medRxiv preprint doi: https://doi.org/10.1101/2021.03.02.21252801; this version posted March 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

A Two-Sample Robust Bayesian Mendelian Randomization Method Accounting for Linkage Disequilibrium and Idiosyncratic Pleiotropy with Applications to the COVID-19 Outcome

Anqi Wang, Zhonghua Liu*

Department of Statistics and Actuarial Science, University of Hong Kong, Hong Kong SAR, China

* To whom correspondence should be addressed: zhhliu@hku.hk

March 29, 2021

ABSTRACT

Mendelian randomization (MR) is a statistical method exploiting genetic variants as instrumental variables to estimate the causal effect of modifiable risk factors on an outcome of interest. Despite wide uses of various popular two-sample MR methods based on genome-wide association study summary level data, however, those methods could suffer from potential power loss or/and biased inference when the chosen genetic variants are in linkage disequilibrium (LD), and also have relatively large direct effects on the outcome whose distribution might be heavy-tailed which is commonly referred to as the idiosyncratic pleiotropy phenomenon. To resolve those two issues, we propose a novel Robust Bayesian Mendelian Randomization (RBMR) model that uses the more robust multivariate generalized t-distribution (Arellano-Valle and Bolfarine, 1995) to model such direct effects in a probabilistic model framework which can also incorporate the LD structure explicitly. The generalized t-distribution can be represented as a Gaussian scaled mixture so that our model parameters can be estimated by the EM-type algorithms. We compute the standard errors by calibrating the evidence lower bound using the likelihood ratio test. Through extensive simulation studies, we show that our RBMR has robust performance compared to other competing methods. We also apply our RBMR method to two benchmark data sets and find that RBMR has smaller bias and standard errors. Using our proposed RBMR method, we find that coronary artery disease is associated with increased risk of critically

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.03.02.21252801; this version posted March 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

ill coronavirus disease 2019 (COVID-19). We also develop a user-friendly R package *RBMR* (https://github.com/AnqiWang2021/RBMR) for public use.

Key Words: COVID-19 outcome; Mendelian randomization; Idiosyncratic pleiotropy; Linkage disequilibrium, Multivariate generalized t-distribution; EM-type algorithm

medRxiv preprint doi: https://doi.org/10.1101/2021.03.02.21252801; this version posted March 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

1. Introduction

Mendelian randomization (MR) is a useful statistical method that leverages genetic variants as instrumental variables (IVs) for assessing the causal effect of a modifiable risk factor on a health outcome of interest even in the presence of unmeasured confounding factors (Ebrahim and Smith, 2008; Lawlor et al., 2008; Evans and Davey Smith, 2015). Because of the inborn nature of genetic variants, the associations between genetic variants and phenotypes after adjusting for possible population stratification will not be confounded by the environmental factors, socio-economic status and life styles after birth. Genome-wide association studies (GWAS) have identified tens of thousands of common genetic variants associated with thousands of complex traits and diseases (MacArthur et al., 2017). Those GWAS summary level data contain rich information about genotype-phenotype associations (https://www.ebi.ac.uk/gwas/), and thus provide us valuable resources for MR studies. Therefore, we have seen a boost of two-sample MR method developments and applications based on GWAS summary statistics recently due to the increasing availability of candidate genetic variant IVs for thousands of phenotypes. (Burgess et al., 2013; Bowden et al., 2015; Pickrell et al., 2016). In particular, a genetic variant serving as a valid IV must satisfy the following three core assumptions (Martens et al., 2006; Lawlor et al., 2008):

1. **Relevance:** The genetic variant must be associated (not necessarily causally) with the exposure;

2. **Effective Random Assignment:** The genetic variant must be independent of any (measured or unmeasured) confounders of the exposure-outcome relationship;

3. **Exclusion Restriction:** The genetic variant must affect the outcome only through the exposure, that is, the genetic variant must have no direct effect on the outcome not mediated by the exposure.

When these three core IV assumptions hold, the inverse variance weighted (IVW) (Ehret et al., 2011) method can be simply used to obtain unbiased causal effect estimate of the exposure on the outcome. However, among those three core assumptions, only the IV relevance assumption can be empirically tested, for example, by checking the empirical association strength between the candidate IV and the exposure using the GWAS catalog (https://www.ebi.ac.uk/gwas/). The association between the IV and the exposure must

medRxiv preprint doi: https://doi.org/10.1101/2021.03.02.21252801; this version posted March 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

be strong enough (the IV explains a large amount of the variation of the exposure variable) to ensure unbiased causal effect estimate. The problem of weak IVs has been studied previously in the econometric literature (Bound et al., 1995; Hansen et al., 2008). In MR settings, the method that uses genetic score by combining multiple weak IVs together to increase the IV-exposure association strength to reduce weak IV bias has also been proposed (Evans et al., 2013). Unfortunately, the other two IV core assumptions cannot be empirically tested and might be violated in practice. Violation of the exclusion restriction assumption can occur when the genetic variant indeed has a non-null direct effect on the outcome not mediated by the exposure, referred to as systematic pleiotropy (Solovieff et al., 2013; Verbanck et al., 2018; Zhao et al., 2020b). However, very often, genetic variants might have relatively large direct effects whose distribution exhibit heavy-tailed pattern, a phenomenon referred to as the idiosyncratic pleiotropy in this paper.

