.29), as expected, only the row corresponding to FG (i.e., FG being the exposure) showed more precise estimates when more IVs for FG were used (right panel), while for other total effect estimates their precision remained the same. For the estimated direct network (Table A.30), the row corresponding to FG also gave more precise estimates, but not for other direct effect estimates.

Table A.29: Set-up (a): standard deviations ($\times 10^{-3}$) of estimated total networks across 100 replicates with 0 extra IV (left) and 50 extra IVs (right) for FG.

	BMI	LDL	FG	TG	CAD	Stroke
BMI		4.0	4.4	4.0	5.1	3.8
LDL	1.1		1.2	1.3	1.6	2.1
FG	7.0	7.0		6.9	7.6	6.5
TG	2.3	3.6	2.3		3.9	5.2
CAD	1.3	2.4	1.6	1.5		1.2
Stroke	5.4	4.5	4.5	4.5	4.4	

	BMI	LDL	FG	TG	CAD	Stroke
BMI		4.0	4.4	4.0	5.1	3.8
LDL	1.1		1.2	1.3	1.6	2.1
FG	2.6	2.4		3.0	2.6	2.3
TG	2.3	3.6	2.3		3.9	5.2
CAD	1.3	2.4	1.6	1.5		1.2
Stroke	5.4	4.5	4.5	4.5	4.4	

Table A.30: Set-up (a): standard deviations ($\times 10^{-3}$) of estimated direct networks across 100 replicates with 0 extra IV (left) and 50 extra IVs (right) for FG.

	BMI	LDL	FG	TG	CAD	Stroke
BMI		4.4	4.4	4.2	5.8	4.4
LDL	1.0		1.3	1.5	1.7	2.1
FG	6.9	7.0		7.1	8.6	6.8
TG	2.3	3.6	2.3		4.5	5.2
CAD	1.7	2.5	1.7	1.6		1.3
Stroke	5.4	4.5	4.5	4.4	4.9	

	BMI	LDL	FG	TG	CAD	Stroke
BMI		4.3	4.4	4.1	5.8	4.5
LDL	1.0		1.3	1.5	1.7	2.1
FG	2.6	2.6		3.1	2.9	2.4
TG	2.3	3.6	2.3		4.5	5.2
CAD	1.7	2.5	1.7	1.6		1.3
Stroke	5.4	4.5	4.5	4.4	4.9	

For the purpose of simulation, we also considered the scenario with a larger BMI $\rightarrow$ FG effect of 1, which originally was 0.07 in Set-up (a). We performed the similar analysis and compared the standard deviations of estimated causal networks with and without extra IVs for FG. Again, the precision of the total effect estimates remained the same between the left and right tables in Table A.31 except for those using FG as the exposure. But for the estimated direct network (Table A.32), the rows corresponding to both FG and BMI showed more precise estimates. This result was not completely surprising. Consider a relevant but simpler example of mediation analysis $A \rightarrow B \rightarrow C$ (e.g., BMI, FG and CAD being traits A, B, C respectively). Denote x the direct effect estimate of $A \rightarrow B$, y the direct effect estimate of $B \rightarrow C$, z the total effect estimate of $A \rightarrow C$. Then the direct effect estimate of $A \rightarrow C$ is $z - xy$. Assuming independence among x, y, z, then $\text{Var}(z - xy) = \text{Var}(z) + \text{Var}(xy) = \text{Var}(z) + \text{Var}(x)E(y)^2 + \text{Var}(y)E(x)^2 + \text{Var}(x)\text{Var}(y)$. Adding extra IVs for FG reduced $\text{Var}(y)$, and when $E(x)$ became larger, the term $\text{Var}(y)E(x)^2$ contributed more in the reduction of $\text{Var}(z - xy)$. When $E(x) = 0.07$, this term became negligible and we didn't observe any increase in the precision of the direct effect estimates of BMI to other traits as shown in Table A.30; when $E(x) = 1$, we observed more precise estimates in the row corresponding to BMI in Table A.32. In general, when we have more (valid) IVs for a trait, the precision of the direct effect estimates starting from that trait is likely to increase, while the precision for other direct effect estimates may

or may not change much, depending on the underlying relationship among the traits.

Table A.31: Set-up (a) with a larger BMI → FG effect: standard deviations ($\times 10^{-3}$) of estimated total networks across 100 replicates with 0 extra IV (left) and 50 extra IVs (right) for FG.

	BMI	LDL	FG	TG	CAD	Stroke
BMI		4.0	5.8	4.0	5.6	3.9
LDL	1.1		1.2	1.3	1.6	2.1
FG	7.0	7.0		6.9	7.6	6.5
TG	2.3	3.6	2.3		3.9	5.2
CAD	1.3	2.4	1.6	1.5		1.2
Stroke	5.4	4.5	4.5	4.5	4.4	

	BMI	LDL	FG	TG	CAD	Stroke
BMI		4.0	5.8	4.0	5.6	3.9
LDL	1.1		1.2	1.3	1.6	2.1
FG	2.6	2.4		3.0	2.6	2.3
TG	2.3	3.6	2.3		3.9	5.2
CAD	1.3	2.4	1.6	1.5		1.2
Stroke	5.4	4.5	4.5	4.5	4.4	

Table A.32: Set-up (a) with a larger BMI → FG effect: standard deviations ($\times 10^{-3}$) of estimated direct networks across 100 replicates with 0 extra IV (left) and 50 extra IVs (right) for FG.

	BMI	LDL	FG	TG	CAD	Stroke
BMI		8.3	9.5	8.5	11.1	7.7
LDL	1.0		1.7	1.5	1.7	2.1
FG	6.9	7.0		7.1	9.2	6.8
TG	2.3	3.6	3.3		4.5	5.2
CAD	1.7	2.5	2.4	1.6		1.3
Stroke	5.4	4.5	6.6	4.4	5.0	

	BMI	LDL	FG	TG	CAD	Stroke
BMI		5.0	6.2	5.1	7.3	5.4
LDL	1.0		1.7	1.5	1.7	2.1
FG	2.6	2.6		3.1	3.1	2.4
TG	2.3	3.6	3.3		4.5	5.2
CAD	1.7	2.5	2.4	1.6		1.3
Stroke	5.4	4.5	6.7	4.4	4.9	

A.5 More results from the real data analysis

A.5.1 GWAS summary data

Table A.33: 17 GWAS summary data used in the real data analysis.

GWAS Trait	ID in IEU data base	Reference
LDL	ebi-a-GCST002222	(Willer et al., 2013)
HDL	ebi-a-GCST002223	(Willer et al., 2013)
Triglycerides	ebi-a-GCST002216	(Willer et al., 2013)
Height	ieu-a-89	(Wood et al., 2014)
BMI	ieu-a-835	(Locke et al., 2015)
Birth weight	ieu-a-1083	(Horikoshi et al., 2016)
DBP	ukb-a-359	Neale Lab
SBP	ukb-a-360	Neale Lab
Fasting glucose	ebi-a-GCST000568	(Dupuis et al., 2010)
Smoke	ieu-b-25	(Liu et al., 2019)
Alcohol	ieu-b-73	(Liu et al., 2019)
CAD	ebi-a-GCST005195	(van der Harst and Verweij, 2018)
Stroke	ebi-a-GCST005838	(Malik et al., 2018)
T2D	ieu-a-26	(Morris et al., 2012)
Asthma	ebi-a-GCST006862	(Demenais et al., 2018)
AFib	ebi-a-GCST006414	(Nielsen et al., 2018)
Alzheimer's disease	Not available	(Jansen et al., 2019)

A.5.2 Results by Graph-MRcML-d0

As mentioned in the main text, given the presence of potential cycles in the underlying direct graph, the diagonal elements of the total graph may not be all zeros. So applying Graph-MRcML-d0 might be problematic. In particular, the final estimated direct graph by Graph-MRcML-d0 among the 17 traits had a spectral radius greater than one (so did every estimate from the 2000 perturbed datasets), violating

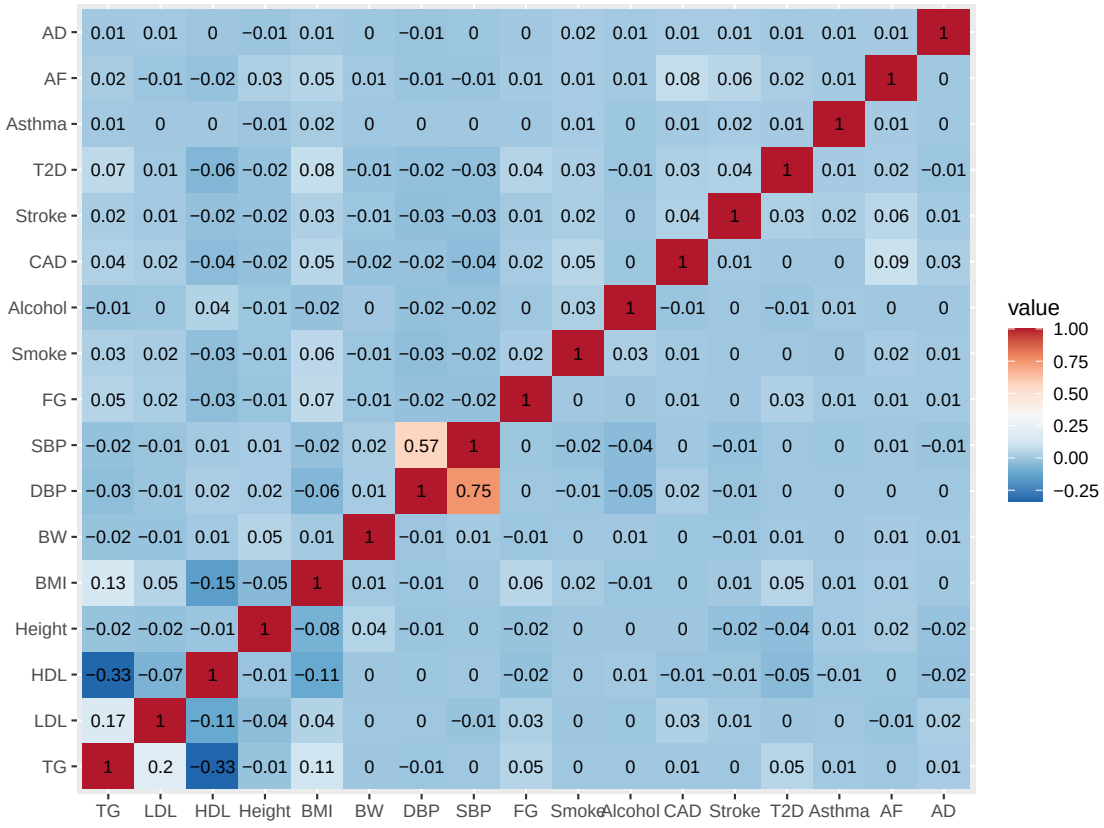


Figure A.18: The estimated correlation matrix $\mathbf{P}$ for the 17 traits by bivariate LDSC (lower triangle) and by null Z-scores (upper triangle).

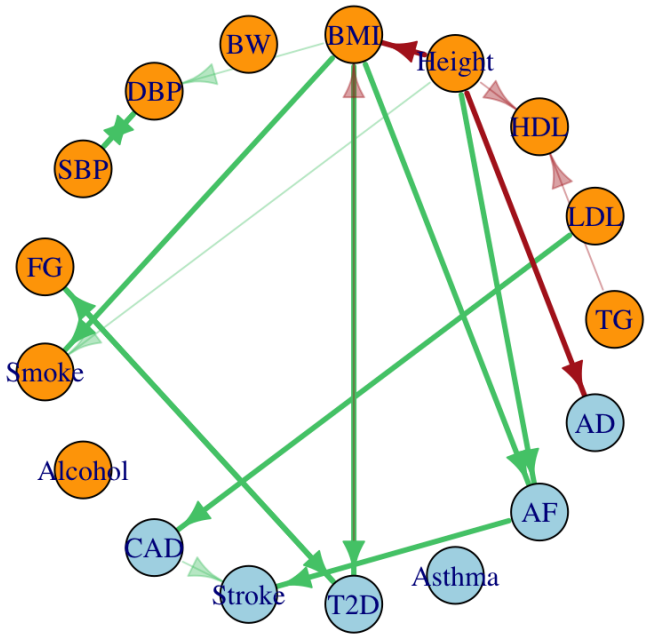


Figure A.19: Estimated direct graph by Graph-MRcML-d0. The spectral radius of the graph was greater than one. The dark-solid edges are identified at the Bonferroni-adjusted significance level, while the light-colored ones are marginally significant at a less stringent level of 6.5e-3.

the key assumption of the graph deconvolution algorithm, thus the result shown in Figure A.19 was not trustworthy. We found that when we excluded traits with potential cycles, e.g., one of the two blood pressures traits (SBP and DBP), one of the two lipid traits (HDL and TG), and T2D, this spectral radius problem was much alleviated. For example, if we removed SBP, TG and T2D from the graph, the spectral radius of the final estimated direct graph among the 14 traits by Graph-MRcML-d0 was smaller than one, and so were those from all 2000 estimates. Furthermore, as shown in Figure A.20, the estimated direct graphs by Graph-MRcML-d0 (left) and Graph-MRcML-d1 (right) had similar structures. Again, we would like to point out that the problem was perhaps due to that Graph-MRcML-d0 simply specified the diagonal elements of the total graph to zeros while there were cycles in the underlying direct graph estimate (e.g. DBP $\leftrightarrow$ SBP, HDL $\leftrightarrow$ TG and T2D $\leftrightarrow$ FG, etc.). At the

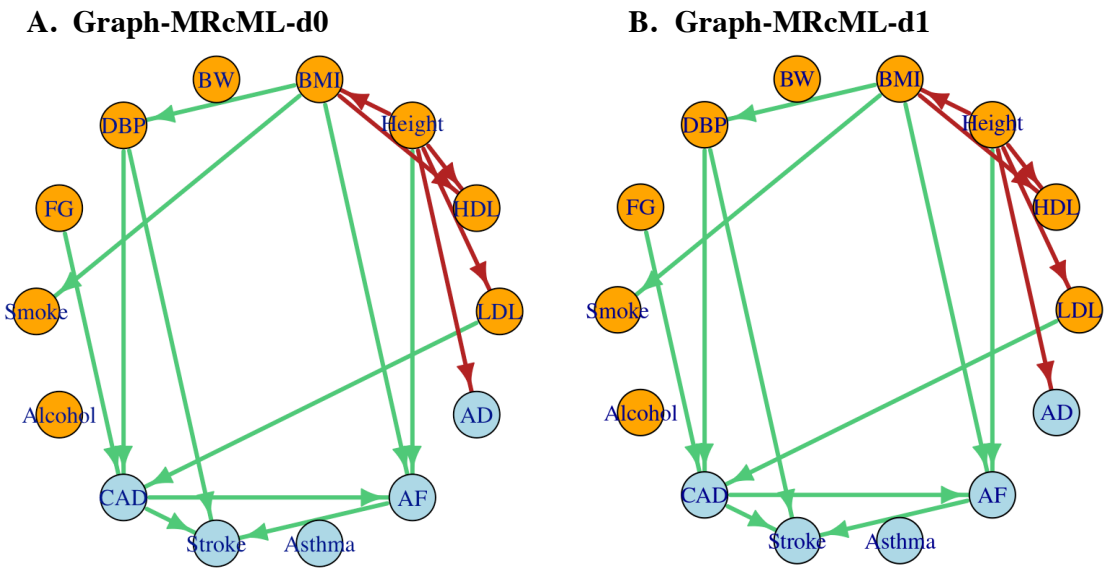


Figure A.20: Estimated direct graphs by Graph-MRcML-d0 (left) and Graph-MRcML-d1 (right) among 14 traits (after excluding SBP, TG and T2D).

same time, it is also possible that the bi-directional MR analysis between DBP and SBP (and between HDL and TG, etc.) was not reliable as there were many shared common IVs between the two highly related traits, many of which might be invalid IVs that MR-cML-BIC-C failed to select out, leading to biased estimation.

A.5.3 More results on MVMR analysis

Table A.34: Results for MVMR analysis estimating direct effect of 15 exposures (excluding AD) on CAD.

			MVMR-IVW			MVMR-Robust		
Trait	F-statistics	#snps	effect size	SE	pval	effect size	SE	pval
FG	2.74	6	0.13	0.09	1.4e-1	0.14	0.10	1.5e-1
TG	10.87	22	0.15	0.05	2.0e-3	0.14	0.04	6.6e-4
LDL	9.98	23	0.38	0.04	1.4e-23	0.36	0.07	2.6e-7
HDL	19.33	39	-0.03	0.03	4.3e-1	-0.03	0.04	5.1e-1
Stroke	1.71	6	0.23	0.05	1.8e-5	0.18	0.06	2.2e-3
AF	6.29	29	0.11	0.02	2.0e-5	0.10	0.03	4.1e-4
Asthma	3.09	6	-0.01	0.03	7.9e-1	0.01	0.02	8.0e-1
BW	3.03	10	-0.06	0.09	4.9e-1	-0.04	0.07	5.9e-1
T2D	1.61	1	0.00	0.02	9.6e-1	0.00	0.02	8.7e-1
BMI	4.74	18	0.07	0.08	3.7e-1	0.12	0.07	6.1e-2
Height	7.64	194	-0.04	0.03	1.8e-1	-0.05	0.03	1.5e-1
Smoke	4.67	5	-0.00	0.05	1.0e+0	0.02	0.09	8.7e-1
Alcohol	3.00	6	-0.13	0.18	4.5e-1	-0.21	0.16	1.7e-1
DBP	9.66	35	0.27	0.11	1.2e-2	0.09	0.14	5.1e-1
SBP	5.85	33	0.15	0.12	2.4e-1	0.26	0.15	8.1e-2

Table A.35: Results for MVMR analysis estimating direct effect of 5 exposures on CAD.

			MVMR-IVW			MVMR-Robust		
Trait	F-statistics	#snps	effect size	SE	pval	effect size	SE	pval
TG	18.80	22	0.18	0.04	4.2e-5	0.18	0.03	1.5e-8
LDL	20.33	28	0.41	0.04	7.2e-27	0.41	0.08	6.3e-8
BMI	7.71	23	0.13	0.07	7.9e-2	0.20	0.06	6.0e-4
Height	36.50	245	-0.02	0.03	5.4e-1	-0.02	0.02	3.2e-1
SBP	7.95	34	0.51	0.09	3.2e-8	0.46	0.12	8.4e-5

A.5.4 Relationships among HDL, TG and glycemic traits

As pointed out by one reviewer, a recent study (Zhu et al., 2022) identified strong causal effects of fasting insulin (FI) on HDL and TG, while we didn't observe any

direct effect of fasting glucose (FG) on TG and HDL as shown in Fig. 5B in the main text, and in a smaller network with the three traits of interest Figure A.21A. With an estimated number of effective test of 5, we identified a bi-directional effect between HDL and TG, and a negative effect of HDL on FG, at the Bonferroni-adjusted significance level (p-value < 0.01). We further performed analysis using FI, instead of FG, to study the causal relationships among HDL, TG and FI. Following Zhu et al. (2022), we obtained the GWAS summary data for FI from Chen et al. (2021). As shown in Figure A.21B, we observed a significant positive effect of FI on TG and a negative effect on HDL, consistent with the findings in Figure 4B of Zhu et al. (2022). We didn't identify any direct effect from HDL or TG to FI, while Zhu et al. (2022) also observed direct effects from HDL and TG to FI, though with a much smaller magnitude of effect size than those in the reverse direction (Figure 4A in Zhu et al. (2022)). This may partly due to the different GWAS summary data (with a smaller sample size) we used for lipid traits and different analysis approaches. We note that the FI GWAS summary data provided by Chen et al. (2021) was adjusted for BMI, and using such GWAS data may introduce bias in the MR estimates (Hartwig et al., 2021). To avoid such potential bias, we conducted an analysis using FI GWAS summary data without adjusting for BMI (Scott et al., 2012). As shown in Figure A.21C, the positive direct effect of FI on TG stayed significant (after the Bonferroni adjustment) while the negative effect of FI on HDL became only marginally significant with p-value $\approx$ 0.021. In summary, given the complex relationships between glycemic traits and lipid traits, more investigations are warranted.

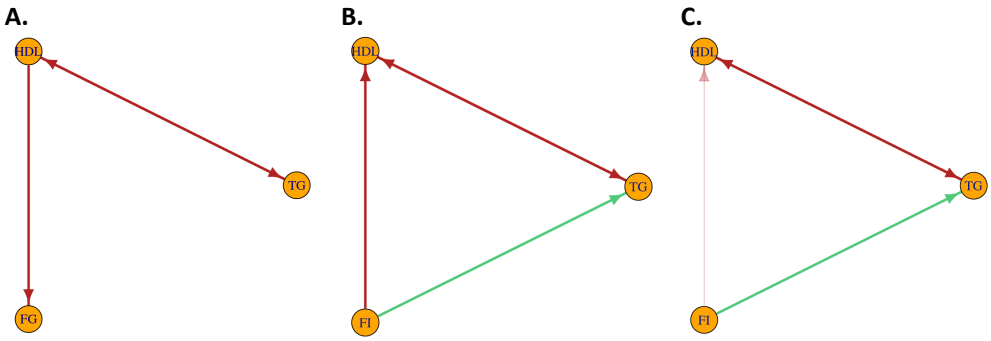


Figure A.21: Estimated direct causal networks for HDL, TG and (A) FG; (B) FI using BMI adjusted FI GWAS; (C) FI. The dark-solid edges are identified at the Bonferroni-adjusted significance level, while the light-colored ones are marginally significant at the nominal level of 0.05.

Appendix B

Supplementary material for Chapter 3

B.1 Supplemental figures

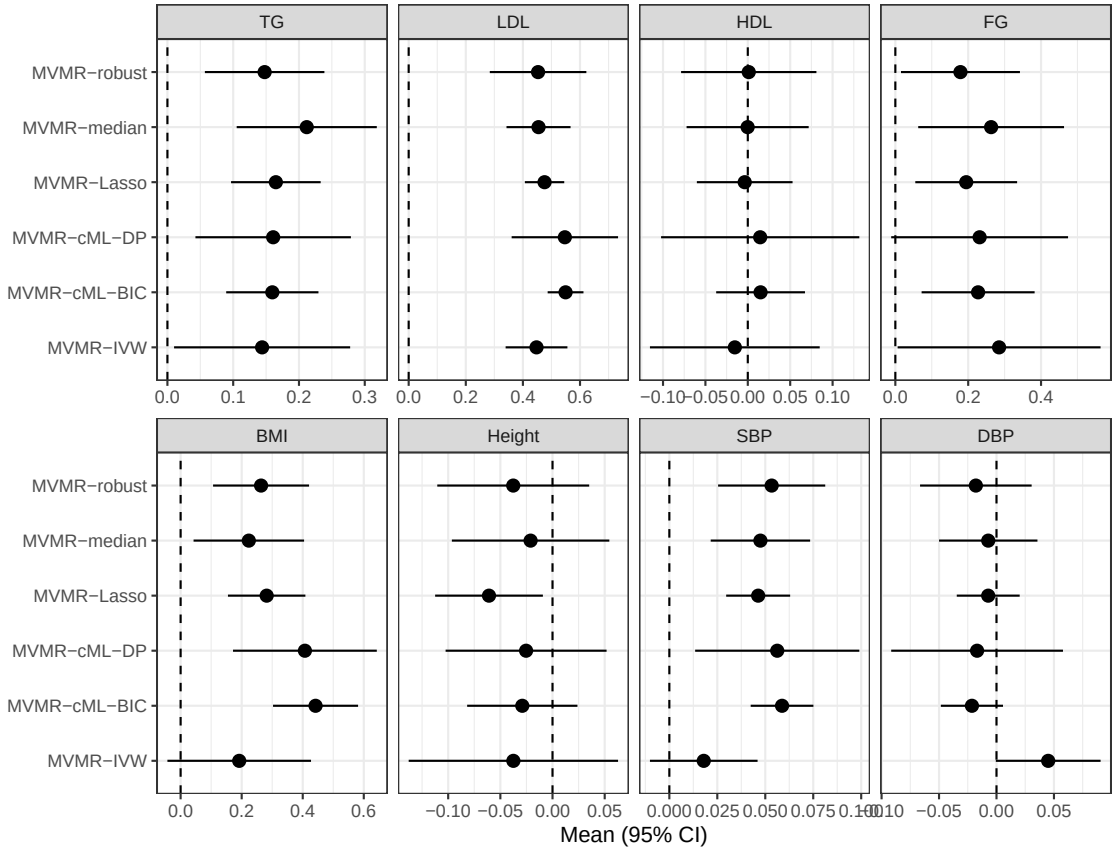


Figure B.1: **Real data analysis results**. The estimated effects (and 95% confidence intervals) of each of the 8 risk factors on CAD by various MVMR methods.

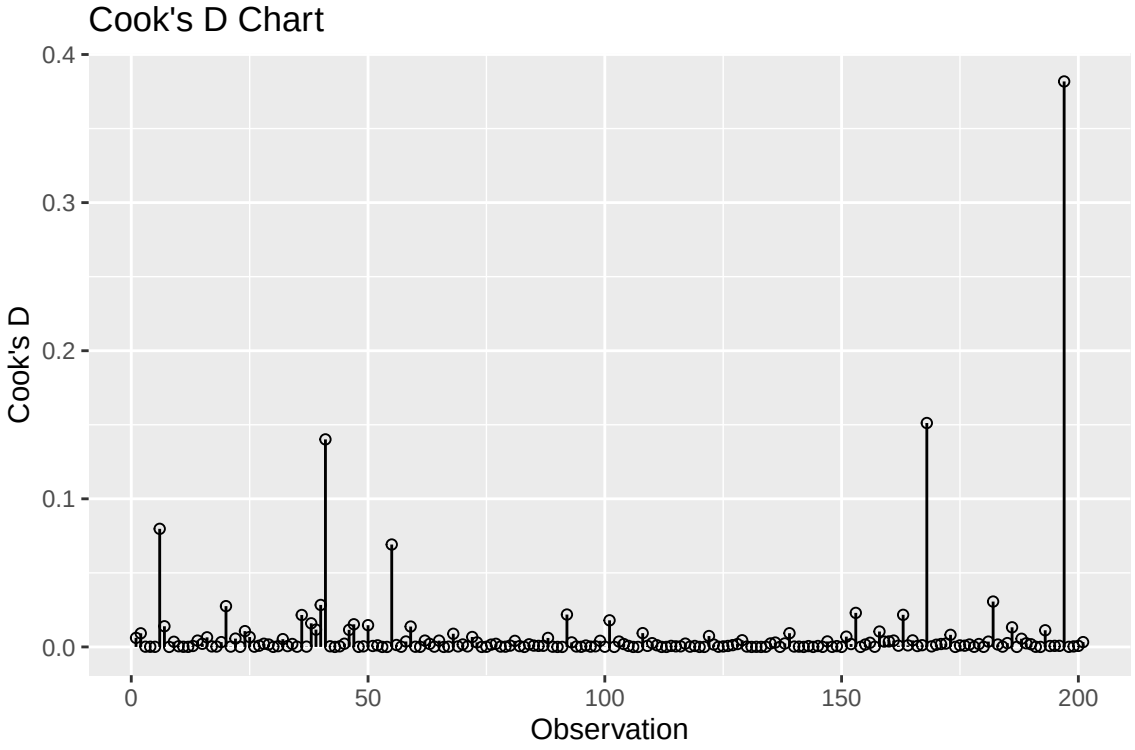


Figure B.2: **Cook's distance applied to MVMR-IVW.** The suggested cut-off for Cook'sdistance is 0.92.

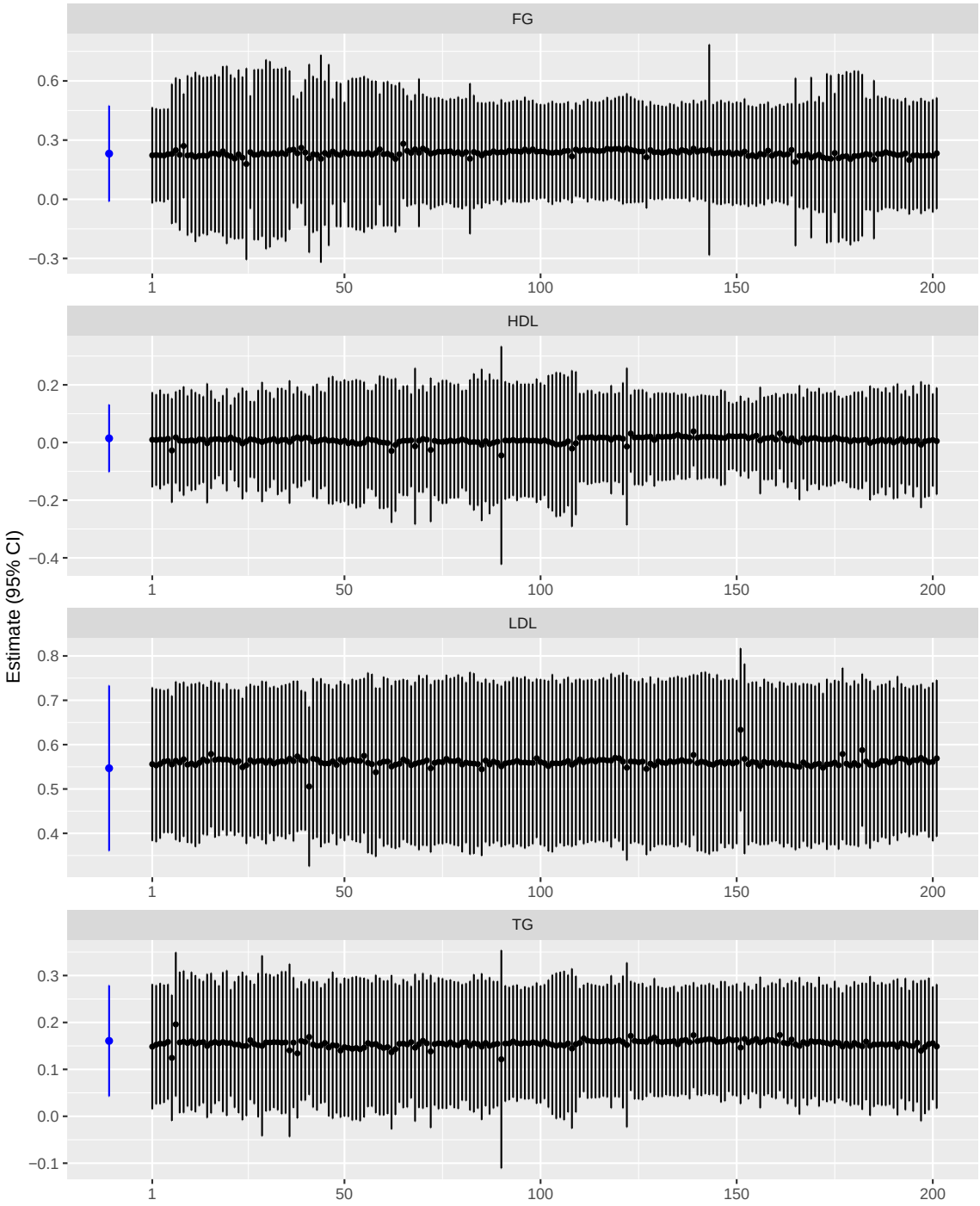


Figure B.3: **Leave-one-out analysis with MVMR-cML-DP**. Traits from top to bottom: FG, HDL, HDL and TG.

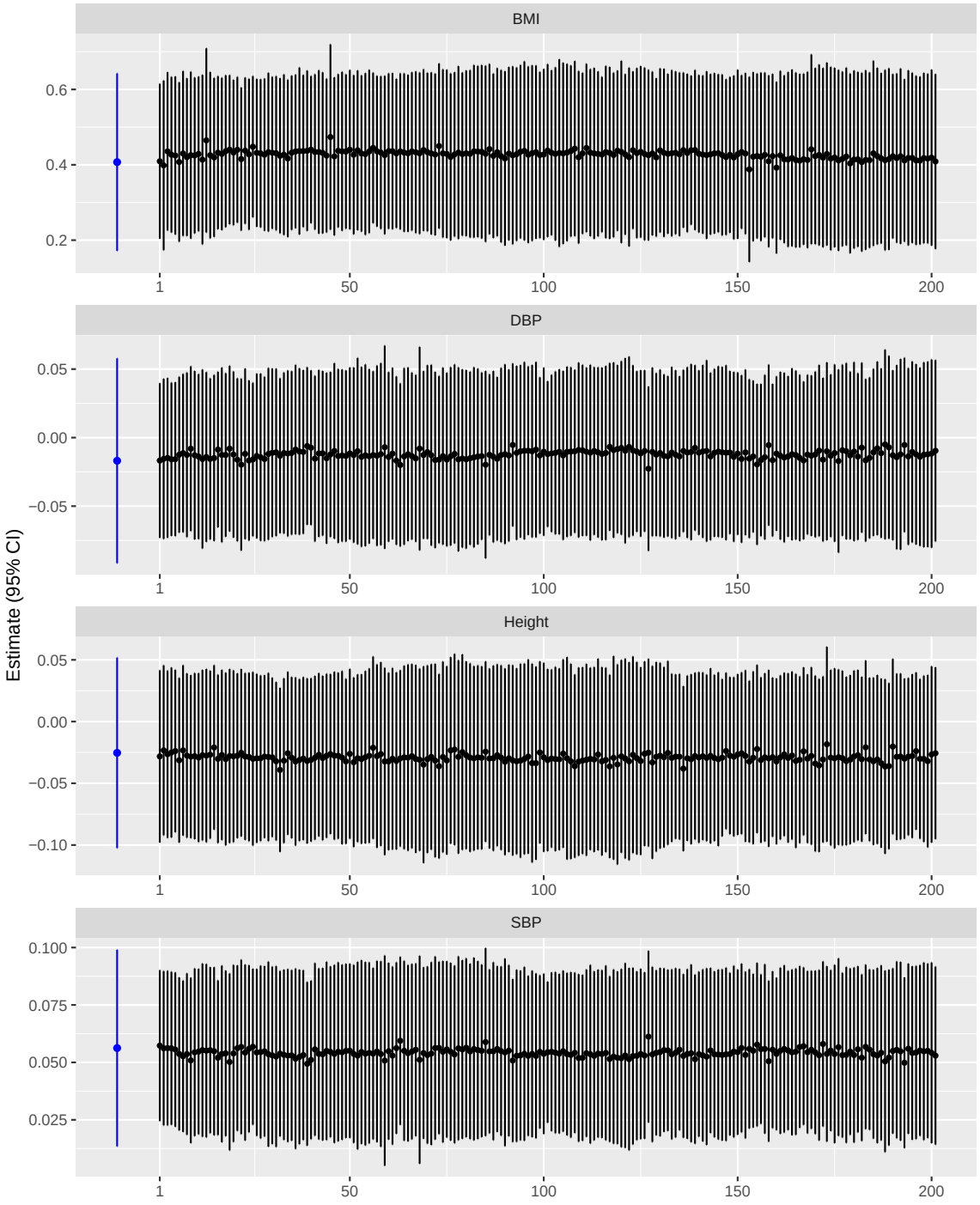


Figure B.4: **Leave-one-out analysis with MVMR-cML-DP**. Traits from top to bottom: BMI, DBP, Height and SBP.

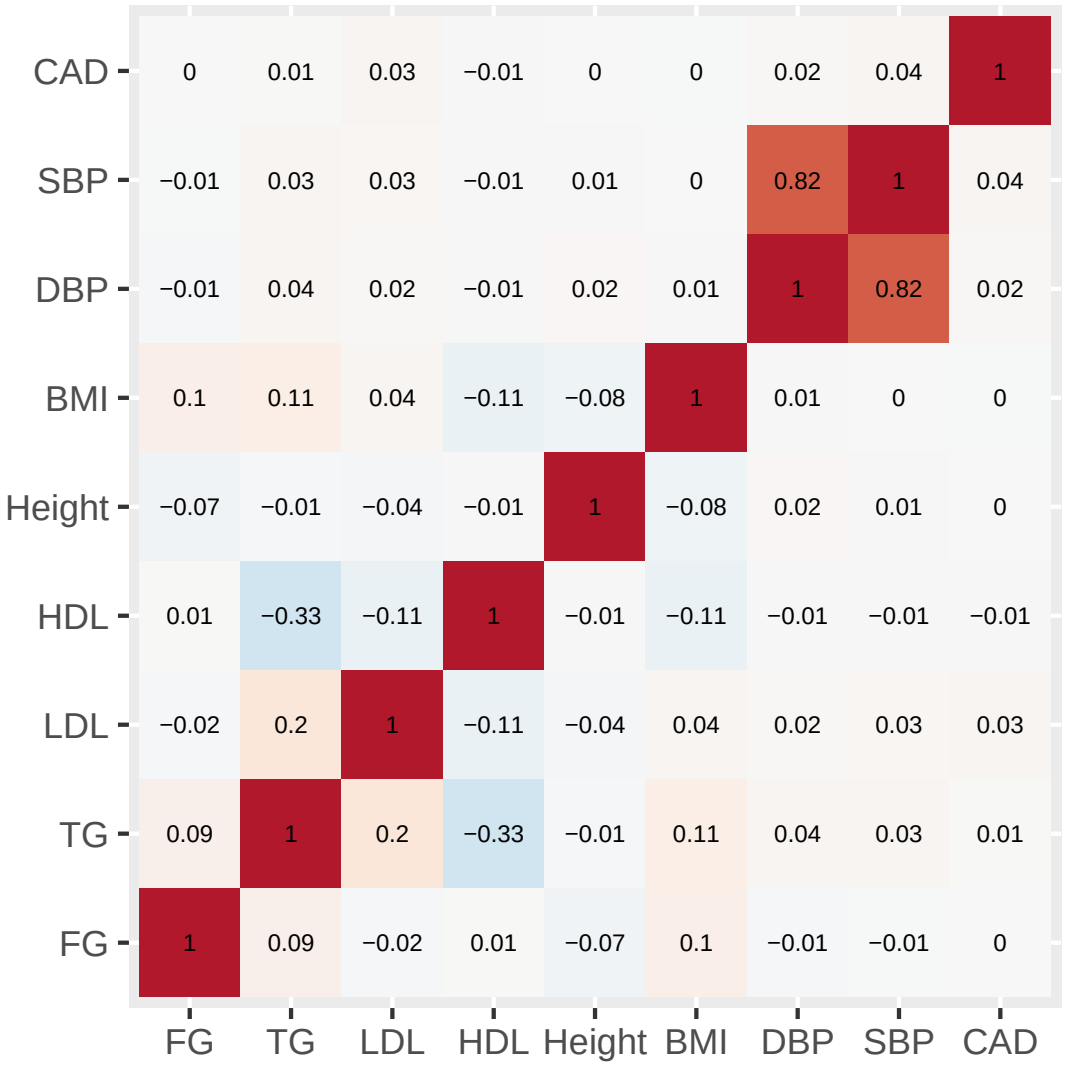


Figure B.5: The estimated correlation matrix for the 9 GWAS data by bivariate LDSC.

B.2 Supplemental tables

	30% invalid						50% invalid					
Method	Mean	SD	SE	Cov	Power	MSE	Mean	SD	SE	Cov	Power	MSE
Scenario 1: Balanced pleiotropy, InSIDE met												
MVMR-cML-BIC	0.098	0.076	0.054	0.872	0.496	0.006	0.101	0.237	0.064	0.688	0.502	0.056
MVMR-cML-DP	0.096	0.076	0.083	0.966	0.292	0.006	0.094	0.179	0.189	0.960	0.120	0.032
MVMR-Egger	0.106	0.441	0.412	0.922	0.078	0.195	0.075	0.574	0.537	0.912	0.074	0.330
MVMR-IVW	0.104	0.390	0.359	0.920	0.084	0.152	0.094	0.500	0.473	0.940	0.064	0.250
MVMR-Lasso	0.099	0.092	0.058	0.862	0.484	0.008	0.099	0.262	0.084	0.670	0.510	0.069
MVMR-median	0.097	0.113	0.084	0.914	0.300	0.013	0.091	0.279	0.122	0.736	0.346	0.078
MVMR-robust	0.101	0.097	0.100	0.929	0.351	0.009	0.092	0.371	0.399	0.925	0.100	0.138
Scenario 2: Directional pleiotropy, InSIDE met												
MVMR-cML-BIC	0.096	0.080	0.055	0.864	0.438	0.006	0.100	0.243	0.066	0.706	0.432	0.059
MVMR-cML-DP	0.098	0.073	0.085	0.974	0.216	0.005	0.100	0.193	0.173	0.928	0.150	0.037
MVMR-Egger	0.110	0.489	0.467	0.946	0.056	0.240	0.152	0.623	0.576	0.920	0.092	0.391
MVMR-IVW	0.164	0.422	0.411	0.928	0.092	0.182	0.237	0.551	0.503	0.926	0.104	0.322
MVMR-Lasso	0.109	0.121	0.064	0.878	0.440	0.015	0.149	0.366	0.091	0.582	0.488	0.136
MVMR-median	0.113	0.155	0.091	0.898	0.320	0.024	0.148	0.379	0.138	0.678	0.390	0.146
MVMR-robust	0.099	0.102	0.100	0.926	0.316	0.010	0.145	0.418	0.460	0.929	0.090	0.176
Scenario 3: Directional pleiotropy, InSIDE violated												
MVMR-cML-BIC	0.102	0.076	0.054	0.880	0.490	0.006	0.104	0.203	0.065	0.718	0.442	0.041
MVMR-cML-DP	0.101	0.072	0.083	0.978	0.294	0.005	0.105	0.142	0.179	0.962	0.108	0.020
MVMR-Egger	0.138	0.498	0.442	0.910	0.102	0.249	0.189	0.584	0.552	0.918	0.104	0.349
MVMR-IVW	0.133	0.427	0.382	0.922	0.108	0.184	0.189	0.483	0.481	0.932	0.082	0.241
MVMR-Lasso	0.105	0.124	0.059	0.872	0.486	0.015	0.141	0.267	0.086	0.696	0.488	0.073
MVMR-median	0.105	0.148	0.086	0.906	0.308	0.022	0.143	0.290	0.129	0.758	0.334	0.086
MVMR-robust	0.102	0.102	0.101	0.905	0.345	0.010	0.151	0.366	0.391	0.927	0.102	0.137

Table B.1: **Simulation results for θ_2 when $\boldsymbol{\theta} = (\mathbf{0.2}, \mathbf{0.1}, \mathbf{0.3}, \mathbf{0.4})^{\mathbf{T}}$.** Mean and standard deviation (SD) of estimates, mean standard error (SE) and coverage rate (Cov), power, mean squared error (MSE) when $\theta_2 = 0.1$.