To address those possible violations of the IV core assumptions and potential risk, many efforts have been made recently. The MR-Egger regression method introduced an intercept term to capture the presence of unbalanced systematic pleiotropy under the Instrument Strength Independent of Direct Effect (InSIDE) assumption (Bowden et al., 2015). However, MR-Egger would be biased when there exists idiosyncratic pleiotropy. Zhu et al. (2018) proposed the GSMR method that removes suspected genetic variants with relatively large direct effects and also takes the LD structure into account by using the generalized least squares approach. However, removal of a large number of relatively large direct effects might lead to efficiency loss. Zhao et al. (2020b) proposed MR-RAPS to improve statistical power for causal inference and limit the influence of relatively large direct effects by using the adjusted profile likelihood and robust loss functions assuming that those SNP IVs are independent. However, this independent IV assumption might not hold in practice because SNPs within proximity tend to be correlated. Cheng et al. (2020) proposed a two-sample MR method named MR-LDP that built a Bayesian probabilistic model accounting for systematic pleiotropy and LD structures among SNP IVs. One drawback of the MR-LDP method is that it cannot handle relatively large direct effects well.

To overcome the limitations of those aforementioned methods, we propose a more robust method named 'Robust Bayesian Mendelian Randomization (RBMR)' accounting for LD, systematic and idiosyncratic pleiotropy simultaneously in a unified framework. Specifically, to account for LD, we first estimate the LD correlation matrix of SNP IVs and then

medRxiv preprint doi: https://doi.org/10.1101/2021.03.02.21252801; this version posted March 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

explicitly include it in the model likelihood. To account for idiosyncratic pleiotropy, we propose to model the direct effects using the more robust multivariate generalized t-distribution (Arellano-Valle and Bolfarine, 1995; Frahm, 2004) which will be shown to have improved performance than using the Gaussian distribution when the idiosyncratic pleiotropy is present. Moreover, this more robust distribution can be represented as a Gaussian scaled mixture to facilitate model parameter estimation using the parameter expanded variational Bayesian expectation maximization algorithm (PX-VBEM) (Yang et al., 2020) which combines the VB-EM (Beal et al., 2003) and the PX-EM (Liu et al., 1998) together. We further calculate the standard error by calibrating the evidence lower bound (ELBO) according to a nice property of the likelihood ratio test (LRT). Both extensive simulation studies in Section 3 and analysis of two real benchmark data sets in Section 4 show that our proposed RBMR method outperforms competitors. We also find that coronary artery disease (CAD) is associated with increased risk of critically ill COVID-19 outcome.

2. Methods

2.1 The Linear Structural Model

Suppose that we have J possibly correlated genetic variants (for example, single-nucleotide polymorphisms, or SNPs) $G_j, j = 1, 2, \ldots, J$, the exposure variable X, the outcome variable Y of interest and unknown confounding factors U. Let δ_X and δ_Y denote the effects of confounders U on exposure X and outcome Y respectively. The coefficients γ_j $(j = 1, 2, \ldots, J)$ denote the SNP-exposure true effects. Suppose that all the IVs are valid, then the exposure can be represented as a linear structural function of the SNPs, confounders and an independent random noise term e_X. The outcome can be represented as a linear structural function of the exposure, confounders and the independent random noise term e_Y. The true effect size of the exposure on the outcome is denoted as β_0. Then, we have the following linear structural equation models (Bowden et al., 2015):

$$X = \sum_{j=1}^{J} G_j \gamma_j + U\delta_X + e_X, \quad Y = \beta_0 X + U\delta_Y + e_Y. \tag{2.1}$$

Let Γ_j $(j = 1, 2, \ldots, J)$ be the true effects of SNPs on the outcome. With valid IVs, we have

$$\Gamma_j = \beta_0 \gamma_j. \tag{2.2}$$

medRxiv preprint doi: https://doi.org/10.1101/2021.03.02.21252801; this version posted March 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

To accommodate possible violations of the exclusion restriction assumption, we now consider the following modified linear structural functions (Bowden et al., 2015):

$$X = \sum_{j=1}^{J} G_j \gamma_j + U\delta_X + e_X, \quad Y = \sum_{j=1}^{J} G_j \alpha_j + \beta_0 X + U\delta_Y e_Y, \tag{2.3}$$

where the coefficients α_j $(j = 1, 2, \ldots, J)$ represent the direct effects of the SNPs on the outcome. Then we have

$$\Gamma_j = \beta_0 \gamma_j + \alpha_j. \tag{2.4}$$

So far, many existing MR methods assign the Gaussian distribution on each direct effect α_j, that is $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \sigma_0^2 \mathbf{I_J})$ (Zhao et al., 2020b; Cheng et al., 2020; Zhao et al., 2020a), where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_J]^{\mathrm{T}}$ is a J-dimensional vector of direct effects. However, real genetic data might contain some relatively large direct effects whose distribution can be heavy-tailed, and thus the Gaussian distribution might not be a good fit. Therefore, we propose to assign the multivariate generalized t-distribution on $\boldsymbol{\alpha}$ (Arellano-Valle and Bolfarine, 1995; Kotz and Nadarajah, 2004), which is a robust alternative to the Gaussian distribution (Frahm, 2004).