	30% invalid						50% invalid					
Method	Mean	SD	SE	Cov	Power	MSE	Mean	SD	SE	Cov	Power	MSE
Scenario 1: Balanced pleiotropy, InSIDE met												
MVMR-cML-BIC	0.303	0.076	0.053	0.854	0.984	0.006	0.303	0.200	0.063	0.692	0.934	0.040
MVMR-cML-DP	0.303	0.074	0.083	0.964	0.908	0.006	0.298	0.168	0.191	0.968	0.526	0.028
MVMR-Egger	0.296	0.429	0.409	0.930	0.150	0.184	0.290	0.565	0.542	0.942	0.106	0.319
MVMR-IVW	0.291	0.375	0.359	0.936	0.180	0.141	0.304	0.494	0.471	0.940	0.136	0.245
MVMR-Lasso	0.304	0.100	0.058	0.846	0.974	0.010	0.287	0.285	0.084	0.644	0.840	0.081
MVMR-median	0.301	0.119	0.083	0.918	0.906	0.014	0.285	0.303	0.121	0.724	0.738	0.092
MVMR-robust	0.303	0.093	0.099	0.907	0.841	0.009	0.305	0.372	0.399	0.917	0.250	0.139
Scenario 2: Directional pleiotropy, InSIDE met												
MVMR-cML-BIC	0.302	0.076	0.055	0.860	0.984	0.006	0.295	0.193	0.065	0.694	0.912	0.037
MVMR-cML-DP	0.302	0.074	0.083	0.964	0.916	0.006	0.297	0.166	0.169	0.942	0.566	0.028
MVMR-Egger	0.306	0.465	0.466	0.946	0.138	0.216	0.337	0.641	0.580	0.918	0.130	0.413
MVMR-IVW	0.363	0.430	0.411	0.922	0.194	0.189	0.424	0.559	0.506	0.912	0.164	0.328
MVMR-Lasso	0.305	0.143	0.064	0.824	0.946	0.021	0.344	0.392	0.090	0.588	0.812	0.156
MVMR-median	0.314	0.177	0.091	0.878	0.894	0.032	0.354	0.397	0.138	0.672	0.680	0.160
MVMR-robust	0.303	0.108	0.099	0.893	0.829	0.012	0.363	0.409	0.459	0.933	0.225	0.172
Scenario 3: Directional pleiotropy, InSIDE violated												
MVMR-cML-BIC	0.298	0.075	0.054	0.868	0.988	0.006	0.307	0.206	0.065	0.724	0.940	0.043
MVMR-cML-DP	0.296	0.074	0.081	0.968	0.908	0.006	0.299	0.156	0.180	0.966	0.548	0.024
MVMR-Egger	0.356	0.459	0.438	0.938	0.182	0.214	0.338	0.594	0.553	0.920	0.132	0.354
MVMR-IVW	0.350	0.402	0.385	0.952	0.180	0.164	0.332	0.498	0.484	0.932	0.134	0.249
MVMR-Lasso	0.299	0.110	0.059	0.874	0.978	0.012	0.312	0.312	0.087	0.696	0.840	0.098
MVMR-median	0.303	0.126	0.085	0.932	0.902	0.016	0.314	0.337	0.129	0.772	0.714	0.113
MVMR-robust	0.297	0.098	0.094	0.899	0.820	0.010	0.325	0.369	0.389	0.933	0.255	0.137

Table B.2: **Simulation results for θ_3 when $\boldsymbol{\theta} = (\mathbf{0.2}, \mathbf{0.1}, \mathbf{0.3}, \mathbf{0.4})^{\mathbf{T}}$.** Mean and standard deviation (SD) of estimates, mean standard error (SE) and coverage rate (Cov), power, mean squared error (MSE) when $\theta_3 = 0.3$.

	30% invalid						50% invalid					
Method	Mean	SD	SE	Cov	Power	MSE	Mean	SD	SE	Cov	Power	MSE
Scenario 1: Balanced pleiotropy, InSIDE met												
MVMR-cML-BIC	0.401	0.071	0.053	0.874	0.998	0.005	0.384	0.202	0.063	0.670	0.952	0.041
MVMR-cML-DP	0.400	0.070	0.082	0.978	0.974	0.005	0.388	0.159	0.186	0.962	0.668	0.025
MVMR-Egger	0.404	0.424	0.407	0.932	0.196	0.180	0.436	0.538	0.534	0.944	0.156	0.291
MVMR-IVW	0.407	0.385	0.358	0.934	0.266	0.148	0.454	0.468	0.468	0.948	0.196	0.222
MVMR-Lasso	0.402	0.083	0.057	0.892	0.994	0.007	0.416	0.265	0.083	0.642	0.910	0.070
MVMR-median	0.405	0.106	0.083	0.922	0.978	0.011	0.426	0.298	0.120	0.726	0.856	0.090
MVMR-robust	0.398	0.087	0.099	0.933	0.899	0.008	0.426	0.385	0.398	0.938	0.344	0.149
Scenario 2: Directional pleiotropy, InSIDE met												
MVMR-cML-BIC	0.400	0.079	0.055	0.852	0.998	0.006	0.416	0.201	0.067	0.716	0.966	0.041
MVMR-cML-DP	0.399	0.077	0.084	0.956	0.974	0.006	0.408	0.168	0.173	0.960	0.732	0.028
MVMR-Egger	0.414	0.491	0.469	0.946	0.198	0.241	0.391	0.612	0.578	0.918	0.136	0.375
MVMR-IVW	0.470	0.425	0.408	0.928	0.262	0.186	0.473	0.541	0.509	0.930	0.200	0.298
MVMR-Lasso	0.400	0.110	0.063	0.866	0.990	0.012	0.419	0.381	0.093	0.632	0.878	0.145
MVMR-median	0.401	0.140	0.089	0.884	0.946	0.020	0.436	0.394	0.144	0.710	0.800	0.156
MVMR-robust	0.400	0.104	0.102	0.897	0.901	0.011	0.445	0.439	0.475	0.912	0.294	0.195
Scenario 3: Directional pleiotropy, InSIDE violated												
MVMR-cML-BIC	0.403	0.072	0.054	0.876	0.998	0.005	0.390	0.188	0.065	0.738	0.972	0.035
MVMR-cML-DP	0.401	0.072	0.083	0.966	0.966	0.005	0.392	0.155	0.180	0.970	0.672	0.024
MVMR-Egger	0.467	0.455	0.435	0.926	0.244	0.212	0.433	0.588	0.548	0.916	0.182	0.347
MVMR-IVW	0.457	0.392	0.381	0.940	0.274	0.157	0.438	0.504	0.486	0.922	0.192	0.255
MVMR-Lasso	0.403	0.111	0.059	0.872	0.988	0.012	0.411	0.298	0.087	0.650	0.906	0.089
MVMR-median	0.409	0.127	0.087	0.910	0.962	0.016	0.410	0.317	0.128	0.726	0.824	0.101
MVMR-robust	0.402	0.092	0.100	0.917	0.905	0.009	0.405	0.399	0.398	0.919	0.334	0.159

Table B.3: **Simulation results for θ_4 when $\boldsymbol{\theta} = (\mathbf{0.2}, \mathbf{0.1}, \mathbf{0.3}, \mathbf{0.4})^{\mathbf{T}}$.** Mean and standard deviation (SD) of estimates, mean standard error (SE) and coverage rate (Cov), power, mean squared error (MSE) when $\theta_4 = 0.4$.

	30% invalid						50% invalid					
Method	Mean	SD	SE	Cov	Power	MSE	Mean	SD	SE	Cov	Power	MSE
Scenario 1: Balanced pleiotropy, InSIDE met												
MVMR-cML-BIC	-0.102	0.059	0.044	0.884	0.646	0.003	-0.100	0.154	0.054	0.730	0.566	0.024
MVMR-cML-DP	-0.102	0.058	0.069	0.972	0.386	0.003	-0.099	0.121	0.144	0.958	0.148	0.015
MVMR-Egger	-0.092	0.440	0.412	0.920	0.100	0.193	-0.123	0.575	0.537	0.916	0.090	0.331
MVMR-IVW	-0.097	0.390	0.359	0.916	0.102	0.152	-0.106	0.502	0.473	0.936	0.078	0.252
MVMR-Lasso	-0.101	0.082	0.049	0.882	0.582	0.007	-0.104	0.249	0.075	0.676	0.514	0.062
MVMR-median	-0.100	0.097	0.070	0.922	0.382	0.009	-0.108	0.271	0.109	0.732	0.376	0.074
MVMR-robust	-0.101	0.080	0.081	0.913	0.457	0.006	-0.107	0.370	0.406	0.921	0.104	0.137
Scenario 2: Directional pleiotropy, InSIDE met												
MVMR-cML-BIC	-0.105	0.056	0.045	0.908	0.628	0.003	-0.090	0.196	0.055	0.748	0.584	0.038
MVMR-cML-DP	-0.104	0.057	0.069	0.976	0.398	0.003	-0.092	0.146	0.146	0.962	0.198	0.021
MVMR-Egger	-0.087	0.489	0.466	0.952	0.064	0.240	-0.045	0.623	0.576	0.918	0.066	0.391
MVMR-IVW	-0.036	0.422	0.410	0.930	0.062	0.182	0.038	0.550	0.503	0.924	0.064	0.322
MVMR-Lasso	-0.097	0.113	0.054	0.872	0.556	0.013	-0.056	0.354	0.083	0.596	0.554	0.128
MVMR-median	-0.089	0.131	0.076	0.900	0.338	0.017	-0.052	0.370	0.125	0.666	0.394	0.139
MVMR-robust	-0.101	0.076	0.075	0.902	0.454	0.006	-0.056	0.410	0.452	0.929	0.102	0.170
Scenario 3: Directional pleiotropy, InSIDE violated												
MVMR-cML-BIC	-0.100	0.060	0.044	0.894	0.600	0.004	-0.097	0.147	0.053	0.772	0.552	0.022
MVMR-cML-DP	-0.101	0.059	0.067	0.978	0.396	0.003	-0.090	0.117	0.143	0.960	0.152	0.014
MVMR-Egger	-0.061	0.496	0.441	0.910	0.092	0.248	-0.010	0.583	0.552	0.918	0.080	0.348
MVMR-IVW	-0.068	0.426	0.382	0.926	0.082	0.183	-0.012	0.482	0.481	0.928	0.078	0.240
MVMR-Lasso	-0.098	0.099	0.050	0.876	0.570	0.010	-0.067	0.258	0.076	0.702	0.450	0.068
MVMR-median	-0.099	0.129	0.073	0.908	0.370	0.017	-0.058	0.282	0.114	0.754	0.332	0.081
MVMR-robust	-0.099	0.080	0.079	0.907	0.435	0.006	-0.051	0.363	0.405	0.927	0.090	0.134

Table B.4: **Simulation results for θ_2 when $\boldsymbol{\theta} = (\mathbf{0}, -\mathbf{0.1}, \mathbf{0.1}, \mathbf{0.2})^{\mathbf{T}}$.** Mean and standard deviation (SD) of estimates, mean standard error (SE) and coverage rate (Cov), power, mean squared error (MSE) when $\theta_2 = -0.1$.

	30% invalid						50% invalid					
Method	Mean	SD	SE	Cov	Power	MSE	Mean	SD	SE	Cov	Power	MSE
Scenario 1: Balanced pleiotropy, InSIDE met												
MVMR-cML-BIC	0.102	0.061	0.044	0.880	0.670	0.004	0.111	0.167	0.053	0.718	0.564	0.028
MVMR-cML-DP	0.101	0.060	0.068	0.964	0.402	0.004	0.104	0.132	0.146	0.950	0.152	0.018
MVMR-Egger	0.098	0.429	0.408	0.930	0.084	0.184	0.093	0.567	0.542	0.942	0.064	0.322
MVMR-IVW	0.092	0.376	0.358	0.930	0.092	0.141	0.105	0.496	0.471	0.936	0.074	0.246
MVMR-Lasso	0.105	0.089	0.049	0.872	0.588	0.008	0.085	0.270	0.073	0.674	0.524	0.073
MVMR-median	0.104	0.102	0.070	0.932	0.398	0.010	0.086	0.294	0.108	0.718	0.398	0.087
MVMR-robust	0.101	0.083	0.080	0.883	0.455	0.007	0.105	0.369	0.390	0.919	0.116	0.136
Scenario 2: Directional pleiotropy, InSIDE met												
MVMR-cML-BIC	0.103	0.057	0.045	0.874	0.614	0.003	0.091	0.184	0.055	0.732	0.538	0.034
MVMR-cML-DP	0.102	0.059	0.068	0.968	0.380	0.003	0.094	0.138	0.140	0.954	0.178	0.019
MVMR-Egger	0.109	0.463	0.465	0.942	0.066	0.215	0.138	0.642	0.580	0.914	0.086	0.414
MVMR-IVW	0.164	0.429	0.410	0.926	0.096	0.188	0.223	0.560	0.506	0.914	0.104	0.328
MVMR-Lasso	0.108	0.131	0.053	0.864	0.572	0.017	0.141	0.386	0.083	0.592	0.546	0.151
MVMR-median	0.114	0.154	0.077	0.892	0.366	0.024	0.150	0.390	0.125	0.662	0.420	0.155
MVMR-robust	0.107	0.076	0.073	0.888	0.464	0.006	0.163	0.401	0.465	0.935	0.096	0.164
Scenario 3: Directional pleiotropy, InSIDE violated												
MVMR-cML-BIC	0.097	0.065	0.044	0.874	0.610	0.004	0.108	0.168	0.054	0.746	0.510	0.028
MVMR-cML-DP	0.099	0.060	0.066	0.970	0.406	0.004	0.102	0.130	0.145	0.958	0.154	0.017
MVMR-Egger	0.159	0.459	0.437	0.936	0.082	0.214	0.141	0.592	0.553	0.918	0.090	0.353
MVMR-IVW	0.151	0.403	0.385	0.948	0.076	0.165	0.133	0.497	0.483	0.928	0.082	0.248
MVMR-Lasso	0.098	0.092	0.050	0.898	0.578	0.008	0.118	0.300	0.074	0.662	0.528	0.090
MVMR-median	0.103	0.107	0.071	0.922	0.378	0.011	0.115	0.328	0.115	0.740	0.378	0.108
MVMR-robust	0.096	0.078	0.075	0.889	0.431	0.006	0.121	0.360	0.389	0.919	0.104	0.130

Table B.5: **Simulation results for θ_3 when $\boldsymbol{\theta} = (\mathbf{0}, -\mathbf{0.1}, \mathbf{0.1}, \mathbf{0.2})^{\mathbf{T}}$.** Mean and standard deviation (SD) of estimates, mean standard error (SE) and coverage rate (Cov), power, mean squared error (MSE) when $\theta_3 = 0.1$.

	30% invalid						50% invalid					
Method	Mean	SD	SE	Cov	Power	MSE	Mean	SD	SE	Cov	Power	MSE
Scenario 1: Balanced pleiotropy, InSIDE met												
MVMR-cML-BIC	0.202	0.055	0.044	0.882	0.976	0.003	0.190	0.172	0.053	0.718	0.908	0.030
MVMR-cML-DP	0.202	0.055	0.067	0.972	0.850	0.003	0.192	0.124	0.143	0.962	0.448	0.015
MVMR-Egger	0.206	0.425	0.407	0.934	0.100	0.181	0.237	0.540	0.534	0.940	0.086	0.293
MVMR-IVW	0.206	0.386	0.357	0.938	0.114	0.149	0.253	0.468	0.468	0.942	0.104	0.222
MVMR-Lasso	0.200	0.071	0.048	0.886	0.946	0.005	0.217	0.254	0.073	0.654	0.788	0.065
MVMR-median	0.199	0.088	0.070	0.948	0.816	0.008	0.225	0.289	0.108	0.736	0.662	0.084
MVMR-robust	0.197	0.068	0.080	0.915	0.748	0.005	0.222	0.379	0.399	0.931	0.179	0.144
Scenario 2: Directional pleiotropy, InSIDE met												
MVMR-cML-BIC	0.202	0.058	0.045	0.904	0.962	0.003	0.201	0.168	0.055	0.736	0.866	0.028
MVMR-cML-DP	0.201	0.059	0.068	0.960	0.816	0.003	0.209	0.129	0.141	0.954	0.456	0.017
MVMR-Egger	0.216	0.490	0.468	0.946	0.098	0.241	0.192	0.613	0.578	0.920	0.084	0.376
MVMR-IVW	0.270	0.425	0.407	0.932	0.134	0.185	0.272	0.542	0.508	0.928	0.122	0.299
MVMR-Lasso	0.200	0.102	0.052	0.884	0.916	0.010	0.225	0.370	0.085	0.630	0.778	0.138
MVMR-median	0.200	0.122	0.075	0.910	0.788	0.015	0.236	0.389	0.130	0.700	0.620	0.153
MVMR-robust	0.196	0.076	0.075	0.886	0.764	0.006	0.248	0.427	0.464	0.925	0.148	0.185
Scenario 3: Directional pleiotropy, InSIDE violated												
MVMR-cML-BIC	0.200	0.058	0.044	0.904	0.964	0.003	0.194	0.142	0.054	0.768	0.878	0.020
MVMR-cML-DP	0.201	0.058	0.066	0.966	0.852	0.003	0.192	0.123	0.144	0.966	0.476	0.015
MVMR-Egger	0.268	0.454	0.434	0.926	0.134	0.211	0.234	0.588	0.548	0.916	0.112	0.347
MVMR-IVW	0.256	0.391	0.380	0.944	0.142	0.156	0.238	0.505	0.485	0.922	0.100	0.256
MVMR-Lasso	0.202	0.083	0.051	0.882	0.930	0.007	0.204	0.291	0.076	0.654	0.754	0.085
MVMR-median	0.210	0.114	0.073	0.918	0.824	0.013	0.210	0.313	0.115	0.720	0.634	0.098
MVMR-robust	0.201	0.072	0.076	0.883	0.777	0.005	0.201	0.390	0.406	0.923	0.182	0.152

Table B.6: **Simulation results for θ_4 when $\boldsymbol{\theta} = (\mathbf{0}, -\mathbf{0.1}, \mathbf{0.1}, \mathbf{0.2})^{\mathbf{T}}$.** Mean and standard deviation (SD) of estimates, mean standard error (SE) and coverage rate (Cov), power, mean squared error (MSE) when $\theta_4 = 0.2$.

	$\theta_1 = 0$						$\theta_2 = 0$					
Method	Mean	SD	SE	Cov	Type-I	MSE	Mean	SD	SE	Cov	Type-I	MSE
MVMR-cML-BIC	1.31E-03	0.045	0.040	0.916	0.084	2.02E-03	-8.98E-04	0.045	0.040	0.934	0.066	2.01E-03
MVMR-cML-DP	1.59E-03	0.046	0.060	0.986	0.014	2.16E-03	-1.17E-03	0.045	0.061	0.984	0.016	2.07E-03
MVMR-Egger	1.99E-03	0.054	0.054	0.940	0.060	2.93E-03	-2.68E-04	0.043	0.044	0.950	0.050	1.81E-03
MVMR-IVW	9.83E-04	0.037	0.038	0.952	0.048	1.40E-03	-5.22E-04	0.037	0.038	0.956	0.044	1.38E-03
MVMR-Lasso	1.97E-03	0.044	0.132	0.978	0.022	1.97E-03	-2.03E-03	0.045	0.140	0.978	0.022	2.03E-03
MVMR-median	1.54E-03	0.045	0.051	0.974	0.026	2.00E-03	-7.86E-04	0.044	0.052	0.980	0.020	1.93E-03
MVMR-robust	1.15E-03	0.039	0.040	0.936	0.064	1.49E-03	-5.04E-04	0.038	0.040	0.952	0.048	1.45E-03

* The average conditional F-statistics across 500 replicates for X_1 and X_2 are 6.73 and 6.70 (with SD 1.44 and 1.41) respectively.

Table B.7: **Simulation results for weak IVs without pleiotropy when $\boldsymbol{\theta} = (\mathbf{0}, \mathbf{0})^{\mathbf{T}}$.** Mean and standard deviation (SD) of estimates, mean standard error (SE) and coverage rate (Cov), type-I error, mean squared error (MSE).

	$\theta_1 = 0$						$\theta_2 = 0$					
Method	Mean	SD	SE	Cov	Type-I	MSE	Mean	SD	SE	Cov	Type-I	MSE
MVMR-cML-BIC	1.61E-04	0.039	0.036	0.942	0.058	1.50E-03	-2.24E-04	0.027	0.025	0.936	0.064	7.47E-04
MVMR-cML-DP	7.06E-04	0.040	0.051	0.992	0.008	1.57E-03	-6.42E-04	0.028	0.036	0.992	0.008	7.82E-04
MVMR-Egger	5.33E-04	0.035	0.036	0.962	0.038	1.23E-03	-4.91E-04	0.025	0.025	0.968	0.032	6.02E-04
MVMR-IVW	5.38E-04	0.034	0.035	0.974	0.026	1.17E-03	-4.71E-04	0.024	0.025	0.966	0.034	5.82E-04
MVMR-Lasso	-6.90E-05	0.042	0.122	0.978	0.022	1.73E-03	-3.76E-05	0.029	0.086	0.978	0.022	8.54E-04
MVMR-median	-1.12E-04	0.042	0.043	0.956	0.044	1.78E-03	-5.46E-06	0.030	0.030	0.960	0.040	8.87E-04
MVMR-robust	5.80E-04	0.035	0.037	0.946	0.054	1.25E-03	-5.60E-04	0.025	0.026	0.942	0.058	6.21E-04

* The average conditional F-statistics across 500 replicates for X_1 and X_2 are 9.35 and 9.38 (with SD 1.95 and 1.96) respectively.

Table B.8: **Simulation results for conditionally weak IVs without pleiotropy when $\boldsymbol{\theta} = (\mathbf{0}, \mathbf{0})^{\mathbf{T}}$.** Mean and standard deviation (SD) of estimates, mean standard error (SE) and coverage rate (Cov), type-I error, mean squared error (MSE).

	$\theta_1 = 0$						$\theta_2 = 0$					
Method	Mean	SD	SE	Cov	Type-I	MSE	Mean	SD	SE	Cov	Type-I	MSE
MVMR-cML-BIC	-3.66E-03	0.102	0.083	0.898	0.102	1.05E-02	2.38E-03	0.104	0.084	0.888	0.112	1.09E-02
MVMR-cML-DP	-2.52E-03	0.104	0.152	0.986	0.014	1.08E-02	6.68E-04	0.108	0.153	0.994	0.006	1.16E-02
MVMR-Egger	9.20E-02	1.657	1.554	0.940	0.060	2.75E+00	9.56E-02	1.442	1.287	0.928	0.072	2.09E+00
MVMR-IVW	-1.08E-02	1.167	1.099	0.930	0.070	1.36E+00	1.67E-02	1.194	1.109	0.938	0.062	1.43E+00
MVMR-Lasso	-1.27E-04	0.087	0.078	0.922	0.078	7.58E-03	-2.29E-04	0.089	0.080	0.928	0.072	7.98E-03
MVMR-median	-1.19E-03	0.124	0.113	0.926	0.074	1.55E-02	4.16E-04	0.128	0.114	0.942	0.058	1.63E-02
MVMR-robust	-9.16E-04	0.092	0.096	0.934	0.066	8.49E-03	4.78E-04	0.091	0.096	0.960	0.040	8.37E-03

* The average conditional F-statistics across 500 replicates for X_1 and X_2 are 6.73 and 6.70 (with SD 1.44 and 1.41) respectively.

Table B.9: **Simulation results for weak IVs with pleiotropy when $\boldsymbol{\theta} = (\mathbf{0}, \mathbf{0})^{\mathbf{T}}$.** Mean and standard deviation (SD) of estimates, mean standard error (SE) and coverage rate (Cov), type-I error, mean squared error (MSE).

	$\theta_1 = 0$						$\theta_2 = 0$					
Method	Mean	SD	SE	Cov	Type-I	MSE	Mean	SD	SE	Cov	Type-I	MSE
MVMR-cML-BIC	1.39E-03	0.076	0.056	0.870	0.130	5.84E-03	-5.94E-04	0.054	0.039	0.868	0.132	2.91E-03
MVMR-cML-DP	-3.53E-04	0.078	0.093	0.978	0.022	6.03E-03	5.74E-04	0.055	0.065	0.980	0.020	2.98E-03
MVMR-Egger	1.33E-02	0.634	0.599	0.936	0.064	4.03E-01	-1.10E-02	0.444	0.419	0.932	0.068	1.98E-01
MVMR-IVW	1.91E-02	0.623	0.588	0.932	0.068	3.88E-01	-1.40E-02	0.436	0.414	0.932	0.068	1.90E-01
MVMR-Lasso	1.68E-03	0.066	0.055	0.906	0.094	4.40E-03	-8.38E-04	0.047	0.039	0.910	0.090	2.21E-03
MVMR-median	5.17E-04	0.089	0.077	0.940	0.060	7.94E-03	-2.43E-04	0.063	0.054	0.934	0.066	3.95E-03
MVMR-robust	-1.20E-03	0.071	0.069	0.924	0.076	5.05E-03	1.31E-03	0.050	0.049	0.922	0.078	2.50E-03

* The average conditional F-statistics across 500 replicates for X_1 and X_2 are 9.35 and 9.38 (with SD 1.95 and 1.96) respectively.

Table B.10: **Simulation results for conditionally weak IVs with pleiotropy when $\boldsymbol{\theta} = (\mathbf{0}, \mathbf{0})^{\mathbf{T}}$.** Mean and standard deviation (SD) of estimates, mean standard error (SE) and coverage rate (Cov), type-I error, mean squared error (MSE).

	$\theta_1 = 0.5$						$\theta_2 = -0.3$					
Method	Mean	SD	SE	Cov	Power	MSE	Mean	SD	SE	Cov	Power	MSE
MVMR-cML-BIC	0.527	0.165	0.113	0.896	0.988	0.028	-0.332	0.170	0.114	0.898	0.798	0.030
MVMR-cML-DP	0.529	0.145	0.199	0.984	0.820	0.022	-0.334	0.150	0.203	0.988	0.432	0.024
MVMR-Egger	0.548	1.589	1.524	0.944	0.064	2.527	-0.141	1.386	1.262	0.934	0.076	1.946
MVMR-IVW	0.435	1.118	1.077	0.932	0.070	1.253	-0.224	1.144	1.087	0.948	0.060	1.314
MVMR-Lasso	0.436	0.118	0.102	0.866	0.970	0.018	-0.239	0.121	0.104	0.856	0.664	0.018
MVMR-median	0.436	0.156	0.147	0.898	0.840	0.029	-0.234	0.164	0.149	0.908	0.390	0.031
MVMR-robust	0.434	0.126	0.133	0.910	0.866	0.020	-0.235	0.128	0.135	0.914	0.512	0.021

* The average conditional F-statistics across 500 replicates for X_1 and X_2 are 6.73 and 6.70 (with SD 1.44 and 1.41) respectively.

Table B.11: **Simulation results for weak IVs with pleiotropy when $\boldsymbol{\theta} = (\mathbf{0.5}, \mathbf{-0.3})^{\mathbf{T}}$.** Mean and standard deviation (SD) of estimates, mean standard error (SE) and coverage rate (Cov), power, mean squared error (MSE).

	$\theta_1 = 0.5$						$\theta_2 = -0.3$					
Method	Mean	SD	SE	Cov	Power	MSE	Mean	SD	SE	Cov	Power	MSE
MVMR-cML-BIC	0.519	0.110	0.080	0.862	0.998	0.012	-0.313	0.077	0.056	0.866	0.996	0.006
MVMR-cML-DP	0.524	0.108	0.137	0.992	0.956	0.012	-0.316	0.076	0.096	0.990	0.920	0.006
MVMR-Egger	0.465	0.618	0.592	0.936	0.150	0.383	-0.277	0.433	0.415	0.930	0.116	0.188
MVMR-IVW	0.471	0.606	0.582	0.926	0.154	0.368	-0.280	0.425	0.409	0.930	0.124	0.181
MVMR-Lasso	0.450	0.092	0.074	0.838	0.998	0.011	-0.265	0.065	0.052	0.838	0.976	0.005
MVMR-median	0.453	0.117	0.112	0.928	0.964	0.016	-0.266	0.082	0.079	0.918	0.916	0.008
MVMR-robust	0.446	0.104	0.104	0.892	0.942	0.014	-0.262	0.073	0.074	0.892	0.876	0.007

* The average conditional F-statistics across 500 replicates for X_1 and X_2 are 9.35 and 9.38 (with SD 1.95 and 1.96) respectively.

Table B.12: **Simulation results for conditionally weak IVs with pleiotropy when $\boldsymbol{\theta} = (\mathbf{0.5}, \mathbf{-0.3})^{\mathbf{T}}$.** Mean and standard deviation (SD) of estimates, mean standard error (SE) and coverage rate (Cov), power, mean squared error (MSE).

	$\theta_1 = 0.1$ ($\theta_{1T} = 0.2$)				$\theta_2 = 0.2$			
Method	Mean	SD	SE	Power	Mean	SD	SE	Power
MVMR-cML-BIC	0.101	0.019	0.018	0.994	0.200	0.014	0.013	1.000
MVMR-cML-DP	0.101	0.020	0.021	0.992	0.199	0.014	0.015	1.000
MVMR-Egger	0.100	0.021	0.020	0.988	0.199	0.017	0.017	1.000
MVMR-IVW	0.100	0.019	0.018	0.994	0.199	0.014	0.013	1.000
MVMR-Lasso	0.100	0.023	0.045	0.729	0.199	0.017	0.031	0.980
MVMR-median	0.101	0.023	0.025	0.980	0.199	0.017	0.018	1.000
MVMR-robust	0.100	0.020	0.019	0.994	0.199	0.014	0.014	1.000
UVMR-cML-BIC	0.350	0.089	0.011	1.000	NA	NA	NA	NA
UVMR-cML-DP	0.298	0.045	0.019	1.000	NA	NA	NA	NA
UVMR-IVW	0.344	0.029	0.031	1.000	NA	NA	NA	NA

Table B.13: **Simulation results for mediation analysis when $\mathbf{K = 0}$, $\mathbf{m_1 = 1}$.** Mean and standard deviation (SD) of estimates, mean standard error (SE) and power. The total causal effect of X_1 is $\theta_{1T} = 0.2$.

	$\theta_1 = 0.1$ ($\theta_{1T} = 0.2$)				$\theta_2 = 0.2$			
Method	Mean	SD	SE	Power	Mean	SD	SE	Power
MVMR-cML-BIC	0.100	0.016	0.015	1.000	0.199	0.019	0.017	1.000
MVMR-cML-DP	0.100	0.016	0.018	1.000	0.199	0.020	0.021	0.998
MVMR-Egger	0.100	0.021	0.021	0.992	0.199	0.020	0.019	1.000
MVMR-IVW	0.100	0.016	0.016	1.000	0.199	0.019	0.018	1.000
MVMR-Lasso	0.100	0.020	0.037	0.851	0.199	0.025	0.039	0.974
MVMR-median	0.100	0.019	0.021	0.998	0.199	0.023	0.024	1.000
MVMR-robust	0.102	0.026	0.017	0.984	0.196	0.047	0.019	0.992
UVMR-cML-BIC	0.208	0.013	0.009	1.000	NA	NA	NA	NA
UVMR-cML-DP	0.206	0.012	0.010	1.000	NA	NA	NA	NA
UVMR-IVW	0.245	0.018	0.024	1.000	NA	NA	NA	NA

Table B.14: **Simulation results for mediation analysis when $\mathbf{K = 0}$, $\mathbf{m_1 = 14}$.** Mean and standard deviation (SD) of estimates, mean standard error (SE) and power. The total causal effect of X_1 is $\theta_{1T} = 0.2$.

	$\theta_1 = 0.1$ ($\theta_{1T} = 0.2$)				$\theta_2 = 0.2$			
Method	Mean	SD	SE	Power	Mean	SD	SE	Power
MVMR-cML-BIC	0.102	0.043	0.023	0.954	0.196	0.080	0.038	0.946
MVMR-cML-DP	0.105	0.036	0.045	0.762	0.189	0.066	0.081	0.766
MVMR-Egger	0.103	0.032	0.027	0.948	0.194	0.052	0.039	0.938
MVMR-IVW	0.103	0.029	0.023	0.956	0.194	0.052	0.037	0.940
MVMR-Lasso	0.112	0.050	0.085	0.704	0.176	0.095	0.142	0.748
MVMR-median	0.102	0.031	0.028	0.922	0.196	0.055	0.044	0.928
MVMR-robust	0.132	0.107	0.036	0.930	0.136	0.213	0.060	0.856
UVMR-cML-BIC	0.202	0.009	0.008	1.000	NA	NA	NA	NA
UVMR-cML-DP	0.201	0.009	0.009	1.000	NA	NA	NA	NA
UVMR-IVW	0.215	0.012	0.016	1.000	NA	NA	NA	NA

Table B.15: **Simulation results for mediation analysis when $\mathbf{K = 0}$, $\mathbf{m_1 = 18}$.** Mean and standard deviation (SD) of estimates, mean standard error (SE) and power. The total causal effect of X_1 is $\theta_{1T} = 0.2$.

	$\theta_1 = 0.1$ ($\theta_{1T} = 0.2$)				$\theta_2 = 0.2$			
Method	Mean	SD	SE	Power	Mean	SD	SE	Power
MVMR-cML-BIC	0.100	0.017	0.016	0.998	0.200	0.018	0.018	1.000
MVMR-cML-DP	0.100	0.017	0.020	0.990	0.200	0.018	0.021	0.994
MVMR-Egger	0.175	0.337	0.351	0.086	0.051	0.192	0.322	0.048
MVMR-IVW	0.287	0.246	0.269	0.166	0.093	0.151	0.309	0.044
MVMR-Lasso	0.100	0.018	0.017	0.996	0.200	0.019	0.018	1.000
MVMR-median	0.105	0.022	0.023	0.984	0.197	0.022	0.025	0.998
MVMR-robust	0.100	0.017	0.016	0.996	0.200	0.018	0.016	1.000
UVMR-cML-BIC	0.209	0.015	0.010	1.000	NA	NA	NA	NA
UVMR-cML-DP	0.207	0.014	0.012	1.000	NA	NA	NA	NA
UVMR-IVW	0.354	0.144	0.145	0.708	NA	NA	NA	NA

Table B.16: **Simulation results for mediation analysis when $\mathbf{K = 3}$, $\mathbf{m_1 = 11}$.** Mean and standard deviation (SD) of estimates, mean standard error (SE) and power. The total causal effect of X_1 is $\theta_{1T} = 0.2$.

	$\theta_1 = 0.1$ ($\theta_{1T} = 0.2$)				$\theta_2 = 0.2$			
Method	Mean	SD	SE	Power	Mean	SD	SE	Power
MVMR-cML-BIC	0.099	0.031	0.024	0.950	0.202	0.054	0.038	0.962
MVMR-cML-DP	0.102	0.028	0.048	0.736	0.195	0.049	0.087	0.744
MVMR-Egger	0.182	0.379	0.465	0.042	0.039	0.418	0.647	0.012
MVMR-IVW	0.280	0.279	0.390	0.040	0.076	0.351	0.633	0.014
MVMR-Lasso	0.103	0.040	0.024	0.962	0.193	0.075	0.039	0.946
MVMR-median	0.106	0.029	0.030	0.904	0.195	0.047	0.045	0.926
MVMR-robust	0.126	0.107	0.026	0.960	0.148	0.210	0.039	0.932
UVMR-cML-BIC	0.202	0.010	0.009	1.000	NA	NA	NA	NA
UVMR-cML-DP	0.201	0.010	0.010	1.000	NA	NA	NA	NA
UVMR-IVW	0.324	0.143	0.144	0.654	NA	NA	NA	NA

Table B.17: **Simulation results for mediation analysis when $\mathbf{K = 3}$, $\mathbf{m_1 = 15}$.** Mean and standard deviation (SD) of estimates, mean standard error (SE) and power. The total causal effect of X_1 is $\theta_{1T} = 0.2$.

	$\theta_1 = 0.1$ ($\theta_{1T} = 0.2$)				$\theta_2 = 0.2$			
Method	Mean	SD	SE	Power	Mean	SD	SE	Power
MVMR-cML-BIC	0.102	0.030	0.022	0.972	0.199	0.026	0.020	0.998
MVMR-cML-DP	0.102	0.029	0.032	0.866	0.199	0.025	0.027	0.992
MVMR-Egger	0.389	0.580	0.638	0.062	-0.261	0.344	0.589	0.004
MVMR-IVW	0.699	0.438	0.495	0.270	-0.142	0.288	0.571	0.002
MVMR-Lasso	0.201	0.290	0.035	0.920	0.144	0.181	0.029	0.946
MVMR-median	0.227	0.290	0.055	0.900	0.130	0.180	0.042	0.892
MVMR-robust	0.269	0.339	0.491	0.123	0.105	0.204	0.287	0.319
UVMR-cML-BIC	0.320	0.748	0.019	0.998	NA	NA	NA	NA
UVMR-cML-DP	0.209	0.064	0.025	0.928	NA	NA	NA	NA
UVMR-IVW	0.591	0.249	0.266	0.628	NA	NA	NA	NA

Table B.18: **Simulation results for mediation analysis when $\mathbf{K = 10}$, $\mathbf{m_1 = 4}$.** Mean and standard deviation (SD) of estimates, mean standard error (SE) and power. The total causal effect of X_1 is $\theta_{1T} = 0.2$.

Trait	#IVs	F-statistics on full set of 201 IVs	F-statistics on selected set of valid IVs by MVMR-cML-BIC
TG	15	25.26	26.52
LDL	19	24.20	20.91
HDL	26	35.83	37.05
BMI	14	8.15	8.33
Height	73	43.56	43.86
FG	14	11.92	12.08
SBP	66	17.19	15.49
DBP	79	18.06	16.02

Table B.19: **Conditional F-statistics for the 8 exposures in the MVMR analysis.**

B.3 Theory

B.3.1 Proof of model identification condition in MVMR

Proof of Theorem 3.1. ($\Rightarrow$) Suppose we have a set of solution $(\hat{\boldsymbol{\theta}}, \hat{\mathbf{r}} = (\hat{r}_1, \ldots, \hat{r}_m))$ with $\widehat{\mathcal{V}} = \{i : \hat{r}_i = 0\}$, we show that this set of solution must be equal to the ground truth when Eq (3.3) in the main text holds. As $\mathbf{B}_{\mathcal{V}^*}$ has full column rank, the identifiability of $(\boldsymbol{\theta}^*, \mathbf{r}^*)$ is equivalent to the identifiability of $\boldsymbol{\theta}^*$.

If $\hat{\boldsymbol{\theta}} \neq \boldsymbol{\theta}^*$, for any $i \in \mathcal{V}^*$, we have $b_{Yi} = \mathbf{b}_{Xi}^T \boldsymbol{\theta}^* = \mathbf{b}_{Xi}^T \hat{\boldsymbol{\theta}} + \hat{r}_i$, which simplifies to $\hat{r}_i = \mathbf{b}_{Xi}^T \mathbf{c}$, where $\mathbf{c} = \boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}} \neq \mathbf{0}$. This implies $\mathcal{V}^* \subseteq \{i : \hat{r}_i = \mathbf{b}_{Xi}^T \mathbf{c}\}$. As any solution to Eq (3.2) in the main text is obtained from the largest set(s) of IVs with $r_i = 0$, we have $|\mathcal{V}^*| \leq |\{i : \hat{r}_i = \mathbf{b}_{Xi}^T \mathbf{c}\}| \leq |\widehat{\mathcal{V}}|$. Similarly, for any $i \in \widehat{\mathcal{V}}$, we have $b_{Yi} = \mathbf{b}_{Xi}^T \hat{\boldsymbol{\theta}} = \mathbf{b}_{Xi}^T \boldsymbol{\theta}^* + r_i^*$, which simplifies to $r_i^* = \mathbf{b}_{Xi}^T \mathbf{c}$, where $\mathbf{c} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \neq \mathbf{0}$. This implies $\widehat{\mathcal{V}} \subseteq \{i : r_i^* = \mathbf{b}_{Xi}^T \mathbf{c}\}$. Combining the plurality condition Eq (3.3) in the main text, we have $|\widehat{\mathcal{V}}| \leq |\{i : r_i^* = \mathbf{b}_{Xi}^T \mathbf{c}\}| < |\mathcal{V}^*|$. Then we have $|\mathcal{V}^*| \leq |\widehat{\mathcal{V}}|$ and $|\widehat{\mathcal{V}}| < |\mathcal{V}^*|$, which leads to contradiction.

($\Leftarrow$) If $\mathbf{B}_{\mathcal{V}^*}$ does not have full column rank, we could find another $\tilde{\theta} \neq \theta^*$ such that $\mathbf{b}_{Y\mathcal{V}^*} = \mathbf{B}_{\mathcal{V}^*} \theta^* = \mathbf{B}_{\mathcal{V}^*} \tilde{\theta}$, thus the true parameters are not identifiable. When Eq (3.3) does not hold, i.e., $|\mathcal{V}^*| \leq \max_{\mathbf{c} \neq \mathbf{0}} \left|\{i : r_i^* = \mathbf{b}_{Xi}^T \mathbf{c}\}\right|$, let $\mathbf{c}_m \neq \mathbf{0}$ denote one of the maximizing $\mathbf{c}$ for $\left|\{i : r_i^* = \mathbf{b}_{Xi}^T \mathbf{c}\}\right|$. Then we can always find a solution to Eq (3.2) where $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + \mathbf{c}_m$, $\hat{\mathbf{r}} = \mathbf{r}^* - \mathbf{b}_{Xi}^T \mathbf{c}_m$ and $\widehat{\mathcal{V}} = \{i : \hat{r}_i = 0\}$. We can easily verify that $\mathbf{b}_{Xi}^T \hat{\boldsymbol{\theta}} + \hat{r}_i = \mathbf{b}_{Xi}^T \boldsymbol{\theta}^* + \mathbf{b}_{Xi}^T \mathbf{c}_m + r_i^* - \mathbf{b}_{Xi}^T \mathbf{c}_m = b_{Yi}, i = 1, \ldots, m$. Furthermore, we also have the constraint that

$$
\begin{aligned}
|\widehat{\mathcal{V}}| = |\{i : \hat{r}_i = 0\}| = |\{i : r_i^* = \mathbf{b}_{Xi}^T \mathbf{c}_m\}| &\geq \max_{\mathbf{c} \neq \mathbf{0}} |\{i : r_i^* = \mathbf{b}_{Xi}^T (\mathbf{c}_m + \mathbf{c})\}| \\
&= \max_{\mathbf{c} \neq \mathbf{0}} |\{i : \hat{r}_i = \mathbf{b}_{Xi}^T \mathbf{c}\}|,
\end{aligned}
$$

Therefore, we find a solution to Eq (3.2) not equal to the ground truth, and the model

is not identifiable. □

B.3.2 Proof of Theorem 3.2

To prove Theorem 3.2, we first prove the selection consistency of BIC (Lemma B.1), i.e, it will select the correct set of invalid IVs as the GWAS sample size N goes to infinity. Then we prove the second part of Theorem 3.2, i.e., the estimation consistency and asymptotic normality based on the set of (selected) valid IVs.

Denote $K_0 = m - m_0$ the number of true invalid IVs.

Lemma B.1

With Assumption 3.2 to Assumption 3.3 satisfied, if $K_0 \in \mathcal{K}$, we have $P(\hat{K} = K_0) \to 1$ and $P(\hat{B}_{\hat{K}} = B_0) \to 1$ as $N \to \infty$.