2.2 The Robust Bayesian MR Model

Let $\left\{\widehat{\gamma}_j, \widehat{\sigma}^2_{X_j}\right\}_{j=1,\ldots,J}$ and $\left\{\widehat{\Gamma}_j, \widehat{\sigma}^2_{Y_j}\right\}_{j=1,\ldots,J}$ be the GWAS summary statistics for the exposure and the outcome respectively, where $\left\{\widehat{\sigma}^2_{X_j}, \widehat{\sigma}^2_{Y_j}\right\}$ are the corresponding estimated standard errors. Many existing MR methods assume that IVs are independent from each other (Ehret et al., 2011; Bowden et al., 2015; Zhao et al., 2020b), and the uncorrelated SNPs can be chosen by using a tool called LD clumping (Hemani et al., 2016; Purcell et al., 2007), which might remove many SNP IVs and thus cause efficiency loss. To include more SNP IVs even if they are in LD, we need to account for the LD structure explicitly. To achieve this goal, we use a reference panel sample to assist with reconstructing LD matrix, such as the 1000 Genome Project Phase 1 (N=379) (Consortium et al., 2012). We first apply the LDetect method to partition the whole genome into Q blocks (Berisa and Pickrell, 2016) and then estimate the LD matrix $\boldsymbol{\Theta}$ using the estimator $\widehat{\boldsymbol{\Theta}}^{(k)} (k = 1, 2, \ldots, Q)$ first proposed by Rothman (2012). Then, the distributions of $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\Gamma}}$ are given by

$$\widehat{\boldsymbol{\gamma}} | \boldsymbol{\gamma}, \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}} \sim \mathcal{N}\left(\widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}} \widehat{\boldsymbol{\Theta}} \widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}}^{-1} \boldsymbol{\gamma}, \widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}} \widehat{\boldsymbol{\Theta}} \widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}}\right), \tag{2.5}$$

$$\widehat{\boldsymbol{\Gamma}} | \boldsymbol{\Gamma}, \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}} \sim \mathcal{N}\left(\widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}} \widehat{\boldsymbol{\Theta}} \widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}}^{-1} \boldsymbol{\Gamma}, \widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}} \widehat{\boldsymbol{\Theta}} \widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}}\right), \tag{2.6}$$

medRxiv preprint doi: https://doi.org/10.1101/2021.03.02.21252801; this version posted March 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

where $\widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}} = \text{diag}\left([\widehat{\sigma}_{X_1}, \ldots, \widehat{\sigma}_{X_J}]\right)$ and $\widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}} = \text{diag}\left([\widehat{\sigma}_{Y_1}, \ldots, \widehat{\sigma}_{Y_J}]\right)$ are both diagonal matrices (Zhu and Stephens, 2017).

To account for the presence of idiosyncratic pleiotropy, we propose to model the direct effects $\boldsymbol{\alpha}$ using the more robust multivariate generalized t-distribution (Arellano-Valle and Bolfarine, 1995; Kotz and Nadarajah, 2004; Ala-Luhtala and Piché, 2016) whose density function is given by

$$\begin{aligned} t_J(\boldsymbol{\alpha}|\boldsymbol{\Sigma}, \alpha_w, \beta_w) &= \frac{f(\alpha_w + J/2)}{|\boldsymbol{\Sigma}|^{1/2} f(\alpha_w)(2\pi\beta_w)^{J/2}} \left[1 + \frac{1}{2\beta_w}(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha})\right]^{-(\alpha_w + J/2)} \\ &= \int \mathcal{N}(\boldsymbol{\alpha}|\mathbf{0}, \boldsymbol{\Sigma}/w)\mathcal{G}(w|\alpha_w, \beta_w)\mathrm{d}w, \end{aligned} \tag{2.7}$$

where $\mathcal{N}(\boldsymbol{\alpha}|\mathbf{0}, \boldsymbol{\Sigma}/w)$ denotes the J-dimensional Gaussian distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}/w$, $\boldsymbol{\Sigma} = \sigma_0^2\mathbf{I_J}$ is a $J \times J$ diagonal matrix, and $\mathcal{G}(w|\alpha_w, \beta_w)$ is the Gamma distribution of a univariate positive variable w referred to as a weight variable

$$\mathcal{G}(w|\alpha_w, \beta_w) = \frac{{\beta_w}^{\alpha_w}}{f(\alpha_w)} w^{\alpha_w - 1} e^{-\beta_w w}, \tag{2.8}$$

where f denotes the Gamma function. When $\alpha_w = \beta_w = \nu/2$ in equation (2.8), the distribution in equation (2.7) reduces to a multivariate t-distribution, where ν is the degree of freedom. Gaussian scaled mixture representation enables the use of EM-type algorithms for statistical inference, such as the PX-VBEM (Yang et al., 2020) described in Section 2.3.

Then we denote the distribution of the latent variable $\boldsymbol{\gamma}$ as

$$\boldsymbol{\gamma}|\boldsymbol{\sigma}^2 \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\sigma}^2\right), \tag{2.9}$$

where $\boldsymbol{\sigma}^2 = \sigma^2\mathbf{I_J}$ is a $J \times J$ diagonal matrix. By assuming that $\boldsymbol{\gamma}$, $\boldsymbol{\alpha}$ and w are latent variables, the complete data likelihood can be written as

$$\begin{aligned} \Pr\left(\widehat{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{\gamma}}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, w|\widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}}, \widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}}, \widehat{\boldsymbol{\Theta}}; \boldsymbol{\theta}, \boldsymbol{h}\right) =& \mathcal{N}\left(\widehat{\boldsymbol{\Gamma}}|\widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}}\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}}^{-1}\left(\beta_0\boldsymbol{\gamma} + \boldsymbol{\alpha}\right), \widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}}\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}}\right)\mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I_J}\right) \\ &\times\mathcal{N}\left(\widehat{\boldsymbol{\gamma}}|\widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}}\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}}^{-1}\boldsymbol{\gamma}, \widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}}\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}}\right)\mathcal{N}\left(\boldsymbol{\alpha}|\mathbf{0}, \sigma_0^2\mathbf{I_J}/w\right)\mathcal{G}\left(w|\alpha_w, \beta_w\right). \end{aligned} \tag{2.10}$$

2.3 Estimation and Inference

The standard expectation-maximization (EM) algorithm (Dempster et al., 1977) is a popular choice for finding the maximum likelihood estimate in the presence of missing (latent) variables. However, one difficulty for implementing the EM algorithm is to calculate the marginal likelihood function which might involve difficult integration with respect to the

medRxiv preprint doi: https://doi.org/10.1101/2021.03.02.21252801; this version posted March 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

distributions of the latent variables. In addition, the original EM algorithm might be slow (Liu et al., 1998). To address these numerical issues, we utilize a parameter expanded variational Bayesian expectation-maximization algorithm, namely, PX-VBEM (Yang et al., 2020), by replacing the EM algorithm in VB-EM (Beal et al., 2003) with PX-EM algorithm (Liu et al., 1998) to accelerate the speed of convergence. To start with, for the purpose of applying the PX-EM algorithm, the distribution of $\widehat{\boldsymbol{\gamma}}$ in equation (2.5) can be rewritten as follows:

$$\widehat{\boldsymbol{\gamma}}|\boldsymbol{\gamma}, \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}} \sim \mathcal{N}\left(\zeta \widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}} \widehat{\boldsymbol{\Theta}} \widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}}^{-1} \boldsymbol{\gamma}, \widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}} \widehat{\boldsymbol{\Theta}} \widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}}\right). \tag{2.11}$$

We also rewrite the complete data likelihood in equation (2.10) as:

$$\begin{aligned}\Pr\left(\widehat{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{\gamma}}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, w | \widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}}, \widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}}, \widehat{\boldsymbol{\Theta}}; \boldsymbol{\theta}, \boldsymbol{h}\right) =& \mathcal{N}\left(\widehat{\boldsymbol{\Gamma}} | \widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}} \widehat{\boldsymbol{\Theta}} \widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}}^{-1} \left(\beta_0 \boldsymbol{\gamma} + \boldsymbol{\alpha}\right), \widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}} \widehat{\boldsymbol{\Theta}} \widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}}\right) \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}_{\mathbf{J}}\right) \\ &\times \mathcal{N}\left(\widehat{\boldsymbol{\gamma}} | \zeta \widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}} \widehat{\boldsymbol{\Theta}} \widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}}^{-1} \boldsymbol{\gamma}, \widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}} \widehat{\boldsymbol{\Theta}} \widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}}\right) \mathcal{N}\left(\mathbf{0}, \sigma_0^2 \mathbf{I}_{\mathbf{J}} / w\right) \mathcal{G}\left(w | \alpha_w, \beta_w\right),\end{aligned} \tag{2.12}$$

where the expanded model parameters for RBMR are $\boldsymbol{\theta} \stackrel{\text{def}}{=} \{\beta_0, \sigma_0^2, \sigma^2, \zeta\}$. Let $q(\boldsymbol{\gamma}, \boldsymbol{\alpha}, w)$ be a variational posterior distribution. The logarithm of the marginal likelihood can be decomposed into two parts,

$$\begin{aligned}&\log \Pr\left(\widehat{\boldsymbol{\gamma}}, \widehat{\Gamma} | \widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}}, \widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}}, \widehat{\boldsymbol{\Theta}}; \boldsymbol{\theta}, \boldsymbol{h}\right) \\ &= \mathbb{E}_{q(\boldsymbol{\gamma}, \boldsymbol{\alpha}, w)}\left[\log \Pr\left(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\Gamma}} | \widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}}, \widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}}, \widehat{\boldsymbol{\Theta}}; \boldsymbol{\theta}, \boldsymbol{h}\right)\right] \\ &= \mathcal{L}(q) + \mathbb{KL}(q \| p),\end{aligned} \tag{2.13}$$

where

$$\begin{aligned}\mathcal{L}(q) &= \mathbb{E}_{q(\boldsymbol{\gamma}, \boldsymbol{\alpha}, w)}\left[\log \frac{\Pr\left(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\Gamma}}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, w | \widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}}, \widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}}, \widehat{\boldsymbol{\Theta}}; \boldsymbol{\theta}, \boldsymbol{h}\right)}{q\left(\boldsymbol{\gamma}, \boldsymbol{\alpha}, w\right)}\right], \\ \mathbb{KL}(q \| p) &= \mathbb{E}_{q(\boldsymbol{\gamma}, \boldsymbol{\alpha}, w)}\left[\log \frac{q\left(\boldsymbol{\gamma}, \boldsymbol{\alpha}, w\right)}{p\left(\boldsymbol{\gamma}, \boldsymbol{\alpha}, w | \widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}}, \widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}}, \widehat{\boldsymbol{\Theta}}; \boldsymbol{\theta}, \boldsymbol{h}\right)}\right].\end{aligned} \tag{2.14}$$