Proof. First, we show $P(\hat{B}_{K_0} = B_0) \to 1$, which is equivalent to show for any $B_1 \subseteq \{1, \cdots, m\}$ such that $|B_1| = K_0$ and $B_1 \neq B_0$, $P(\hat{B}_{K_0} = B_1) \to 0$ as $N \to \infty$. We have

$$\begin{aligned} &P(\hat{B}_{K_0} = B_1) \\ &\leq P\Big\{ \min_{\widetilde{\boldsymbol{\theta}}, \widetilde{\mathbf{b}}_{Xi}} \sum_{i \in B_1^c} (\hat{\boldsymbol{\beta}}_i - \widetilde{\mathbf{b}}_i)^T \boldsymbol{\Sigma}_i^{-1} (\hat{\boldsymbol{\beta}}_i - \widetilde{\mathbf{b}}_i) \leq \sum_{i \in B_0^c} (\hat{\boldsymbol{\beta}}_i - \mathbf{b}_i)^T \boldsymbol{\Sigma}_i^{-1} (\hat{\boldsymbol{\beta}}_i - \mathbf{b}_i) \Big\}. \end{aligned}$$

Note that, for $i \in B_0^c$, $\hat{\boldsymbol{\beta}}_i - \mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_i)$. So for any $\epsilon > 0$, there exists $C > 0$ such that

$$P\Big\{ \sum_{i \in B_0^c} (\hat{\boldsymbol{\beta}}_i - \mathbf{b}_i)^T \boldsymbol{\Sigma}_i^{-1} (\hat{\boldsymbol{\beta}}_i - \mathbf{b}_i) > C \Big\} < \frac{\epsilon}{2}. \tag{B.1}$$

Then we have

$$P\Big\{ \min_{\widetilde{\boldsymbol{\theta}},\widetilde{\mathbf{b}}_{Xi}} \sum_{i\in B_1^c} (\hat{\boldsymbol{\beta}}_i - \widetilde{\mathbf{b}}_i)^T \boldsymbol{\Sigma}_i^{-1} (\hat{\boldsymbol{\beta}}_i - \widetilde{\mathbf{b}}_i) \le \sum_{i\in B_0^c} (\hat{\boldsymbol{\beta}}_i - \mathbf{b}_i)^T \boldsymbol{\Sigma}_i^{-1} (\hat{\boldsymbol{\beta}}_i - \mathbf{b}_i) \Big\}$$
$$\le P\Big\{ \min_{\widetilde{\boldsymbol{\theta}},\widetilde{\mathbf{b}}_{Xi}} \sum_{i\in B_1^c} (\hat{\boldsymbol{\beta}}_i - \widetilde{\mathbf{b}}_i)^T \boldsymbol{\Sigma}_i^{-1} (\hat{\boldsymbol{\beta}}_i - \widetilde{\mathbf{b}}_i) \le C \Big\} + P\Big\{ \sum_{i\in B_0^c} (\hat{\boldsymbol{\beta}}_i - \mathbf{b}_i)^T \boldsymbol{\Sigma}_i^{-1} (\hat{\boldsymbol{\beta}}_i - \mathbf{b}_i) > C \Big\}.$$

With Eq (B.1), we only need to prove, as $N \to \infty$,

$$P\Big\{ \min_{\widetilde{\boldsymbol{\theta}},\widetilde{\mathbf{b}}_{Xi}} \sum_{i\in B_1^c} (\hat{\boldsymbol{\beta}}_i - \widetilde{\mathbf{b}}_i)^T \boldsymbol{\Sigma}_i^{-1} (\hat{\boldsymbol{\beta}}_i - \widetilde{\mathbf{b}}_i) \le C \Big\} < \epsilon/2.$$

After profiling out $\widetilde{\mathbf{b}}_{Xi}$'s, we get

$$\min_{\widetilde{\boldsymbol{\theta}},\widetilde{\mathbf{b}}_{Xi}} \sum_{i\in B_1^c} (\hat{\boldsymbol{\beta}}_i - \widetilde{\mathbf{b}}_i)^T \boldsymbol{\Sigma}_i^{-1} (\hat{\boldsymbol{\beta}}_i - \widetilde{\mathbf{b}}_i) = \min_{\widetilde{\boldsymbol{\theta}}} \sum_{i\in B_1^c} \frac{(\hat{\beta}_{Yi} - \widetilde{\boldsymbol{\theta}}^T \hat{\boldsymbol{\beta}}_{Xi})^2}{\widetilde{\boldsymbol{\gamma}}^T \boldsymbol{\Sigma}_i \widetilde{\boldsymbol{\gamma}}},$$

where $\widetilde{\boldsymbol{\gamma}}^T = [\widetilde{\boldsymbol{\theta}}^T, -1]$.

Hence

$$P\Big\{ \min_{\widetilde{\boldsymbol{\theta}},\widetilde{\mathbf{b}}_{Xi}} \sum_{i\in B_1^c} (\hat{\boldsymbol{\beta}}_i - \widetilde{\mathbf{b}}_i)^T \boldsymbol{\Sigma}_i^{-1} (\hat{\boldsymbol{\beta}}_i - \widetilde{\mathbf{b}}_i) \le C \Big\} = P\Big\{ \min_{\widetilde{\boldsymbol{\theta}}} \sum_{i\in B_1^c} \frac{(\hat{\beta}_{Yi} - \widetilde{\boldsymbol{\theta}}^T \hat{\boldsymbol{\beta}}_{Xi})^2}{\widetilde{\boldsymbol{\gamma}}^T \boldsymbol{\Sigma}_i \widetilde{\boldsymbol{\gamma}}} \le C \Big\}.$$

We have $\frac{(\hat{\beta}_{Yi} - \widetilde{\boldsymbol{\theta}}^T \hat{\boldsymbol{\beta}}_{Xi})}{\sqrt{\widetilde{\boldsymbol{\gamma}}^T \boldsymbol{\Sigma}_i \widetilde{\boldsymbol{\gamma}}}} \sim \mathcal{N}(\frac{\boldsymbol{\theta}^{*T} \mathbf{b}_{Xi} + r_i^* - \widetilde{\boldsymbol{\theta}}^T \mathbf{b}_{Xi}}{\sqrt{\widetilde{\boldsymbol{\gamma}}^T \boldsymbol{\Sigma}_i \widetilde{\boldsymbol{\gamma}}}}, 1)$, so $\sum_{i\in B_1^c} \frac{(\hat{\beta}_{Yi} - \widetilde{\boldsymbol{\theta}}^T \hat{\boldsymbol{\beta}}_{Xi})^2}{\widetilde{\boldsymbol{\gamma}}^T \boldsymbol{\Sigma}_i \widetilde{\boldsymbol{\gamma}}}$ follows non-central χ^2 distribution with non-centrality parameter $\lambda_{\widetilde{\theta}}$ depending on $\widetilde{\theta}$

$$\lambda_{\widetilde{\theta}} = \sum_{i\in B_1^c} \frac{(\boldsymbol{\theta}^{*T} \mathbf{b}_{Xi} + r_i^* - \widetilde{\boldsymbol{\theta}}^T \mathbf{b}_{Xi})^2}{\widetilde{\boldsymbol{\gamma}}^T \boldsymbol{\Sigma}_i \widetilde{\boldsymbol{\gamma}}}. \tag{B.2}$$

Notice that there exists $i \in B_1^c$ such that the nominator of the $i-$th summand in

Eq (B.2) is strictly greater than 0. This is because by Assumption 3.2, there is no $\widetilde{\boldsymbol{\theta}}$ making $\boldsymbol{\theta}^{*T}\mathbf{b}_{Xi} + r_i^* - \widetilde{\boldsymbol{\theta}}^T\mathbf{b}_{Xi} = 0$ for all $i \in B_1^c$. Furthermore, by Assumption 3.3, the denominator of the $i-$th summand in Eq (B.2) goes to 0 as $N \to \infty$. Hence we have $\min_{\widetilde{\theta}} \lambda_{\widetilde{\theta}} \to \infty$ as $N \to \infty$. Then as N large enough, we have

$$P\Big\{ \min_{\widetilde{\boldsymbol{\theta}},\widetilde{\mathbf{b}}_{Xi}} \sum_{i\in B_1^c} (\hat{\boldsymbol{\beta}}_i - \widetilde{\mathbf{b}}_i)^T \boldsymbol{\Sigma}_i^{-1} (\hat{\boldsymbol{\beta}}_i - \widetilde{\mathbf{b}}_i) \le C \Big\} < \epsilon/2. \tag{B.3}$$

Combining (Eq (B.1)) and (Eq (B.3)), we get $P(\hat{B}_{K_0} = B_0) \to 1$ as $N \to \infty$.

Next, we show $P(\hat{K} = K_0) \to 1$. For any $K_1 < K_0$, we have

$$\begin{aligned} P(\hat{K} = K_1) &\le P\big\{ \mathrm{BIC}(K_1) \le \mathrm{BIC}(K_0) \big\} \\ &= P\Big\{ 2l\left(\widehat{\boldsymbol{\theta}}(K_0), \widehat{\mathbf{b}}_{Xi}(K_0), \hat{r}_i(K_0)\right) - 2l\left(\widehat{\boldsymbol{\theta}}(K_1), \widehat{\mathbf{b}}_{Xi}(K_1), \hat{r}_i(K_1)\right) \\ &\le \log(N)(K_0 - K_1) \Big\}. \end{aligned}$$

As we have shown $P(\hat{B}_{K_0} = B_0) \to 1$, with probability goes to 1 we have

$$\begin{aligned} & 2l\left(\widehat{\boldsymbol{\theta}}(K_0), \widehat{\mathbf{b}}_{Xi}(K_0), \hat{r}_i(K_0)\right) - 2l\left(\widehat{\boldsymbol{\theta}}(K_1), \widehat{\mathbf{b}}_{Xi}(K_1), \hat{r}_i(K_1)\right) \\ \ge & \min_{\widetilde{\boldsymbol{\theta}},\widetilde{\mathbf{b}}_{Xi}} \sum_{i\in \hat{B}_{K_1}^c} (\hat{\boldsymbol{\beta}}_i - \widetilde{\mathbf{b}}_i)^T \boldsymbol{\Sigma}_i^{-1} (\hat{\boldsymbol{\beta}}_i - \widetilde{\mathbf{b}}_i) - \sum_{i\in B_0^c} (\hat{\boldsymbol{\beta}}_i - \mathbf{b}_i)^T \boldsymbol{\Sigma}_i^{-1} (\hat{\boldsymbol{\beta}}_i - \mathbf{b}_i). \end{aligned}$$

Then we get

$$\begin{aligned} P(\hat{K} = K_1) \le \sum_{|B|=K_1} P\Big\{ & \min_{\widetilde{\theta},\widetilde{b}_{Xi}} \sum_{i\in B^c} (\hat{\boldsymbol{\beta}}_i - \widetilde{\mathbf{b}}_i)^T \boldsymbol{\Sigma}_i^{-1} (\hat{\boldsymbol{\beta}}_i - \widetilde{\mathbf{b}}_i) \\ & \le \sum_{i\in B_0^c} (\hat{\boldsymbol{\beta}}_i - \mathbf{b}_i)^T \boldsymbol{\Sigma}_i^{-1} (\hat{\boldsymbol{\beta}}_i - \mathbf{b}_i) + \log(N)(K_0 - K_1) \Big\}. \end{aligned}$$

Similar as above, we get

$$\min_{\widetilde{\boldsymbol{\theta}},\widetilde{\mathbf{b}}_{Xi}} \sum_{i\in B^c} (\hat{\boldsymbol{\beta}}_i - \widetilde{\mathbf{b}}_i)^T \boldsymbol{\Sigma}_i^{-1} (\hat{\boldsymbol{\beta}}_i - \widetilde{\mathbf{b}}_i) = \min_{\widetilde{\boldsymbol{\theta}}} \sum_{i\in B^c} \frac{(\hat{\beta}_{Yi} - \widetilde{\boldsymbol{\theta}}^T \hat{\boldsymbol{\beta}}_{Xi})^2}{\widetilde{\boldsymbol{\gamma}}^T \boldsymbol{\Sigma}_i \widetilde{\boldsymbol{\gamma}}},$$

which follows non-central χ^2 distribution with non-centrality parameter $\lambda_{\widetilde{\theta}}$. Similarly, since $K_1 < K_0$ (to apply Assumption 3.2), and with Assumption 3.3 we have $\lambda_{\widetilde{\theta}} = O(N)$. So for any $|B| = K_1$, we get

$$\begin{aligned} P\Big\{ &\min_{\widetilde{\theta},\widetilde{b}_{Xi}} \sum_{i\in B^c} (\hat{\boldsymbol{\beta}}_i - \widetilde{\mathbf{b}}_i)^T \boldsymbol{\Sigma}_i^{-1} (\hat{\boldsymbol{\beta}}_i - \widetilde{\mathbf{b}}_i) \\ &\leq \sum_{i\in B_0^c} (\hat{\boldsymbol{\beta}}_i - \mathbf{b}_i)^T \boldsymbol{\Sigma}_i^{-1} (\hat{\boldsymbol{\beta}}_i - \mathbf{b}_i) + \log(N)(K_0 - K_1) \Big\} \to 0 \end{aligned}$$

This gives us $P(\hat{K} = K_1) \to 0$ for any $K_1 < K_0$. For any $K_1 > K_0$, we have

$$P(\hat{K} = K_1) \leq P\Big\{ \log(N)(K_1 - K_0) \leq \sum_{i\in B_0^c} (\hat{\boldsymbol{\beta}}_i - \mathbf{b}_i)^T \boldsymbol{\Sigma}_i^{-1} (\hat{\boldsymbol{\beta}}_i - \mathbf{b}_i) \Big\}$$

Since each $(\hat{\boldsymbol{\beta}}_i - \mathbf{b}_i)^T \boldsymbol{\Sigma}_i^{-1} (\hat{\boldsymbol{\beta}}_i - \mathbf{b}_i)$ follows a central χ^2 distribution with degree of freedom $(L+1)$, and both the number of IVs and the number of exposures are treated as fixed, we get $P(\hat{K} = K_1) \to 0$ for any $K_1 > K_0$. So we have $P(\hat{K} = K_0) \to 1$ as $N \to \infty$. □

Now assume all m IVs are valid, the profile likelihood is:

$$\tilde{l}(\boldsymbol{\theta}) = \max_{\{\mathbf{b}_{Xi}\}} l(\theta, \{\mathbf{b}_{Xi}\}) = -\frac{1}{2} \sum_{i=1}^{m} \frac{(\hat{\beta}_{Yi} - \boldsymbol{\theta}^T \hat{\boldsymbol{\beta}}_{Xi})^2}{\boldsymbol{\gamma}^T \boldsymbol{\Sigma}_i \boldsymbol{\gamma}}, \tag{B.4}$$

where $\boldsymbol{\gamma}^T = [\boldsymbol{\theta}^T, -1]$.

Lemma B.2

Under the model that all m IVs are valid and Assumption 3.1 to Assumption 3.3, the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ is consistent, that is, $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}^*$ as $N \to \infty$.

Proof. Let $e_i = \hat{\beta}_{Yi} - b_{Yi}$, $\boldsymbol{\epsilon}_i = \widehat{\boldsymbol{\beta}}_{Xi} - \mathbf{b}_{Xi}$. After some algebra, we have

$$\tilde{l}(\boldsymbol{\theta}) = -\frac{1}{2}\sum_{i=1}^{m} \frac{(\mathbf{b}_{Xi}^T(\boldsymbol{\theta}^* - \boldsymbol{\theta}))^2 + (e_i - \boldsymbol{\theta}^T\boldsymbol{\epsilon}_i)^2 + 2\mathbf{b}_{Xi}^T(\boldsymbol{\theta}^* - \boldsymbol{\theta})(e_i - \boldsymbol{\theta}^T\boldsymbol{\epsilon}_i)}{\boldsymbol{\gamma}^T\boldsymbol{\Sigma}_i\boldsymbol{\gamma}},$$
$$\tilde{l}(\boldsymbol{\theta}^*) = -\frac{1}{2}\sum_{i=1}^{m} \frac{(e_i - \boldsymbol{\theta}^{*T}\boldsymbol{\epsilon}_i)^2}{{\boldsymbol{\gamma}_0}^T\boldsymbol{\Sigma}_i\boldsymbol{\gamma}_0},$$

where ${\boldsymbol{\gamma}_0}^T = [\boldsymbol{\theta}^{*T}, -1]$. Notice that $e_i - \boldsymbol{\theta}^T\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \boldsymbol{\gamma}^T\boldsymbol{\Sigma}_i\boldsymbol{\gamma})$ and $e_i - \boldsymbol{\theta}^{*T}\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, {\boldsymbol{\gamma}_0}^T\boldsymbol{\Sigma}_i\boldsymbol{\gamma}_0)$, and we have $\sum_{i=1}^m (e_i - \boldsymbol{\theta}^T\boldsymbol{\epsilon}_i)^2/\boldsymbol{\gamma}^T\boldsymbol{\Sigma}_i\boldsymbol{\gamma}$ and $\sum_{i=1}^m (e_i - \boldsymbol{\theta}^{*T}\boldsymbol{\epsilon}_i)^2/{\boldsymbol{\gamma}_0}^T\boldsymbol{\Sigma}_i\boldsymbol{\gamma}_0$ follow a χ^2 distribution with degree of freedom m. And they are both $O_p(1)$ as we consider a fixed number of IVs (i.e., fixed m). For any $\boldsymbol{\theta}$ such that $|\boldsymbol{\theta} - \boldsymbol{\theta}^*| > \epsilon$, there exists $i \in [m]$ such that $\mathbf{b}_{Xi}^T(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \neq 0$ and $(\mathbf{b}_{Xi}^T(\boldsymbol{\theta} - \boldsymbol{\theta}^*))^2/\boldsymbol{\gamma}^T\boldsymbol{\Sigma}_i\boldsymbol{\gamma} = O(N)$, $\mathbf{b}_{Xi}^T(\boldsymbol{\theta}^* - \boldsymbol{\theta})(e_i - \boldsymbol{\theta}^T\boldsymbol{\epsilon}_i)/\boldsymbol{\gamma}^T\boldsymbol{\Sigma}_i\boldsymbol{\gamma} = O_p(\sqrt{N})$. Therefore, when $N \to \infty$, we have

$$P\Big\{\tilde{l}(\boldsymbol{\theta}^*) - \tilde{l}(\boldsymbol{\theta}) > 0\Big\} = P\Big\{\frac{1}{2}\sum_{i=1}^{m} \frac{(\mathbf{b}_{Xi}^T(\boldsymbol{\theta}^* - \boldsymbol{\theta}))^2 + (e_i - \boldsymbol{\theta}^T\boldsymbol{\epsilon}_i)^2 + 2\mathbf{b}_{Xi}^T(\boldsymbol{\theta}^* - \boldsymbol{\theta})(e_i - \boldsymbol{\theta}^T\boldsymbol{\epsilon}_i)}{\boldsymbol{\gamma}^T\boldsymbol{\Sigma}_i\boldsymbol{\gamma}}$$
$$- \frac{1}{2}\sum_{i=1}^{m} \frac{(e_i - \boldsymbol{\theta}^{*T}\boldsymbol{\epsilon}_i)^2}{{\boldsymbol{\gamma}_0}^T\boldsymbol{\Sigma}_i\boldsymbol{\gamma}_0} > 0\Big\} \to 1. \qquad \square$$

Next, we study the asymptotic normality of $\widehat{\boldsymbol{\theta}}$. Define the profile score vector to be the derivative of the profile log-likelihood:

$$\boldsymbol{\psi}(\boldsymbol{\theta}) := \frac{\partial \tilde{l}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{m} \frac{(\hat{\beta}_{Yi} - \boldsymbol{\theta}^T\widehat{\boldsymbol{\beta}}_{Xi})A_i\widehat{\boldsymbol{\beta}}_{Xi} + (\hat{\beta}_{Yi} - \boldsymbol{\theta}^T\widehat{\boldsymbol{\beta}}_{Xi})^2(\boldsymbol{\Sigma}_{Xi}\boldsymbol{\theta} - \boldsymbol{\sigma}_i)}{A_i^2}, \tag{B.5}$$

where $A_i = \boldsymbol{\gamma}^T\boldsymbol{\Sigma}_i\boldsymbol{\gamma}$, $\boldsymbol{\Sigma}_{Xi}$ is the covariance matrix of $\widehat{\boldsymbol{\beta}}_{Xi} = (\hat{\beta}_{X_1 i}, \ldots, \hat{\beta}_{X_L i})^T$, and $\boldsymbol{\sigma}_i$

is the column vector with the l-th element being the covariance between $\hat{\beta}_{X_l i}$ and $\hat{\beta}_{Y i}$.

The Taylor expansion of the l-th component of $\boldsymbol{\psi}(\widehat{\boldsymbol{\theta}})$ around the truth $\boldsymbol{\theta}^*$ can be expressed as:

$$\begin{aligned} 0 = \boldsymbol{\psi}_l(\widehat{\boldsymbol{\theta}}) &= \boldsymbol{\psi}_l(\boldsymbol{\theta}^*) + \boldsymbol{\psi}_l'(\boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) + \frac{1}{2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T \boldsymbol{\psi}_l''(\widetilde{\boldsymbol{\theta}}_l)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \\ &= \boldsymbol{\psi}_l(\boldsymbol{\theta}^*) + \left(\boldsymbol{\psi}_l'(\boldsymbol{\theta}^*) + \frac{1}{2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T \boldsymbol{\psi}_l''(\widetilde{\boldsymbol{\theta}}_l) \right) (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*), \end{aligned}$$

where $\widetilde{\boldsymbol{\theta}}_l$ is on the line segment joining $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$, $l = 1, \ldots, L$. Putting these L equations together, we have

$$0 = \boldsymbol{\psi}(\boldsymbol{\theta}^*) + (\boldsymbol{\psi}'(\boldsymbol{\theta}^*) + \frac{1}{2}\widetilde{\mathbf{Q}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*), \tag{B.6}$$

where $\widetilde{\mathbf{Q}}$ is a $L \times L$ matrix with $l-$th row given by $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T \boldsymbol{\psi}_l''(\widetilde{\boldsymbol{\theta}}_l)$.

Lemma B.3

Under the assumptions in Lemma B.2, we have

$$\mathbf{V}^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

where

$$\mathbf{V} = \mathbb{E}[-\boldsymbol{\psi}'(\boldsymbol{\theta}^*)] = \sum_{i=1}^{m} \frac{\mathbf{b}_{Xi}\mathbf{b}_{Xi}^T}{\boldsymbol{\gamma}_0^T \boldsymbol{\Sigma}_i \boldsymbol{\gamma}_0}.$$

Proof. We first show $\mathbf{V}^{-\frac{1}{2}}\boldsymbol{\psi}(\boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Denote $A_{0i} = \boldsymbol{\gamma}_0^T \boldsymbol{\Sigma}_i \boldsymbol{\gamma}_0$, we can rewrite Eq (B.5) to obtain

$$\boldsymbol{\psi}(\boldsymbol{\theta}^*) = \sum_{i=1}^{m} \frac{(e_i - \boldsymbol{\theta}^{*T}\boldsymbol{\epsilon}_i)\mathbf{b}_{Xi}}{A_{0i}} + \sum_{i=1}^{m} \frac{(e_i - \boldsymbol{\theta}^{*T}\boldsymbol{\epsilon}_i)\left(\boldsymbol{\epsilon}_i A_{0i} + (e_i - \boldsymbol{\theta}^{*T}\boldsymbol{\epsilon}_i)\boldsymbol{\xi}_i\right)}{A_{0i}^2}, \tag{B.7}$$

where $\boldsymbol{\xi}_i = \boldsymbol{\Sigma}_{Xi}\boldsymbol{\theta}^* - \boldsymbol{\sigma}_i$. Since $\mathbb{E}[e_i - \boldsymbol{\theta}^{*T}\boldsymbol{\epsilon}_i] = 0$ and $\mathrm{Var}(e_i - \boldsymbol{\theta}^{*T}\boldsymbol{\epsilon}_i) = {\boldsymbol{\gamma}_0}^T \boldsymbol{\Sigma}_i \boldsymbol{\gamma}_0$, the

first term on the right hand side is distributed as $\mathcal{N}(\mathbf{0}, \mathbf{V})$ and $\mathbf{V} = O(N)$. The second term is negligible compared to the first term since it has variance $O(m)$. Therefore, $\mathbf{V}^{-\frac{1}{2}}\boldsymbol{\psi}(\boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I})$.