Given that the $\mathcal{L}(q)$ is an evidence lower bound (ELBO) of the marginal log-likelihood, the non-negative Kullback-Leibler (KL) divergence $\mathbb{KL}(q \| p)$ is equal to zero if and only if the variational posterior distribution is equal to the true posterior distribution. Minimizing the KL divergence is equivalent to maximizing ELBO. Before calculating the maximization of ELBO, due to the fact that latent variables are independent of each other, the decomposition

medRxiv preprint doi: https://doi.org/10.1101/2021.03.02.21252801; this version posted March 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

form of the posterior distribution $q(\boldsymbol{\gamma}, \boldsymbol{\alpha}, w)$ is obtained using the mean field assumption (Blei et al., 2017),

$$q(\boldsymbol{\gamma}, \boldsymbol{\alpha}, w) = \prod_{j=1}^{J} q\left(\gamma_j\right) \prod_{j=1}^{J} q\left(\alpha_j\right) q(w). \tag{2.15}$$

In the PX-VB-E step, the optimal variational posterior distributions for $\boldsymbol{\gamma}$, $\boldsymbol{\alpha}$ and w can be written as:

$$\begin{gathered} q\left(\boldsymbol{\gamma}|\mu_{\gamma_j}, \sigma^2_{\gamma_j}\right) = \prod_{j=1}^{J} \mathcal{N}\left(\mu_{\gamma_j}, \sigma^2_{\gamma_j}\right),\ q\left(\boldsymbol{\alpha}|\mu_{\alpha_j}, \sigma^2_{\alpha_j}\right) = \prod_{j=1}^{J} \mathcal{N}\left(\mu_{\alpha_j}, \sigma^2_{\alpha_j}\right), \\ q\left(w|\widetilde{\alpha}_w, \widetilde{\beta}_w\right) = \mathcal{G}\left(\widetilde{\alpha}_w, \widetilde{\beta}_w\right). \end{gathered} \tag{2.16}$$

The updating equations for the parameters are given by

$$\begin{aligned} -\frac{1}{2\sigma^2_{\gamma_j}} &= -\frac{\beta_0^2}{2}\frac{\widehat{\boldsymbol{\Theta}}_{jj}}{\sigma^2_{Y_j}} - \frac{\zeta^2\widehat{\boldsymbol{\Theta}}_{jj}}{2\sigma^2_{X_j}} - \frac{1}{2\sigma^2}, \\ \frac{\mu_{\gamma_j}}{\sigma^2_{\gamma_j}} &= \beta_0\frac{\widehat{\Gamma}_j}{\sigma^2_{Y_j}} - \frac{\beta_0^2}{\sigma_{Y_j}}\left(\sum_{j'\neq j}\frac{\left[\gamma_{j'}\right]\widehat{\boldsymbol{\Theta}}_{jj'}}{\sigma_{Y_j}}\right) - \frac{\beta_0}{\sigma_{Y_j}}\left(\sum_{j'=1}^{J}\frac{\left[\alpha_{j'}\right]\widehat{\boldsymbol{\Theta}}_{jj'}}{\sigma_{Y'_j}}\right) + \frac{\zeta\widehat{\gamma}_j}{\sigma^2_{X_j}} - \frac{\zeta^2}{\sigma_{X_j}}\left(\sum_{j'\neq j}\frac{\left[\gamma_{j'}\right]\widehat{\boldsymbol{\Theta}}_{jj'}}{\sigma_{X'_j}}\right), \\ -\frac{1}{2\sigma^2_{\alpha_j}} &= -\frac{1}{2}\frac{\widehat{\boldsymbol{\Theta}}_{jj}}{\sigma^2_{Y_j}} - \frac{[w]}{2\sigma_0^2}, \\ \frac{\mu_{\alpha_j}}{\sigma^2_{\alpha_j}} &= \frac{\widehat{\Gamma}_j}{\sigma^2_{Y_j}} - \frac{\beta_0}{\sigma_{Y_j}}\sum_{j'=1}^{J}\frac{\widehat{\boldsymbol{\Theta}}_{jj'}\left[\gamma_{j'}\right]}{\sigma_{Y'_j}} - \frac{1}{\sigma_{Y_j}}\sum_{j'\neq j}\frac{\left[\alpha_{j'}\right]\widehat{\boldsymbol{\Theta}}_{jj'}}{\sigma_{Y'_j}}, \\ \widetilde{\alpha}_w &= \alpha_w + \frac{J}{2}, \\ \widetilde{\beta}_w &= \beta_w + \sum_{j=1}^{J}\frac{\left[\alpha_j^2\right]}{\sigma_0^2}. \end{aligned} \tag{2.17}$$

where $\left[\gamma_{j'}\right] \stackrel{\text{def}}{=} E_q\left(\gamma_{j'}\right)$, $\left[\alpha_{j'(j)}\right] \stackrel{\text{def}}{=} E_q\left(\alpha_{j'(j)}\right)$ and $[w] \stackrel{\text{def}}{=} E_q\left(w\right)$.

In the PX-VB-M step, by setting the derivate of the ELBO to be zero, the model parameters $\boldsymbol{\theta}$ can be obtained as:

$$\begin{aligned} \beta_0 &= \left\{\boldsymbol{\mu}_{\boldsymbol{\gamma}}^{\mathrm{T}}\widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}}^{-1}\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\gamma}} + \mathrm{Tr}\left(\widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}}^{-1}\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}}^{-1}\mathbf{S}_{\boldsymbol{\gamma}}\right)\right\}^{-1}\left(\widehat{\boldsymbol{\Gamma}}^{\mathrm{T}}\widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}}^{-2}\boldsymbol{\mu}_{\boldsymbol{\gamma}} - \boldsymbol{\mu}_{\boldsymbol{\alpha}}^{\mathrm{T}}\widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}}^{-1}\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\gamma}}\right), \\ \sigma^2 &= \left\{\boldsymbol{\mu}_{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{\mu}_{\boldsymbol{\gamma}} + \mathrm{Tr}\left(\mathbf{S}_{\boldsymbol{\gamma}}\right)\right\}/J, \\ \sigma_0^2 &= \left\{\widetilde{\alpha}_w\left(\boldsymbol{\mu}_{\alpha}^{\mathrm{T}}\boldsymbol{\mu}_{\alpha} + \mathrm{Tr}\left(\mathbf{S}_{\boldsymbol{\alpha}}\right)\right)\right\}/J\widetilde{\beta}_w, \\ \zeta &= \left\{\mu_{\boldsymbol{\gamma}}^{\mathrm{T}}\widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}}^{-1}\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}}^{-1}\mu_{\boldsymbol{\gamma}} + \mathrm{Tr}\left(\widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}}^{-1}\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}}^{-1}\mathbf{S}_{\boldsymbol{\gamma}}\right)\right\}^{-1}\left(\widehat{\boldsymbol{\gamma}}^{\mathbf{T}}\widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}}^{-2}\boldsymbol{\mu}_{\boldsymbol{\gamma}}\right), \end{aligned} \tag{2.18}$$

medRxiv preprint doi: https://doi.org/10.1101/2021.03.02.21252801; this version posted March 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

where $\boldsymbol{\mu}_{\boldsymbol{\gamma}} = (\mu_{\gamma_1}, \ldots, \mu_{\gamma_J})^{\mathrm{T}}$, $\boldsymbol{\mu}_{\boldsymbol{\alpha}} = (\mu_{\alpha_1}, \ldots, \mu_{\alpha_J})^{\mathrm{T}}$, $\mathbf{S}_{\boldsymbol{\gamma}} = \mathrm{diag}\left(\left[\sigma^2_{\gamma_1}, \ldots, \sigma^2_{\gamma_J}\right]\right)$ and $\mathbf{S}_{\boldsymbol{\alpha}} = \mathrm{diag}\left(\left[\sigma^2_{\alpha_1}, \ldots, \sigma^2_{\alpha_J}\right]\right)$. Finally, we use the updated model parameters $\boldsymbol{\theta}$ to construct the evidence lower bound to check the convergence. Since we adopt PX-EM algorithm, the reduction step should be used to process the obtained parameters. More technical details can be found in the Supplementary Materials.

After obtaining an estimate of the causal effect, we further calculate the standard error according to the property of likelihood ratio test (LRT) statistics which asymptotically follows the χ^2_1 under the null hypothesis (Van der Vaart, 2000). We first formulate the statistical tests to examine the association between the risk factor and the outcome.

$$\mathcal{H}_0 : \beta_0 = 0 \qquad \mathcal{H}_a : \beta_0 \neq 0, \tag{2.19}$$

the likelihood ratio test (LRT) statistics for the causal effect is given by:

$$\boldsymbol{\Lambda} = 2\left(\log \Pr\left(\widehat{\gamma}, \widehat{\Gamma} | \widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}}, \widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}}, \widehat{\boldsymbol{\Theta}}; \boldsymbol{h}, \hat{\boldsymbol{\theta}}^{ML}\right) - \log \Pr\left(\widehat{\gamma}, \widehat{\Gamma} | \widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}}, \widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}}, \widehat{\boldsymbol{\Theta}}; \boldsymbol{h}, \hat{\boldsymbol{\theta}}_0^{ML}\right)\right), \tag{2.20}$$

where $\hat{\boldsymbol{\theta}}_0^{ML}$ and $\hat{\boldsymbol{\theta}}^{ML}$ are collections of parameter estimates obtained by maximizing the marginal likelihood under the null hypothesis $\mathcal{H}_0$ and under the alternative hypothesis $\mathcal{H}_a$. We utilize PX-VBEM algorithm to maximize the ELBO to get the $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}_0$ instead of maximizing the marginal likelihood to overcome the computational intractability. Although PX-VBEM produces accurate posterior mean estimates (Blei et al., 2017; Dai et al., 2017; Yang et al., 2018), it would underestimate the marginal variance because we use the estimated posterior distribution from the ELBO to approximate the marginal likelihood in equation (2.20) (Wang and Titterington, 2005). Thus, we calibrate ELBO by plugging our estimates ($\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}_0$) from PX-VBEM into the equation (2.20) to construct the test statistics (Yang et al., 2020):

$$\boldsymbol{\Lambda}' = 2\left(\log \Pr\left(\widehat{\gamma}, \widehat{\Gamma} | \widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}}, \widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}}, \widehat{\boldsymbol{\Theta}}; \boldsymbol{h}, \hat{\boldsymbol{\theta}}\right) - \log \Pr\left(\widehat{\gamma}, \widehat{\Gamma} | \widehat{\boldsymbol{\sigma}}_{\boldsymbol{X}}, \widehat{\boldsymbol{\sigma}}_{\boldsymbol{Y}}, \widehat{\boldsymbol{\Theta}}; \boldsymbol{h}, \hat{\boldsymbol{\theta}}_0\right)\right). \tag{2.21}$$

Then, we can get the well-calibrated standard error as $\widehat{se}(\widehat{\beta}_0) = \widehat{\beta}_0 / \sqrt{\boldsymbol{\Lambda}'}$.