We next show $\mathbf{V}^{\frac{1}{2}}(\boldsymbol{\psi}'(\boldsymbol{\theta}^*) + \frac{1}{2}\widetilde{\mathbf{Q}})^{-1}\mathbf{V}^{\frac{1}{2}} = [\mathbf{V}^{-\frac{1}{2}}(\boldsymbol{\psi}'(\boldsymbol{\theta}^*) + \frac{1}{2}\widetilde{\mathbf{Q}})\mathbf{V}^{-\frac{1}{2}}]^{-1} \xrightarrow{p} \mathbf{I}$, or equivalently $\mathbf{V}^{-\frac{1}{2}}\boldsymbol{\psi}'(\boldsymbol{\theta}^*)\mathbf{V}^{-\frac{1}{2}} + \frac{1}{2}\mathbf{V}^{-\frac{1}{2}}\widetilde{\mathbf{Q}}\mathbf{V}^{-\frac{1}{2}} \xrightarrow{p} \mathbf{I}$. Then applying Slutsky's Theorem with Eq (B.6) and $\mathbf{V}^{-\frac{1}{2}}\boldsymbol{\psi}(\boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I})$, we have $\mathbf{V}^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I})$.

To show $\mathbf{V}^{-\frac{1}{2}}\boldsymbol{\psi}'(\boldsymbol{\theta}^*)\mathbf{V}^{-\frac{1}{2}} \to \mathbf{I}$, since $\mathbb{E}[-\boldsymbol{\psi}'(\boldsymbol{\theta}^*)] = \mathbf{V}$ by definition, it suffices to show the variance of *each* element of $\mathbf{V}^{-\frac{1}{2}}\boldsymbol{\psi}'(\boldsymbol{\theta}^*)\mathbf{V}^{-\frac{1}{2}}$ goes to 0 as $N \to \infty$.

We take the similar argument as in the proof of Lemma B.2 in Zhao et al. (2020). For example, the first element in the $i-$th summand in Eq (B.5), $\boldsymbol{\psi}_{1i}(\boldsymbol{\theta})$, is a homogeneous quadratic polynomial of
$(\tilde{b}_{X_1 i}, \tilde{e}_i, \tilde{\epsilon}_{1i}, \ldots, \tilde{\epsilon}_{Li}) = (\sqrt{N}b_{X_1 i}, \sqrt{N}e_i, \sqrt{N}\epsilon_{1i}, \ldots, \sqrt{N}\epsilon_{Li})$:

$$\boldsymbol{\psi}_{1i}(\boldsymbol{\theta}) = \frac{(NA_i)(\tilde{e}_i - \sum_l \theta_l \tilde{\epsilon}_{li})\tilde{b}_{X_1 i} + (\tilde{e}_i - \sum_l \theta_l \tilde{\epsilon}_{li})(NA_i\tilde{\epsilon}_{1i} + (\tilde{e}_i - \sum_l \theta_l \tilde{\epsilon}_{li})(N\boldsymbol{\xi}_{1i}))}{(NA_i)^2},$$

and thus $\frac{\partial \boldsymbol{\psi}_{1i}(\boldsymbol{\theta})}{\partial \theta_l}$ is also a homogeneous quadratic polynomial of $(\tilde{b}_{X_1 i}, \tilde{e}_i, \tilde{\epsilon}_{1i}, \ldots, \tilde{\epsilon}_{Li})$. Also note that $\tilde{b}_{X_1 i}$ is treated as fixed and other terms such as $\mathrm{Var}(\tilde{e}_i)$, $\mathrm{Var}(\tilde{\epsilon}_{li})$, $\mathrm{Var}(\tilde{e}_i\tilde{\epsilon}_{li})$, $\mathrm{Cov}(\tilde{e}_i, \tilde{\epsilon}_{li})$ are all $O(1)$, we have $\mathrm{Var}(\frac{\partial \boldsymbol{\psi}_{1i}(\boldsymbol{\theta})}{\partial \theta_l}) = O(N)$. Meanwhile, each element in $\mathbf{V}$ is $O(N)$, and thus we can show that the variance of each element of $\mathbf{V}^{-\frac{1}{2}}\boldsymbol{\psi}'(\boldsymbol{\theta}^*)\mathbf{V}^{-\frac{1}{2}}$ goes to 0 as $N \to \infty$, leading to $\mathbf{V}^{-\frac{1}{2}}\boldsymbol{\psi}'(\boldsymbol{\theta}^*)\mathbf{V}^{-\frac{1}{2}} \xrightarrow{p} \mathbf{I}$. We can similarly show $\mathbf{V}^{-\frac{1}{2}}\widetilde{\mathbf{Q}}\mathbf{V}^{-\frac{1}{2}} \xrightarrow{p} \mathbf{0}$, since $\widehat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}^*$, and $\mathrm{Var}(\frac{\partial^2 \boldsymbol{\psi}_l(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k}) = O(N)$ for any $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}^*$, using the homogeneous quadratic polynomial argument. Therefore, $\mathbf{V}^{-\frac{1}{2}}\boldsymbol{\psi}'(\boldsymbol{\theta}^*)\mathbf{V}^{-\frac{1}{2}} + \frac{1}{2}\mathbf{V}^{-\frac{1}{2}}\widetilde{\mathbf{Q}}\mathbf{V}^{-\frac{1}{2}} \xrightarrow{p} \mathbf{I}$. Lastly by Slutsky's Theorem, we have $\mathbf{V}^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I})$. □

B.4 Simulation set-ups with weak IVs

For the first scenario with individually weak IVs, we simulated data as follows:

$$\mathbf{X}_1 = \mathbf{G}\boldsymbol{\gamma}_{X_1} + \mathbf{e}_{X_1},\ \mathbf{X}_2 = \mathbf{G}\boldsymbol{\gamma}_{X_2} + \mathbf{e}_{X_2},\ \mathbf{Y} = \theta_1\mathbf{X}_1 + \theta_2\mathbf{X}_2 + \mathbf{G}\boldsymbol{\alpha} + \mathbf{e}_Y,$$

where the unmeasured confounding was reflected in the correlated error terms $(\mathbf{e}_Y, \mathbf{e}_{X_1}, \mathbf{e}_{X_2})$. Each component of $(\mathbf{e}_Y, \mathbf{e}_{X_1}, \mathbf{e}_{X_2})$ was generated from the multivariate normal distribution $\mathcal{N}\left(\mathbf{0}, \left(\begin{smallmatrix} 1 & 0.8 & -0.7 \\ 0.8 & 1 & -0.7 \\ -0.7 & -0.7 & 1 \end{smallmatrix}\right)\right)$. $\mathbf{G}$ was the genotype matrix with 45 IVs, each generated independently from a binomial distribution with MAF 0.5 and each component of $\boldsymbol{\gamma}_{X_1}, \boldsymbol{\gamma}_{X_2}$ was generated independently from Uniform$(0, 0.1)$. The pleiotropic effect $\boldsymbol{\alpha} = \mathbf{0}$ when all IVs were valid (as considered in the main text), or $\boldsymbol{\alpha}_i \overset{iid.}{\sim} \mathcal{N}(0, 0.5^2)$ for $i = 1, \ldots, 15$ in the presence of invalid IVs.

For the second scenario, we followed the similar approach in Sanderson et al. (2021) to generate the individually strong but conditionally weak IVs as follows:

$$\begin{aligned}
&\mathbf{X}_1 = \mathbf{G}\boldsymbol{\gamma}_{X_1} + 0.9\mathbf{U}_1 + 0.1\mathbf{U}_2 + \mathbf{e}_{X_1}, \mathbf{X}_2 = \mathbf{G}\boldsymbol{\gamma}_{X_2} + 0.1\mathbf{U}_1 + 0.9\mathbf{U}_2 + \mathbf{e}_{X_2}, \\
&\mathbf{Y} = \theta_1\mathbf{X}_1 + \theta_2\mathbf{X}_2 + \mathbf{G}\boldsymbol{\alpha} + 0.5\mathbf{U}_1 + 0.5\mathbf{U}_2 + \mathbf{e}_Y, \\
&\gamma_{X_2 i} \overset{iid.}{\sim} \text{Uniform}(0, 2), \gamma_{X_1 i} = 0.7\gamma_{X_2 i} + \epsilon_i \text{ for } i = 1, \ldots, 45, \\
&\epsilon_1 = 0 \text{ and } \epsilon_i \overset{iid.}{\sim} \mathcal{N}(0, 0.075^2) \text{ for } i = 2, \ldots, 45,
\end{aligned}$$

and each component of $\mathbf{U}_1, \mathbf{U}_2$ was independently and identically distributed (iid) as $\mathcal{N}(0, 2^2)$. Each component of $(\mathbf{e}_Y, \mathbf{e}_{X_1}, \mathbf{e}_{X_2})$ was generated from the multivariate normal distribution $\mathcal{N}\left(\mathbf{0}, \left(\begin{smallmatrix} 1 & 0.8 & -0.7 \\ 0.8 & 1 & -0.7 \\ -0.7 & -0.7 & 1 \end{smallmatrix}\right)\right)$. The pleiotropic effect $\boldsymbol{\alpha} = \mathbf{0}$ when all IVs were valid (as considered in the main text), or $\boldsymbol{\alpha}_i \overset{iid.}{\sim} \mathcal{N}(0, 0.5^2)$ for $i = 1, \ldots, 15$ in the presence of invalid IVs.

For both scenarios, the GWAS summary statistics of the two exposures were computed in one sample with 20000 individuals, and that of the outcome was computed

in another independent sample with 20000 individuals.

B.5 Simulation set-ups for mediation analysis

Corresponding to the three scenarios in the main text, we simulated the GWAS summary statistics as follows:

$$
\begin{aligned}
&\hat{\beta}_{X_1 i} = \gamma_{X_1 i} + \epsilon_{X_1 i}, \quad i = 1, \ldots, 20, \\
&\hat{\beta}_{X_2 i} = 0.5\gamma_{X_1 i} + \epsilon_{X_2 i}, \quad i = 1, \ldots, K + m_1, \\
&\hat{\beta}_{X_2 i} = \gamma_{X_2 i} + 0.5\gamma_{X_1 i} + \epsilon_{X_2 i}, \quad i = K + m_1 + 1, \ldots, 20, \\
&\hat{\beta}_{Y i} = \theta_1 \gamma_{X_1 i} + \theta_2 \gamma_{X_2 i} + \alpha_i + \epsilon_{Y i}, \quad i = 1, \ldots, K, \\
&\hat{\beta}_{Y i} = \theta_1 \gamma_{X_1 i} + \theta_2 \gamma_{X_2 i} + \epsilon_{Y i}, \quad i = K + 1, \ldots, 20,
\end{aligned}
$$

where $(\epsilon_{X_1 i}, \epsilon_{X_2 i})^T \overset{iid.}{\sim} \mathcal{N}\left(\mathbf{0}, 1/n * \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right)$, $\epsilon_{Y i} \overset{iid.}{\sim} \mathcal{N}(0, 1/n)$,
$\gamma_{X_1 i}, \gamma_{X_2 i} \overset{iid.}{\sim} \mathcal{U}(0, 0.22)$, $\alpha_i \overset{iid.}{\sim} \mathcal{N}(0.1, 0.2^2)$. And $\hat{\sigma}^2_{X_l i} = \hat{\sigma}^2_{Y i} = 1/n$, $n = 50\,000$.

B.6 Estimation of correlation matrix in real data analysis

We applied bivariate LDSC regression (Bulik-Sullivan et al., 2015) on every pair of the exposure-exposure and exposure-outcome GWAS summary data, and estimated the correlations $\rho_{ll'}$'s and ρ_{lY}'s between the two GWAS summary estimates as the intercept term from the fitted bivariate LDSC regression model. As shown in the Figure B.5, most of the GWAS estimates were not highly correlated with each other except for a few such as {SBP, DBP}, {TG, HDL, LDL}, the GWAS datasets of which were collected from the same study respectively. We note that for pairs involving the

FG GWAS, the LDSC regression gave warning that *'number of SNPs less than 200k; this is almost always bad.'*, and we set them to be 0 in $\mathbf{\Sigma}_i$. For the pairs with small correlations ($|\rho| < 0.1$), we also set them to be 0 in the subsequent analysis.

Appendix C

Supplementary material for Chapter 4

Recall that our log-likelihood for the proposed model (up to some constant) is:

$$l(\theta, \mathbf{b}_X, \mathbf{r}; \hat{\boldsymbol{\beta}}_X, \hat{\boldsymbol{\beta}}_Y, \boldsymbol{\Sigma}_X, \boldsymbol{\Sigma}_Y) = -\frac{1}{2}\Big[(\hat{\boldsymbol{\beta}}_X - \mathbf{b}_X)^T \boldsymbol{\Sigma}_X^{-1}(\hat{\boldsymbol{\beta}}_X - \mathbf{b}_X) + (\hat{\boldsymbol{\beta}}_Y - \theta\mathbf{b}_X - \mathbf{r})^T \boldsymbol{\Sigma}_Y^{-1}(\hat{\boldsymbol{\beta}}_Y - \theta\mathbf{b}_X - \mathbf{r})\Big]. \quad \text{(C.1)}$$

We will first give the algorithm to solving $\{\hat{\theta}, \hat{\mathbf{b}}_X, \hat{\mathbf{r}}\} = \arg\max_{\theta, \mathbf{b}_X, \mathbf{r}} l(\theta, \mathbf{b}_X, \mathbf{r})$ under the constraint of $r_i = 0$ if $i \in \mathcal{I}$, where $\mathcal{I}$ is the set of unselected variables in Algorithm 1. We introduce the notation of $\mathbf{a}_{\mathcal{A}} = (a_i, i \in \mathcal{A}) \in \mathbb{R}^{|\mathcal{A}|}$ be a sub-vector of $\mathbf{a}$. For the simplicity of notation and without loss of generality, we assume $\mathcal{A} = \{1, 2, \ldots, K\},\ \mathcal{I} = \{K+1, \ldots, m\}$. The first-order partial derivatives of l are given by:

$$\begin{aligned}
\frac{\partial l}{\partial \mathbf{r}_{\mathcal{A}}} &= (\hat{\boldsymbol{\beta}}_Y^T - \theta\mathbf{b}_X^T - \mathbf{r}^T)\boldsymbol{\Sigma}_Y^{-1}\begin{bmatrix}\mathbf{I}_K \\ \mathbf{0}_{(m-K)\times K}\end{bmatrix}, \\
\frac{\partial l}{\partial \mathbf{b}_X} &= \theta(\hat{\boldsymbol{\beta}}_Y^T - \theta\mathbf{b}_X^T - \mathbf{r}^T)\boldsymbol{\Sigma}_Y^{-1} + (\hat{\boldsymbol{\beta}}_X^T - \mathbf{b}_X^T)\boldsymbol{\Sigma}_X^{-1}, \\
\frac{\partial l}{\partial \theta} &= (\hat{\boldsymbol{\beta}}_Y^T - \theta\mathbf{b}_X^T - \mathbf{r}^T)\boldsymbol{\Sigma}_Y^{-1}\mathbf{b}_X,
\end{aligned}$$

Equate the above partial derivatives of log-likelihood to 0, we have

$$\mathbf{r}_{\mathcal{A}}(\theta, \mathbf{b}_X; \hat{\boldsymbol{\beta}}_Y) = \begin{bmatrix} \mathbf{I}_K & \mathbf{A}^{-1}\mathbf{B} \end{bmatrix} (\hat{\boldsymbol{\beta}}_Y - \theta \mathbf{b}_X) = (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{A}}} - \theta \mathbf{b}_{X_{\mathcal{A}}}) + \mathbf{A}^{-1}\mathbf{B}(\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}}} - \theta \mathbf{b}_{X_{\mathcal{I}}}), \quad \text{(C.2)}$$

$$\mathbf{b}_X(\theta, \mathbf{r}; \hat{\boldsymbol{\beta}}_X, \hat{\boldsymbol{\beta}}_Y, \boldsymbol{\Sigma}_X, \boldsymbol{\Sigma}_Y) = (\theta^2 \boldsymbol{\Sigma}_Y^{-1} + \boldsymbol{\Sigma}_X^{-1})^{-1}(\theta \boldsymbol{\Sigma}_Y^{-1}(\hat{\boldsymbol{\beta}}_Y - \mathbf{r}) + \boldsymbol{\Sigma}_X^{-1}\hat{\boldsymbol{\beta}}_X), \quad \text{(C.3)}$$

$$\theta(\mathbf{b}_X, \mathbf{r}; \hat{\boldsymbol{\beta}}_Y, \boldsymbol{\Sigma}_Y) = \frac{(\hat{\boldsymbol{\beta}}_Y^T - \mathbf{r}^T)\boldsymbol{\Sigma}_Y^{-1}\mathbf{b}_X}{\mathbf{b}_X^T \boldsymbol{\Sigma}_Y^{-1}\mathbf{b}_X}. \quad \text{(C.4)}$$

where $\boldsymbol{\Sigma}_Y^{-1} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}$, $\mathbf{A}$ is a $K \times K$ symmetric matrix, $\mathbf{D}$ is a $(m-K) \times (m-K)$ symmetric matrix, $\mathbf{B}$ is a $K \times (m-K)$ matrix, and $\mathbf{C} = \mathbf{B}^T$.

Accordingly, we will use the following coordinate-descent algorithm to solve $\{\hat{\theta}, \hat{\mathbf{b}}_X, \hat{\mathbf{r}}\} = \arg\max_{\theta, \mathbf{b}_X, \mathbf{r}} l(\theta, \mathbf{b}_X, \mathbf{r})$ under the constraint of $r_i = 0$ if $i \in \mathcal{I}$ in Algorithm 1. At $(l+1)$-th iteration,

Step 1: Update $\mathbf{r}_{\mathcal{A}}^{(l+1)} = (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{A}}} - \theta^{(l)}\mathbf{b}_{X_{\mathcal{A}}}^{(l)}) + \mathbf{A}^{-1}\mathbf{B}(\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}}} - \theta^{(l)}\mathbf{b}_{X_{\mathcal{I}}}^{(l)})$ (Eq (C.2)), and $\mathbf{r}_{\mathcal{I}}^{(l+1)} = \mathbf{0}$;

Step 2: Update $\mathbf{b}_X^{(l+1)} = (\theta^{(l)^2}\boldsymbol{\Sigma}_Y^{-1} + \boldsymbol{\Sigma}_X^{-1})^{-1}(\theta^{(l)}\boldsymbol{\Sigma}_Y^{-1}(\hat{\boldsymbol{\beta}}_Y - \mathbf{r}^{(l+1)}) + \boldsymbol{\Sigma}_X^{-1}\hat{\boldsymbol{\beta}}_X)$ (Eq (C.3));

Step 3: Update $\theta^{(l+1)} = \frac{(\hat{\boldsymbol{\beta}}_Y^T - \mathbf{r}^{(l+1)^T})\boldsymbol{\Sigma}_Y^{-1}\mathbf{b}_X^{(l+1)}}{\mathbf{b}_X^{(l+1)^T}\boldsymbol{\Sigma}_Y^{-1}\mathbf{b}_X^{(l+1)}}$ (Eq (C.4)).

It is noted that at the convergence, only the IVs in $\mathcal{I}$ will contribute to the estimation of θ (and $\mathbf{b}_{X_{\mathcal{I}}}$). To see this, by comparing Eq (C.2) and Eq (C.4), after some simple matrix algebra, we have $\theta = \hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}}}^T \boldsymbol{\Sigma}_{Y_{\mathcal{I}}}^{-1}\mathbf{b}_{X_{\mathcal{I}}} / \mathbf{b}_{X_{\mathcal{I}}}^T \boldsymbol{\Sigma}_{Y_{\mathcal{I}}}^{-1}\mathbf{b}_{X_{\mathcal{I}}}$. And by comparing Eq (C.2) and Eq (C.3), we have $\mathbf{b}_{X_{\mathcal{I}}} = (\theta^2\boldsymbol{\Sigma}_{Y_{\mathcal{I}}}^{-1} + \boldsymbol{\Sigma}_{X_{\mathcal{I}}}^{-1})^{-1}(\theta\boldsymbol{\Sigma}_{Y_{\mathcal{I}}}^{-1}(\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}}}) + \boldsymbol{\Sigma}_{X_{\mathcal{I}}}^{-1}\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}}})$. Therefore, we have the following proposition:

Proposition C.1

Denote $\{\hat{\theta}, \hat{\mathbf{b}}_X, \hat{\mathbf{r}}\} = \arg\max_{\theta, \mathbf{b}_X, \mathbf{r}} l(\theta, \mathbf{b}_X, \mathbf{r}; \hat{\boldsymbol{\beta}}_X, \hat{\boldsymbol{\beta}}_Y, \boldsymbol{\Sigma}_X, \boldsymbol{\Sigma}_Y)$ under the constraint of $\mathbf{r}_{\mathcal{I}} = \mathbf{0}_{m-K}$, and $\{\tilde{\theta}, \tilde{\mathbf{b}}_{X_{\mathcal{I}}}\} = \arg\max_{\theta, \mathbf{b}_{X_{\mathcal{I}}}} l(\theta, \mathbf{b}_{X_{\mathcal{I}}}, \mathbf{0}_{m-K}; \hat{\boldsymbol{\beta}}_{X_{\mathcal{I}}}, \hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}}}, \boldsymbol{\Sigma}_{X_{\mathcal{I}}}, \boldsymbol{\Sigma}_{Y_{\mathcal{I}}})$. Then $\hat{\theta} = \tilde{\theta}$ and $\hat{\mathbf{b}}_{X_{\mathcal{I}}} = \tilde{\mathbf{b}}_{X_{\mathcal{I}}}$.

Proposition C.2

Assumption 4.1 is equivalent to the plurality condition stated in Theorem 1 in Guo et al. (2018).