3. Simulation Studies

Although our proposed method is based on summary level data, we still simulate the individual-level data to better mimic real genetic data sets. Specifically, the data sets are generated according to the following models:

$$\mathbf{X} = \mathbf{G}_{\boldsymbol{X}}\boldsymbol{\gamma} + \mathbf{U}_{\boldsymbol{X}}\boldsymbol{\eta}_X + \boldsymbol{\varepsilon}_{\boldsymbol{X}}, \quad \mathbf{Y} = \beta_0\mathbf{X} + \mathbf{G}_{\boldsymbol{Y}}\boldsymbol{\alpha} + \mathbf{U}_{\boldsymbol{Y}}\boldsymbol{\eta}_Y + \boldsymbol{\varepsilon}_{\boldsymbol{Y}}, \tag{3.1}$$

medRxiv preprint doi: https://doi.org/10.1101/2021.03.02.21252801; this version posted March 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

where $\mathbf{X} \in \mathbb{R}^{n_X \times 1}$ is the exposure vector, $\mathbf{Y} \in \mathbb{R}^{n_Y \times 1}$ is the outcome vector, $\mathbf{G_X} \in \mathbb{R}^{n_X \times J}$ and $\mathbf{G_Y} \in \mathbb{R}^{n_Y \times J}$ are the genotype datasets for the exposure $\mathbf{X}$ and the outcome $\mathbf{Y}$, $\mathbf{U_X} \in \mathbb{R}^{n_X \times N_0}$ and $\mathbf{U_Y} \in \mathbb{R}^{n_Y \times N_0}$ are matrices for confounding variables, n_X and n_Y are the corresponding sample sizes of exposure $\mathbf{X}$ and outcome $\mathbf{Y}$, J is the number of genotyped SNPs. The error terms $\boldsymbol{\varepsilon_X}$ and $\boldsymbol{\varepsilon_Y}$ are independent noises from $\mathcal{N}\left(\mathbf{0}, \boldsymbol{\sigma^2_{\varepsilon_X}}\mathbf{I_{n_X}}\right)$ and $\mathcal{N}\left(\mathbf{0}, \boldsymbol{\sigma^2_{\varepsilon_Y}}\mathbf{I_{n_Y}}\right)$ respectively. In model (3.1), β_0 is the true causal effect and $\boldsymbol{\alpha}$ represents the direct effect of the SNPs on the outcome not mediated by the exposure variable.

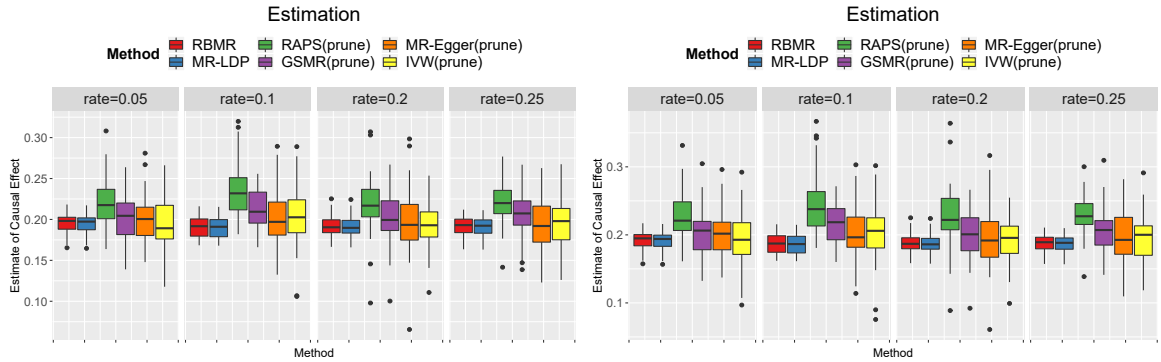
An external reference panel $\mathbf{G}_r \in \mathbb{R}^{n_r \times J}$ is chosen for estimating the LD matrix among SNPs, where n_r is the sample size of the chosen reference panel. The R package named *MR.LDP* is available to generate genotyped matrices $\mathbf{G_X}$, $\mathbf{G_Y}$ and $\mathbf{G_r}$. We fix $n_X = n_Y = 20000$ and $n_r = 2500$. The total number of SNPs is $J = 300$. For confounding variables, each column of $\mathbf{U_X}$ and $\mathbf{U_Y}$ is sampled from a standard normal distribution while each row of corresponding coefficients $\boldsymbol{\eta_X} \in \mathbb{R}^{N_0 \times 1}$ and $\boldsymbol{\eta_Y} \in \mathbb{R}^{N_0 \times 1}$ of confounding variables is obtained from a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{S_\eta})$ where the diagonal elements of $\boldsymbol{S_\eta} \in \mathbb{R}^{2\times 2}$ are 1 and the off-diagonal elements are 0.8. We simulate the following three cases of idiosyncratic pleiotropy:

1. **case 1**: $\Gamma_j = \gamma_j\beta_0 + \alpha_j, \alpha_j \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_0^2),\ j = 1, 2, \ldots, 300$. We randomly select $\{5\%, 10\%,$
   $20\%, 25\%\}$ of IVs so that their direct effects α_js have mean 0 and variance $10{\sigma_0}^2$.
2. **case 2**: $\Gamma_j = \gamma_j\beta_0 + \alpha_j, \alpha_j \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_0^2),\ j = 1, 2, \ldots, 300$. We randomly select $\{5\%, 10\%,$
   $20\%, 25\%\}$ of IVs so that their direct effect α_js have mean $10\sigma_0$ and variance σ_0^2.
3. **case 3**: $\Gamma_j = \gamma_j\beta_0 + \alpha_j,\ \alpha_j \overset{i.i.d}{\sim} t(df),\ j = 1, 2, \ldots, 300$, the values of freedom df are $\{10, 15, 20\}$.