Proof. Denote $\mathcal{S} = \{i : b_{Xi} \neq 0\}$ be the set of relevant IVs, $\mathcal{V} = \{i : b_{Xi} \neq 0, r_i = 0\} = \mathcal{S} \setminus \mathcal{A}_0$ be the set of valid IVs. The plurality condition in (Guo et al., 2018) is $|\mathcal{V}| > \max_{c\neq 0} |\{i \in \mathcal{S} : r_i/b_{Xi} = c\}|$. We will use proof by contradiction to show the equivalence between the two conditions. Now, consider two cases:

(1) When all m IVs are relevant, i.e., $\mathcal{S} = \{1, 2, \ldots, m\}$ and $\mathcal{S}^C = \varnothing$, then $\mathcal{V} = \mathcal{A}_0^C$ and $|\mathcal{V}| = m - K_0$.

($\Rightarrow$): If $\exists \mathcal{A} \neq \mathcal{A}_0$, $|\mathcal{A}| = K_0$ and $\tilde{\theta} \neq \theta$ such that $b_{Yi} = \tilde{\theta} b_{Xi}$ for all $i \in \mathcal{A}^C$, then we have $r_i/b_{Xi} = (b_{Yi} - \theta b_{Xi})/b_{Xi} = \tilde{\theta} - \theta$ for $i \in \mathcal{A}^C$, where $|\mathcal{A}^C| = m - K_0$. Therefore, $\max_{c\neq 0} |\{i \in \mathcal{S} : r_i/b_{Xi} = c\}| \geq m - K_0 = |\mathcal{V}|$.

($\Leftarrow$): If $|\mathcal{V}| \leq \max_{c\neq 0} |\{i \in \mathcal{S} : r_i/b_{Xi} = c\}|$, denote c_m be the maximizing c and $\tilde{\mathcal{V}}$ be one subset of $\{i \in \mathcal{S} : r_i/b_{Xi} = c_m\}$ and $|\tilde{\mathcal{V}}| = m - K_0$. Let $\mathcal{A} = (\mathcal{S} \setminus \mathcal{A}_0) \cup (\mathcal{A}_0 \setminus \tilde{\mathcal{V}})$, then $|\mathcal{A}| = (m - K_0) + (K_0 - m + K_0) = K_0$ and $\mathcal{A}^C = (\mathcal{S} \setminus \mathcal{A}_0)^C \cap (\mathcal{A}_0 \setminus \tilde{\mathcal{V}})^C = (\mathcal{S}^C \cup \mathcal{A}_0) \cap (\mathcal{A}_0^C \cup \tilde{\mathcal{V}}) = (\mathcal{A}_0 \cap \mathcal{A}_0^C) \cup (\mathcal{A}_0 \cap \tilde{\mathcal{V}}) = \tilde{\mathcal{V}}$. Therefore, for $i \in \mathcal{A}^C = \tilde{\mathcal{V}}$, we have $b_{Yi} = \theta b_{Xi} + r_i = (\theta + c_m) b_{Xi}$.

(2) When $\mathcal{S}^C \neq \varnothing$ and $|\mathcal{S}| = s_0$, $|\mathcal{S}^C| = m - s_0$, consider two sub-cases:

(2a) When $\mathcal{A}_0 \setminus \mathcal{S} = \varnothing$, i.e. $\mathcal{A}_0 = \{i : r_i \neq 0, b_{Xi} \neq 0\}$ or $\mathcal{A}_0 \subset \mathcal{S}$, then $|\mathcal{V}| = s_0 - K_0$.

($\Rightarrow$): If $\exists \mathcal{A} \neq \mathcal{A}_0$, $|\mathcal{A}| = K_0$ and $\tilde{\theta} \neq \theta$ such that $b_{Yi} = \tilde{\theta} b_{Xi}$ for all $i \in \mathcal{A}^C$, then we have $r_i/b_{Xi} = (b_{Yi} - \theta b_{Xi})/b_{Xi} = \tilde{\theta} - \theta$ for $i \in \mathcal{A}^C \cap \mathcal{S}$. Therefore, $\max_{c\neq 0} |\{i \in \mathcal{S} : r_i/b_{Xi} = c\}| \geq s_0 - K_0 = |\mathcal{V}|$.

($\Leftarrow$): If $|\mathcal{V}| \leq \max_{c\neq 0} |\{i \in \mathcal{S} : r_i/b_{Xi} = c\}|$, denote c_m be the maximizing c and $\tilde{\mathcal{V}}$ be one subset of $\{i \in \mathcal{S} : r_i/b_{Xi} = c_m\}$ and $|\tilde{\mathcal{V}}| = s_0 - K_0$. Let

$\mathcal{A} = (\mathcal{S} \setminus \mathcal{A}_0) \cup (\mathcal{A}_0 \setminus \tilde{\mathcal{V}})$, then $|\mathcal{A}| = (s_0 - K_0) + (K_0 - s_0 + K_0) = K_0$ and $\mathcal{A}^C = (\mathcal{S} \setminus \mathcal{A}_0)^C \cap (\mathcal{A}_0 \setminus \tilde{\mathcal{V}})^C = (\mathcal{S}^C \cup \mathcal{A}_0) \cap (\mathcal{A}_0^C \cup \tilde{\mathcal{V}}) = (\mathcal{S}^C \cap \mathcal{A}_0^C) \cup (\mathcal{A}_0 \cap \mathcal{A}_0^C) \cup (\mathcal{S}^C \cap \tilde{\mathcal{V}}) \cup (\mathcal{A}_0 \cap \tilde{\mathcal{V}}) = \mathcal{S}^C \cup \tilde{\mathcal{V}}$. Note that for $i \in \mathcal{S}^C$, $b_{Xi} = 0$ and $b_{Yi} = 0$. Therefore, we have $b_{Yi} = \theta b_{Xi} + r_i = (\theta + c_m) b_{Xi}$ for $i \in \mathcal{A}^C = \mathcal{S}^C \cup \tilde{\mathcal{V}}$.

(2b) When $\mathcal{A}_0 \setminus \mathcal{S} = \{i : r_i \neq 0, b_{Xi} = 0\} \neq \varnothing$, further denote $|\mathcal{A}_0 \setminus \mathcal{S}| = d_0$, then $|\mathcal{A}_0 \cap \mathcal{S}| = K_0 - d_0$ and $|\mathcal{V}| = s_0 - (K_0 - d_0)$.

($\Rightarrow$): If $\exists \mathcal{A} \neq \mathcal{A}_0$, $|\mathcal{A}| = K_0$ and $\tilde{\theta} \neq \theta$ such that $b_{Yi} = \tilde{\theta} b_{Xi}$ for all $i \in \mathcal{A}^C$, then we have $r_i / b_{Xi} = (b_{Yi} - \theta b_{Xi}) / b_{Xi} = \tilde{\theta} - \theta$ for $i \in \mathcal{A}^C \cap \mathcal{S}$. We note that since $b_{Yi} = \tilde{\theta} b_{Xi}$ will not hold for $i \in \mathcal{A}_0 \setminus \mathcal{S}$, $\mathcal{A}_0 \setminus \mathcal{S}$ must be a subset of $\mathcal{A}$. Therefore, $\max_{c \neq 0} |\{i \in \mathcal{S} : r_i / b_{Xi} = c\}| \geq s_0 - (K_0 - d_0) = |\mathcal{V}|$.

($\Leftarrow$): If $|\mathcal{V}| \leq \max_{c \neq 0} |\{i \in \mathcal{S} : r_i / b_{Xi} = c\}|$, denote c_m be the maximizing c and $\tilde{\mathcal{V}}$ be one subset of $\{i \in \mathcal{S} : r_i / b_{Xi} = c_m\}$ and $|\tilde{\mathcal{V}}| = s_0 - K_0 + d_0$. Let $\mathcal{A} = (\mathcal{S} \setminus \mathcal{A}_0) \cup (\mathcal{A}_0 \setminus \tilde{\mathcal{V}})$, then $|\mathcal{A}| = (s_0 - K_0 + d_0) + (K_0 - s_0 + K_0 - d_0) = K_0$ and $\mathcal{A}^C = (\mathcal{S} \setminus \mathcal{A}_0)^C \cap (\mathcal{A}_0 \setminus \tilde{\mathcal{V}})^C = (\mathcal{S}^C \cup \mathcal{A}_0) \cap (\mathcal{A}_0^C \cup \tilde{\mathcal{V}}) = (\mathcal{S}^C \cap \mathcal{A}_0^C) \cup (\mathcal{A}_0 \cap \mathcal{A}_0^C) \cup (\mathcal{S}^C \cap \tilde{\mathcal{V}}) \cup (\mathcal{A}_0 \cap \tilde{\mathcal{V}}) = (\mathcal{S}^C \cap \mathcal{A}_0^C) \cup \tilde{\mathcal{V}}$. Note that for $i \in \mathcal{S}^C$, $b_{Xi} = 0$ and for $i \in \mathcal{A}_0^C$, $r_i = 0$, hence for $i \in (\mathcal{S}^C \cap \mathcal{A}_0^C)$, $b_{Xi} = b_{Yi} = 0$. Therefore, we have $b_{Yi} = \theta b_{Xi} + r_i = (\theta + c_m) b_{Xi}$ for $i \in \mathcal{A}^C = (\mathcal{S}^C \cap \mathcal{A}_0^C) \cup \tilde{\mathcal{V}}$. □

To prove Theorem 4.1, we will first prove the selection consistency of BIC, i.e., it will select the correct set of pleiotropic IVs ($\mathcal{A}_0$) as the sample size N goes to infinity. Then we prove the estimation consistency and asymptotic normality based on the set of IVs without horizontal pleiotropy.

Lemma C.1

With Assumption 4.1 to Assumption 4.3 satisfied, if $K_0 \in \mathcal{K}$, we have $P(\hat{K} = K_0) \to 1$ and $P(\hat{\mathcal{A}}_{\hat{K}} = \mathcal{A}_0) \to 1$ as $N \to \infty$.

Proof. First, we show $P(\hat{\mathcal{A}}_{K_0} = \mathcal{A}_0) \to 1$, which is equivalent to show for any $\mathcal{A}_1 \subseteq \{1, \cdots, m\}$ such that $|\mathcal{A}_1| = K_0$ and $\mathcal{A}_1 \neq \mathcal{A}_0$, $P(\hat{\mathcal{A}}_{K_0} = \mathcal{A}_1) \to 0$ as $N \to \infty$. Denote $\mathcal{I}_0 = \mathcal{A}_0^C$, $\mathcal{I}_1 = \mathcal{A}_1^C$, then by Proposition 1, we have

$$
\begin{aligned}
&P(\hat{\mathcal{A}}_{K_0} = \mathcal{A}_1) \\
&\leq P\Big\{ \min_{\tilde{\theta}, \tilde{\mathbf{b}}_{X_{\mathcal{I}_1}}} (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_1}} - \tilde{\mathbf{b}}_{X_{\mathcal{I}_1}})^T \boldsymbol{\Sigma}_{X_{\mathcal{I}_1}}^{-1} (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_1}} - \tilde{\mathbf{b}}_{X_{\mathcal{I}_1}}) + (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_1}} - \tilde{\theta}\tilde{\mathbf{b}}_{X_{\mathcal{I}_1}})^T \boldsymbol{\Sigma}_{Y_{\mathcal{I}_1}}^{-1} (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_1}} - \tilde{\theta}\tilde{\mathbf{b}}_{X_{\mathcal{I}_1}}) \\
&\qquad \leq (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_0}} - \mathbf{b}_{X_{\mathcal{I}_0}})^T \boldsymbol{\Sigma}_{X_{\mathcal{I}_0}}^{-1} (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_1}} - \mathbf{b}_{X_{\mathcal{I}_0}}) + (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_0}} - \theta \mathbf{b}_{X_{\mathcal{I}_0}})^T \boldsymbol{\Sigma}_{Y_{\mathcal{I}_0}}^{-1} (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_0}} - \theta \mathbf{b}_{X_{\mathcal{I}_0}}) \Big\}.
\end{aligned}
$$

Note that, for $i \in \mathcal{I}_0$, $b_{Yi} = \theta b_{Xi}$, and $\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_0}} - \mathbf{b}_{X_{\mathcal{I}_0}} \sim \mathcal{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_{X_{\mathcal{I}_0}})$, $\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_0}} - \theta \mathbf{b}_{X_{\mathcal{I}_0}} \sim \mathcal{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_{Y_{\mathcal{I}_0}})$. So for any $\epsilon > 0$, as $N \to \infty$, there exists $C > 0$ such that

$$
P\Big\{ (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_0}} - \mathbf{b}_{X_{\mathcal{I}_0}})^T \boldsymbol{\Sigma}_{X_{\mathcal{I}_0}}^{-1} (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_1}} - \mathbf{b}_{X_{\mathcal{I}_0}}) + (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_0}} - \theta \mathbf{b}_{X_{\mathcal{I}_0}})^T \boldsymbol{\Sigma}_{Y_{\mathcal{I}_0}}^{-1} (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_0}} - \theta \mathbf{b}_{X_{\mathcal{I}_0}}) > C \Big\} < \frac{\epsilon}{2}. \tag{C.5}
$$

And we have

$$
\begin{aligned}
&P\Big\{ \min_{\tilde{\theta}, \tilde{\mathbf{b}}_{X_{\mathcal{I}_1}}} (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_1}} - \tilde{\mathbf{b}}_{X_{\mathcal{I}_1}})^T \boldsymbol{\Sigma}_{X_{\mathcal{I}_1}}^{-1} (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_1}} - \tilde{\mathbf{b}}_{X_{\mathcal{I}_1}}) + (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_1}} - \tilde{\theta}\tilde{\mathbf{b}}_{X_{\mathcal{I}_1}})^T \boldsymbol{\Sigma}_{Y_{\mathcal{I}_1}}^{-1} (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_1}} - \tilde{\theta}\tilde{\mathbf{b}}_{X_{\mathcal{I}_1}}) \\
&\qquad \leq (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_0}} - \mathbf{b}_{X_{\mathcal{I}_0}})^T \boldsymbol{\Sigma}_{X_{\mathcal{I}_0}}^{-1} (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_1}} - \mathbf{b}_{X_{\mathcal{I}_0}}) + (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_0}} - \theta \mathbf{b}_{X_{\mathcal{I}_0}})^T \boldsymbol{\Sigma}_{Y_{\mathcal{I}_0}}^{-1} (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_0}} - \theta \mathbf{b}_{X_{\mathcal{I}_0}}) \Big\} \\
&\leq P\Big\{ \min_{\tilde{\theta}, \tilde{\mathbf{b}}_{X_{\mathcal{I}_1}}} (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_1}} - \tilde{\mathbf{b}}_{X_{\mathcal{I}_1}})^T \boldsymbol{\Sigma}_{X_{\mathcal{I}_1}}^{-1} (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_1}} - \tilde{\mathbf{b}}_{X_{\mathcal{I}_1}}) + (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_1}} - \tilde{\theta}\tilde{\mathbf{b}}_{X_{\mathcal{I}_1}})^T \boldsymbol{\Sigma}_{Y_{\mathcal{I}_1}}^{-1} (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_1}} - \tilde{\theta}\tilde{\mathbf{b}}_{X_{\mathcal{I}_1}}) \leq C \Big\} \\
&\quad + P\Big\{ (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_0}} - \mathbf{b}_{X_{\mathcal{I}_0}})^T \boldsymbol{\Sigma}_{X_{\mathcal{I}_0}}^{-1} (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_1}} - \mathbf{b}_{X_{\mathcal{I}_0}}) + (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_0}} - \theta \mathbf{b}_{X_{\mathcal{I}_0}})^T \boldsymbol{\Sigma}_{Y_{\mathcal{I}_0}}^{-1} (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_0}} - \theta \mathbf{b}_{X_{\mathcal{I}_0}}) > C \Big\}.
\end{aligned}
$$

After profiling out $\tilde{\mathbf{b}}_{X_{\mathcal{I}_1}}$, we get

$$\begin{aligned}
&\min_{\tilde{\theta},\tilde{\mathbf{b}}_{X_{\mathcal{I}_1}}} (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_1}} - \tilde{\mathbf{b}}_{X_{\mathcal{I}_1}})^T \boldsymbol{\Sigma}_{X_{\mathcal{I}_1}}^{-1} (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_1}} - \tilde{\mathbf{b}}_{X_{\mathcal{I}_1}}) + (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_1}} - \tilde{\theta}\tilde{\mathbf{b}}_{X_{\mathcal{I}_1}})^T \boldsymbol{\Sigma}_{Y_{\mathcal{I}_1}}^{-1} (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_1}} - \tilde{\theta}\tilde{\mathbf{b}}_{X_{\mathcal{I}_1}}) \\
&= \min_{\tilde{\theta}} (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_1}} - \tilde{\theta}\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_1}})^T (\boldsymbol{\Sigma}_{Y_{\mathcal{I}_1}} + \tilde{\theta}^2 \boldsymbol{\Sigma}_{X_{\mathcal{I}_1}})^{-1} (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_1}} - \tilde{\theta}\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_1}}),
\end{aligned}$$

so

$$\begin{aligned}
&P\Big\{ \min_{\tilde{\theta},\tilde{\mathbf{b}}_{X_{\mathcal{I}_1}}} (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_1}} - \tilde{\mathbf{b}}_{X_{\mathcal{I}_1}})^T \boldsymbol{\Sigma}_{X_{\mathcal{I}_1}}^{-1} (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_1}} - \tilde{\mathbf{b}}_{X_{\mathcal{I}_1}}) + (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_1}} - \tilde{\theta}\tilde{\mathbf{b}}_{X_{\mathcal{I}_1}})^T \boldsymbol{\Sigma}_{Y_{\mathcal{I}_1}}^{-1} (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_1}} - \tilde{\theta}\tilde{\mathbf{b}}_{X_{\mathcal{I}_1}}) \le C \Big\} \\
&= P\Big\{ \min_{\tilde{\theta}} (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_1}} - \tilde{\theta}\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_1}})^T (\boldsymbol{\Sigma}_{Y_{\mathcal{I}_1}} + \tilde{\theta}^2 \boldsymbol{\Sigma}_{X_{\mathcal{I}_1}})^{-1} (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_1}} - \tilde{\theta}\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_1}}) \le C \Big\}.
\end{aligned}$$

We have $(\boldsymbol{\Sigma}_{Y_{\mathcal{I}_1}} + \tilde{\theta}^2 \boldsymbol{\Sigma}_{X_{\mathcal{I}_1}})^{-1/2} (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_1}} - \tilde{\theta}\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_1}}) \sim \mathcal{MVN}(\boldsymbol{\mu}_{\tilde{\theta}}, \mathbf{I}_{m-K_0})$, where $\boldsymbol{\mu}_{\tilde{\theta}} = (\boldsymbol{\Sigma}_{Y_{\mathcal{I}_1}} + \tilde{\theta}^2 \boldsymbol{\Sigma}_{X_{\mathcal{I}_1}})^{-1/2} (\theta \mathbf{b}_{X_{\mathcal{I}_1}} + \mathbf{r}_{\mathcal{I}_1} - \tilde{\theta}\mathbf{b}_{X_{\mathcal{I}_1}})$. Hence $(\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_1}} - \tilde{\theta}\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_1}})^T (\boldsymbol{\Sigma}_{Y_{\mathcal{I}_1}} + \tilde{\theta}^2 \boldsymbol{\Sigma}_{X_{\mathcal{I}_1}})^{-1} (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_1}} - \tilde{\theta}\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_1}})$ follows non-central χ^2 distribution with degrees of freedom $(m - K_0)$ and non-centrality parameter $\lambda_{\tilde{\theta}}$ depending on $\tilde{\theta}$:

$$\lambda_{\tilde{\theta}} = (\theta \mathbf{b}_{X_{\mathcal{I}_1}} + \mathbf{r}_{\mathcal{I}_1} - \tilde{\theta}\mathbf{b}_{X_{\mathcal{I}_1}})^T (\boldsymbol{\Sigma}_{Y_{\mathcal{I}_1}} + \tilde{\theta}^2 \boldsymbol{\Sigma}_{X_{\mathcal{I}_1}})^{-1} (\theta \mathbf{b}_{X_{\mathcal{I}_1}} + \mathbf{r}_{\mathcal{I}_1} - \tilde{\theta}\mathbf{b}_{X_{\mathcal{I}_1}}).$$

With Assumption 4.1, there is no $\tilde{\theta}$ making $\theta \cdot b_{Xi} + r_i - \tilde{\theta} \cdot b_{Xi} = 0$ for all $i \in \mathcal{A}_1^c$ simultaneously, and with Assumption 4.3, the elements of $(\boldsymbol{\Sigma}_{Y_{\mathcal{I}_1}} + \tilde{\theta}^2 \boldsymbol{\Sigma}_{X_{\mathcal{I}_1}})^{-1}$ go to infinity as $N \to \infty$. So we have $\min_{\tilde{\theta}} \lambda_{\tilde{\theta}} \to \infty$ as $N \to \infty$. Then as N large enough, we have

$$P\Big\{ \min_{\tilde{\theta}} (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_1}} - \tilde{\theta}\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_1}})^T (\boldsymbol{\Sigma}_{Y_{\mathcal{I}_1}} + \tilde{\theta}^2 \boldsymbol{\Sigma}_{X_{\mathcal{I}_1}})^{-1} (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_1}} - \tilde{\theta}\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_1}}) \le C \Big\} \le \frac{\epsilon}{2}. \tag{C.6}$$

Combining (Eq (C.5)) and (Eq (C.6)), we get $P(\hat{\mathcal{A}}_{K_0} = \mathcal{A}_0) \to 1$ as $N \to \infty$.

Next, we show $P(\hat{K} = K_0) \to 1$. For any $K_1 < K_0$, we have

$$\begin{aligned} P(\hat{K} = K_1) &\leq P\big\{\mathrm{BIC}(K_1) \leq \mathrm{BIC}(K_0)\big\} \\ &= P\Big\{2l\left(\hat{\theta}(K_0), \hat{\mathbf{b}}_X(K_0), \hat{\mathbf{r}}(K_0)\right) - 2l\left(\hat{\theta}(K_1), \hat{\mathbf{b}}_X(K_1), \hat{\mathbf{r}}(K_1)\right) \leq \log(N)(K_0 - K_1)\Big\}. \end{aligned}$$

As we have shown $P(\hat{\mathcal{A}}_{K_0} = \mathcal{A}_0) \to 1$, with probability goes to 1 we have

$$\begin{aligned} &2l\left(\hat{\theta}(K_0), \hat{\mathbf{b}}_X(K_0), \hat{\mathbf{r}}(K_0)\right) - 2l\left(\hat{\theta}(K_1), \hat{\mathbf{b}}_X(K_1), \hat{\mathbf{r}}(K_1)\right) \\ \geq &\min_{\tilde{\theta}, \tilde{\mathbf{b}}_{X_{\mathcal{I}_{K_1}}}} (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_{K_1}}} - \tilde{\mathbf{b}}_{X_{\mathcal{I}_{K_1}}})^T \boldsymbol{\Sigma}^{-1}_{X_{\mathcal{I}_{K_1}}} (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_{K_1}}} - \tilde{\mathbf{b}}_{X_{\mathcal{I}_{K_1}}}) + (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_{K_1}}} - \tilde{\theta}\tilde{\mathbf{b}}_{X_{\mathcal{I}_{K_1}}})^T \boldsymbol{\Sigma}^{-1}_{Y_{\mathcal{I}_{K_1}}} (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_{K_1}}} - \tilde{\theta}\tilde{\mathbf{b}}_{X_{\mathcal{I}_{K_1}}}) \\ &- (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_0}} - \mathbf{b}_{X_{\mathcal{I}_0}})^T \boldsymbol{\Sigma}^{-1}_{X_{\mathcal{I}_0}} (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_1}} - \mathbf{b}_{X_{\mathcal{I}_0}}) + (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_0}} - \theta \mathbf{b}_{X_{\mathcal{I}_0}})^T \boldsymbol{\Sigma}^{-1}_{Y_{\mathcal{I}_0}} (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_0}} - \theta \mathbf{b}_{X_{\mathcal{I}_0}}). \end{aligned}$$

Then we get

$$\begin{aligned} P(\hat{K} = K_1) \leq \sum_{|\mathcal{I}| = m - K_1} P\Big\{ &\min_{\tilde{\theta}, \tilde{\mathbf{b}}_{X_{\mathcal{I}}}} (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}}} - \tilde{\mathbf{b}}_{X_{\mathcal{I}}})^T \boldsymbol{\Sigma}^{-1}_{X_{\mathcal{I}}} (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}}} - \tilde{\mathbf{b}}_{X_{\mathcal{I}}}) + \\ &(\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}}} - \tilde{\theta}\tilde{\mathbf{b}}_{X_{\mathcal{I}}})^T \boldsymbol{\Sigma}^{-1}_{Y_{\mathcal{I}}} (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}}} - \tilde{\theta}\tilde{\mathbf{b}}_{X_{\mathcal{I}}}) \\ \leq &(\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_0}} - \mathbf{b}_{X_{\mathcal{I}_0}})^T \boldsymbol{\Sigma}^{-1}_{X_{\mathcal{I}_0}} (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_1}} - \mathbf{b}_{X_{\mathcal{I}_0}}) + \\ &(\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_0}} - \theta \mathbf{b}_{X_{\mathcal{I}_0}})^T \boldsymbol{\Sigma}^{-1}_{Y_{\mathcal{I}_0}} (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_0}} - \theta \mathbf{b}_{X_{\mathcal{I}_0}}) + \log(N)(K_0 - K_1)\Big\}. \end{aligned}$$

Similar as above, we get

$$\begin{aligned} &\min_{\tilde{\theta}, \tilde{\mathbf{b}}_{X_{\mathcal{I}}}} (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}}} - \tilde{\mathbf{b}}_{X_{\mathcal{I}_1}})^T \boldsymbol{\Sigma}^{-1}_{X_{\mathcal{I}}} (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}}} - \tilde{\mathbf{b}}_{X_{\mathcal{I}}}) + (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}}} - \tilde{\theta}\tilde{\mathbf{b}}_{X_{\mathcal{I}}})^T \boldsymbol{\Sigma}^{-1}_{Y_{\mathcal{I}}} (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}}} - \tilde{\theta}\tilde{\mathbf{b}}_{X_{\mathcal{I}}}) \\ &= \min_{\tilde{\theta}} (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}}} - \tilde{\theta}\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}}})^T (\boldsymbol{\Sigma}^{-1}_{Y_{\mathcal{I}}} + \tilde{\theta}^2 \boldsymbol{\Sigma}^{-1}_{X_{\mathcal{I}}})^{-1} (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}}} - \tilde{\theta}\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}}}), \end{aligned}$$

and $(\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}}} - \tilde{\theta}\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}}})^T (\boldsymbol{\Sigma}_{Y_{\mathcal{I}}} + \tilde{\theta}^2 \boldsymbol{\Sigma}_{X_{\mathcal{I}}})^{-1} (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}}} - \tilde{\theta}\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}}})$ follows non-central χ^2 distribution with degrees of freedom $(m - K_1)$ and non-centrality parameter $\lambda_{\tilde{\theta}}$. Similarly, since

$K_1 < K_0$, and with Assumption 4.1 and Assumption 4.3 we have $\lambda_{\tilde{\theta}} = O(N)$, so for any $|\mathcal{A}| = K_1$, we get

$$P\Big\{ \min_{\tilde{\theta},\tilde{\mathbf{b}}_{X_\mathcal{I}}} (\hat{\boldsymbol{\beta}}_{X_\mathcal{I}} - \tilde{\mathbf{b}}_{X_\mathcal{I}})^T \boldsymbol{\Sigma}_{X_\mathcal{I}}^{-1} (\hat{\boldsymbol{\beta}}_{X_\mathcal{I}} - \tilde{\mathbf{b}}_{X_\mathcal{I}}) + (\hat{\boldsymbol{\beta}}_{Y_\mathcal{I}} - \tilde{\theta}\tilde{\mathbf{b}}_{X_\mathcal{I}})^T \boldsymbol{\Sigma}_{Y_\mathcal{I}}^{-1} (\hat{\boldsymbol{\beta}}_{Y_\mathcal{I}} - \tilde{\theta}\tilde{\mathbf{b}}_{X_\mathcal{I}})$$
$$\leq (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_0}} - \mathbf{b}_{X_{\mathcal{I}_0}})^T \boldsymbol{\Sigma}_{X_{\mathcal{I}_0}}^{-1} (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_1}} - \mathbf{b}_{X_{\mathcal{I}_0}}) + (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_0}} - \theta\mathbf{b}_{X_{\mathcal{I}_0}})^T \boldsymbol{\Sigma}_{Y_{\mathcal{I}_0}}^{-1} (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_0}} - \theta\mathbf{b}_{X_{\mathcal{I}_0}})$$
$$+ \log(N)(K_0 - K_1)\Big\} \to 0, \ as \ N \to \infty.$$

This gives us $P(\hat{K} = K_1) \to 0$ for any $K_1 < K_0$. For any $K_1 > K_0$, we have

$$P(\hat{K} = K_1) \leq P\Big\{ \log(N)(K_1 - K_0) \leq$$
$$(\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_0}} - \mathbf{b}_{X_{\mathcal{I}_0}})^T \boldsymbol{\Sigma}_{X_{\mathcal{I}_0}}^{-1} (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_1}} - \mathbf{b}_{X_{\mathcal{I}_0}}) + (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_0}} - \theta\mathbf{b}_{X_{\mathcal{I}_0}})^T \boldsymbol{\Sigma}_{Y_{\mathcal{I}_0}}^{-1} (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_0}} - \theta\mathbf{b}_{X_{\mathcal{I}_0}})\Big\}$$

Since $(\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_0}} - \mathbf{b}_{X_{\mathcal{I}_0}})^T \boldsymbol{\Sigma}_{X_{\mathcal{I}_0}}^{-1} (\hat{\boldsymbol{\beta}}_{X_{\mathcal{I}_1}} - \mathbf{b}_{X_{\mathcal{I}_0}}) + (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_0}} - \theta\mathbf{b}_{X_{\mathcal{I}_0}})^T \boldsymbol{\Sigma}_{Y_{\mathcal{I}_0}}^{-1} (\hat{\boldsymbol{\beta}}_{Y_{\mathcal{I}_0}} - \theta\mathbf{b}_{X_{\mathcal{I}_0}})$ is a central chi-square with degrees of freedom $2(m - K_0)$, we get $P(\hat{K} = K_1) \to 0$ for any $K_1 > K_0$. So we have $P(\hat{K} = K_0) \to 1$ as $N \to \infty$. □

Suppose we have correctly selected IVs in $\mathcal{A}_0$, then the resulted cMLE of θ is the same as the maximum profile likelihood estimator (MPLE) (profiling out $\mathbf{b}_{X_{\mathcal{I}_0}}$) based on IVs in $\mathcal{I}_0$. And we will prove the consistency and asymptotic normality based on the MPLE. For the simplicity of notation, we assume that all m IVs have $r_i = 0$ from now on, i.e.,

$$\begin{aligned} \hat{\boldsymbol{\beta}}_X &\sim \mathcal{MVN}(\mathbf{b}_X, \boldsymbol{\Sigma}_X), \\ \hat{\boldsymbol{\beta}}_Y &\sim \mathcal{MVN}(\theta\mathbf{b}_X, \boldsymbol{\Sigma}_Y). \end{aligned} \tag{C.7}$$

Then the profile log-likelihood is

$$\tilde{l}(\theta) = -\frac{1}{2}(\hat{\boldsymbol{\beta}}_Y - \theta\hat{\boldsymbol{\beta}}_X)^T (\boldsymbol{\Sigma}_Y + \theta^2\boldsymbol{\Sigma}_X)^{-1} (\hat{\boldsymbol{\beta}}_Y - \theta\hat{\boldsymbol{\beta}}_X). \tag{C.8}$$

The maximum likelihood estimator of θ is given by $\hat{\theta} = \arg\max_\theta \tilde{l}(\theta)$.

Lemma C.2

Under model Eq (C.7) and Assumption 4.1 to Assumption 4.3, the maximum likelihood estimator $\hat{\theta}$ of Eq (C.8) is consistent for the true causal parameter θ_0, that is, $\hat{\theta} \xrightarrow{p} \theta_0$ as $N \to \infty$.

Proof. Let $\mathbf{e} = \hat{\boldsymbol{\beta}}_Y - \mathbf{b}_Y$, $\boldsymbol{\epsilon} = \hat{\boldsymbol{\beta}}_X - \mathbf{b}_X$, $\mathbf{b}_Y = \theta_0 \mathbf{b}_X$. After some algebra, we have

$$\begin{aligned}
-2\tilde{l}(\theta) =& (\mathbf{e} - \theta\boldsymbol{\epsilon})^T(\boldsymbol{\Sigma}_Y + \theta^2\boldsymbol{\Sigma}_X)^{-1}(\mathbf{e} - \theta\boldsymbol{\epsilon}) + \\
& (\theta_0 - \theta)^2 \mathbf{b}_X^T(\boldsymbol{\Sigma}_Y + \theta^2\boldsymbol{\Sigma}_X)^{-1}\mathbf{b}_X + 2(\theta_0 - \theta)\mathbf{b}_X^T(\boldsymbol{\Sigma}_Y + \theta^2\boldsymbol{\Sigma}_X)^{-1}(\mathbf{e} - \theta\boldsymbol{\epsilon}), \\
-2\tilde{l}(\theta_0) =& (\mathbf{e} - \theta_0\boldsymbol{\epsilon})^T(\boldsymbol{\Sigma}_Y + \theta_0^2\boldsymbol{\Sigma}_X)^{-1}(\mathbf{e} - \theta_0\boldsymbol{\epsilon}).
\end{aligned}$$

Notice that $(\mathbf{e} - \theta\boldsymbol{\epsilon})^T(\boldsymbol{\Sigma}_Y + \theta^2\boldsymbol{\Sigma}_X)^{-1}(\mathbf{e} - \theta\boldsymbol{\epsilon})$ and $(\mathbf{e} - \theta_0\boldsymbol{\epsilon})^T(\boldsymbol{\Sigma}_Y + \theta_0^2\boldsymbol{\Sigma}_X)^{-1}(\mathbf{e} - \theta_0\boldsymbol{\epsilon})$ follow a χ^2 with degrees of freedom m, and they are both $O_p(1)$ as we consider a fixed number of IVs. Furthermore, given $||\mathbf{b}_X||_2^2$ is bounded and $\mathbf{b}_X \neq \mathbf{0}$, and by Assumption 3, the elements of $(\boldsymbol{\Sigma}_Y + \theta^2\boldsymbol{\Sigma}_X)^{-1}$ are all $O(N)$, thus we have $(\theta_0 - \theta)^2\mathbf{b}_X^T(\boldsymbol{\Sigma}_Y + \theta^2\boldsymbol{\Sigma}_X)^{-1}\mathbf{b}_X$ is $O_p(N)$, and $(\theta_0 - \theta)\mathbf{b}_X^T(\boldsymbol{\Sigma}_Y + \theta^2\boldsymbol{\Sigma}_X)^{-1}(\mathbf{e} - \theta\boldsymbol{\epsilon})$ is $O_p(\sqrt{N})$, when $|\theta - \theta_0| > \epsilon$.

Therefore, we have $P\left(\tilde{l}(\theta_0) > \sup_{|\theta - \theta_0| > \epsilon} \tilde{l}(\theta)\right) \to 1$ as $N \to \infty$. □

Next, we study the asymptotic normality of $\hat{\theta}$. Define the profile score to be the derivative of the profile log-likelihood:

$$\begin{aligned}
\psi(\theta) := \tilde{l}'(\theta) = \theta(\hat{\boldsymbol{\beta}}_Y - \theta\hat{\boldsymbol{\beta}}_X)^T(\boldsymbol{\Sigma}_Y + \theta^2\boldsymbol{\Sigma}_X)^{-1}\boldsymbol{\Sigma}_X(\boldsymbol{\Sigma}_Y + \theta^2\boldsymbol{\Sigma}_X)^{-1}(\hat{\boldsymbol{\beta}}_Y - \theta\hat{\boldsymbol{\beta}}_X) \\
+(\hat{\boldsymbol{\beta}}_Y - \theta\hat{\boldsymbol{\beta}}_X)^T(\boldsymbol{\Sigma}_Y + \theta^2\boldsymbol{\Sigma}_X)^{-1}\hat{\boldsymbol{\beta}}_X.
\end{aligned} \tag{C.9}$$

The Taylor expansion of $\psi(\hat{\theta})$ around the truth θ_0 can be expressed as:

$$0 = \psi(\hat{\theta}) = \psi(\theta_0) + \psi'(\theta_0)(\hat{\theta} - \theta_0) + \frac{1}{2}\psi''(\tilde{\theta})(\hat{\theta} - \theta_0)^2,$$

where $\tilde{\theta}$ is between $\hat{\theta}$ and θ_0. Then we have

$$\sqrt{V}(\hat{\theta} - \theta_0) = \frac{-\psi(\theta_0)/\sqrt{V}}{\psi'(\theta_0)/V + (1/2)\psi''(\tilde{\theta})(\hat{\theta} - \theta_0)/V}, \tag{C.10}$$

where

$$V = \mathbb{E}[-\psi'(\theta_0)] = \mathbf{b}_X^T(\mathbf{\Sigma}_Y + \theta_0^2\mathbf{\Sigma}_X)^{-1}\mathbf{b}_X. \tag{C.11}$$

The nominator of Eq (C.10) can be proved to converge in distribution to $\mathcal{N}(0,1)$, the first term in the denominator of Eq (C.10) can be proved to converge in probability to 1 and the second term in the denominator an be proved to be negligible given $\hat{\theta}$ is consistent. We first prove these three statements, and by Slutsky's Theorem, the following asymptotic normality of $\hat{\theta}$ can be established.

Lemma C.3

Under the assumptions in Lemma C.2, as $N \to \infty$, we have

$$\sqrt{V}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0,1),$$

where $V = \mathbf{b}_X^T(\mathbf{\Sigma}_Y + \theta_0^2\mathbf{\Sigma}_X)^{-1}\mathbf{b}_X$.

Proof. We first show $(1/\sqrt{V})\psi(\theta_0) \xrightarrow{d} \mathcal{N}(0,1)$.

Denote $\mathbf{e} = \hat{\boldsymbol{\beta}}_Y - \mathbf{b}_Y$, $\boldsymbol{\epsilon} = \hat{\boldsymbol{\beta}}_X - \mathbf{b}_X$, $\mathbf{U}(\theta) = (\mathbf{\Sigma}_Y + \theta^2\mathbf{\Sigma}_X)^{-1}$, $\mathbf{W}(\theta) = -U'(\theta)/(2\theta) = (\mathbf{\Sigma}_Y + \theta^2\mathbf{\Sigma}_X)^{-1}\mathbf{\Sigma}_X(\mathbf{\Sigma}_Y + \theta^2\mathbf{\Sigma}_X)^{-1}$, after some algebra, we can rewrite

Eq (C.9) to obtain

$$\psi(\theta_0) = \mathbf{b}_X^T \mathbf{U}(\theta_0)(\mathbf{e} - \theta_0\boldsymbol{\epsilon}) + \left[\theta_0(\mathbf{e} - \theta_0\boldsymbol{\epsilon})^T \mathbf{W}(\theta_0)(\mathbf{e} - \theta_0\boldsymbol{\epsilon}) + (\mathbf{e} - \theta_0\boldsymbol{\epsilon})^T \mathbf{U}(\theta_0)\boldsymbol{\epsilon}\right] \tag{C.12}$$

Since $\mathbf{e} - \theta_0\boldsymbol{\epsilon} \sim \mathcal{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_Y + \theta_0^2\boldsymbol{\Sigma}_X)$, the first term on the right hand side in Eq (C.12) is distributed as $\mathcal{N}(0, V)$ and $V = \Theta(N||\mathbf{b}_X||_2^2)$. The second term is negligible compared to the first term since it has variance $O(m)$. Therefore, $(1/\sqrt{V})\psi(\theta_0) \to \mathcal{N}(0, 1)$.

We next show $(-1/V)\psi'(\theta_0) \xrightarrow{p} 1$.

Since $\mathbb{E}[-\psi'(\theta_0)] = V$ by definition, it suffices to show $\text{Var}(\psi'(\theta_0)/V) \to 0$. Rewrite Eq (C.9) using $\mathbf{e}$, $\boldsymbol{\epsilon}$, $\mathbf{U}(\theta)$ and $\mathbf{W}(\theta)$, we have

$$\begin{aligned}\psi(\theta) = \theta(\mathbf{e} - \theta\boldsymbol{\epsilon})^T \mathbf{W}(\theta)(\mathbf{e} - \theta\boldsymbol{\epsilon}) + \theta(\theta_0 - \theta)^2 \mathbf{b}_X^T \mathbf{W}(\theta)\mathbf{b}_X + \\ 2\theta(\theta_0 - \theta)(\mathbf{e} - \theta\boldsymbol{\epsilon})^T \mathbf{W}(\theta)\mathbf{b}_X + (\mathbf{e} - \theta\boldsymbol{\epsilon})^T \mathbf{U}(\theta)\boldsymbol{\epsilon} + \\ (\theta_0 - \theta)\mathbf{b}_X^T \mathbf{U}(\theta)\boldsymbol{\epsilon} + \mathbf{b}_X^T \mathbf{U}(\theta)(\mathbf{e} - \theta\boldsymbol{\epsilon}) + (\theta_0 - \theta)\mathbf{b}_X^T \mathbf{U}(\theta)\mathbf{b}_X.\end{aligned} \tag{C.13}$$

Notice that $\psi(\theta)$ (Eq (C.13)) is a homogeneous quadratic polynomial of $(\tilde{\mathbf{b}}_X, \tilde{\mathbf{e}}, \tilde{\boldsymbol{\epsilon}}) = (\sqrt{N}\mathbf{b}_X, \sqrt{N}\mathbf{e}, \sqrt{N}\boldsymbol{\epsilon})$, it is easy to see that $\psi'(\theta)$ is also a homogeneous quadratic polynomial of $(\tilde{\mathbf{b}}_X, \tilde{\mathbf{e}}, \tilde{\boldsymbol{\epsilon}})$. Also given that $\tilde{\mathbf{b}}_X$ is treated as fixed, $\tilde{e}_i \perp\!\!\!\perp \tilde{\epsilon}_j$ and other terms such as $\text{Cov}(\tilde{\mathbf{e}})$, $\text{Cov}(\tilde{\boldsymbol{\epsilon}})$, $\text{Var}(\tilde{\mathbf{e}}^T\tilde{\boldsymbol{\epsilon}})$ are all $O(1)$, we have $\text{Var}(\psi'(\theta)) = O(N||\mathbf{b}_X||_2^2)$ for any θ in a neighbourhood $\mathcal{N}$ of θ_0. Then it is easy to see that $\text{Var}(\psi'(\theta_0)) \ll V^2 = \Theta(N^2||\mathbf{b}_X||^4)$.

Since $\psi'(\theta)$ is a homogeneous quadratic polynomial of $(\tilde{\mathbf{b}}_X, \tilde{\mathbf{e}}, \tilde{\boldsymbol{\epsilon}})$, so is $\psi''(\theta)$, and we can apply the same argument for $\psi'(\theta_0)$ above to $\psi''(\theta)$ and obtain that for a neighborhood $\mathcal{N}$ of θ_0, $\sup_{\theta\in\mathcal{N}}(1/V)\psi''(\theta) = O_p(1)$ and $\psi''(\tilde{\theta})(\hat{\theta} - \theta_0)/V = o_p(1)$

Lastly by Slutsky's Theorem, we have $\sqrt{V}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, 1)$. □

With Lemma C.1 to Lemma C.3, we can conclude Theorem 4.1. Here are a few points we would like to highlight. Our proposed method can only consistently select the invalid IVs with $r_i \neq 0$ (violating the second and the third IV assumptions stated in the main text), but it cannot select the invalid IVs only violating the first IV assumption, i.e. $b_{Xi} = r_i = 0$. One one hand, the first IV assumption can be tested empirically and less prone to be violated by using IVs jointly associated with the exposure. On the other hand, even though we mistakenly include some irrelevant IVs in cisMR-cML, it will not affect the validity of the statistical inference for the causal parameter. This can be seen from the proof of Lemma C.2 and Lemma C.3 that, as long as there are valid IVs such that $\mathbf{b}_X \neq \mathbf{0}$, having a few zero elements in $\mathbf{b}_X$ will not affect the result. Therefore, the proposed method is robust to the violation of any or all of the three IV assumptions under the plurality assumption.