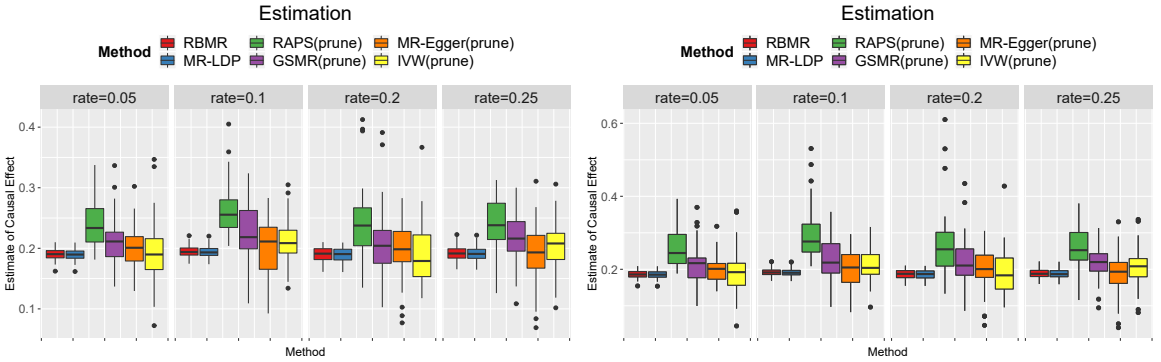
The σ_0^2 in case 1 and case 2 is controlled by the heritability $h_{\boldsymbol{\alpha}}$ due to systematic pleiotropy, $h^2_{\boldsymbol{\alpha}} = \frac{\mathrm{var}(\mathbf{G_Y}\boldsymbol{\alpha})}{\mathrm{var}(\mathbf{Y})}$ which is set to be $h^2_{\boldsymbol{\alpha}}$: 0.05 and 0.07 respectively. The signal magnitude for $\boldsymbol{\gamma}$ is chosen such that the heritability $h^2_{\boldsymbol{\gamma}} = \frac{\mathrm{var}(\beta_0\mathbf{G_X}\boldsymbol{\gamma})}{\mathrm{var}(\mathbf{Y})} = 0.1$. The true causal effect β_0 is set to be 0.2.

We first run single-variant genetic association analysis for the exposure and the outcome respectively, and then we obtain the summary-level statistics $\{\widehat{\gamma}_j, \widehat{\Gamma}_j\}_{j=1,2,\ldots,300}$ with their corresponding standard errors $\{\widehat{\sigma}_{X_j}, \widehat{\sigma}_{Y_j}\}_{j=1,2,\ldots,300}$ for all the three cases. Then we use the

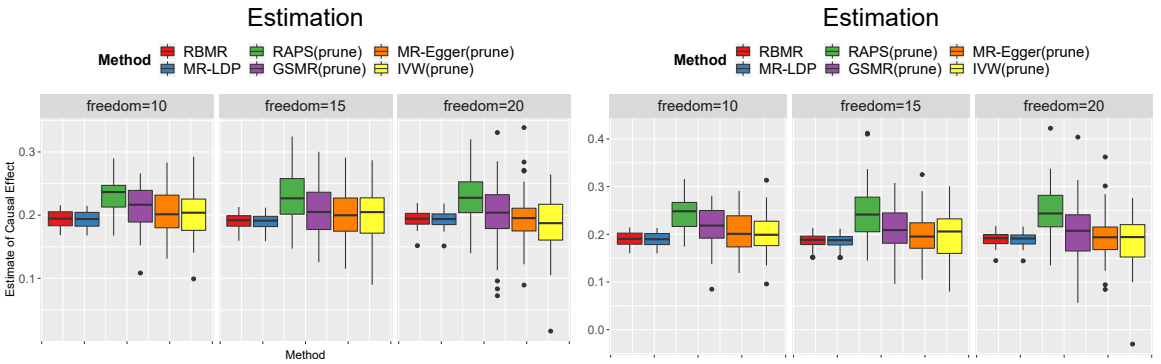
medRxiv preprint doi: https://doi.org/10.1101/2021.03.02.21252801; this version posted March 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.



(a) $h^2_{\boldsymbol{\alpha}}$=0.05 (b) $h^2_{\boldsymbol{\alpha}}$=0.07



(c) $h^2_{\boldsymbol{\alpha}}$=0.05 (d) $h^2_{\boldsymbol{\alpha}}$=0.07



(e) $h^2_{\boldsymbol{\alpha}}$=0.05 (f) $h^2_{\boldsymbol{\alpha}}$=0.07

Figure 3.1: *Comparisons of MR methods at heritability level of 0.05 and 0.07: The Figure (a), (c) and (e) represent comparisons of the causal estimates $\widehat{\beta}_0$ of RBMR, MR-LDP, RAPS, GSMR, MR-Egger and IVW methods at heritability level of 0.05 for three cases of idiosyncratic pleiotropy, respectively. The Figure (b), (d) and (f) represent comparisons of the causal estimates $\widehat{\beta}_0$ of RBMR, MR-LDP, RAPS, GSMR, MR-Egger and IVW methods at heritability level of 0.07 for three cases of idiosyncratic pleiotropy, respectively.*

medRxiv preprint doi: https://doi.org/10.1101/2021.03.02.21252801; this version posted March 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

summary-level data to conduct MR analyses using the proposed RBMR, MR-LDP, MR-Egger, RAPS, GSMR and IVW methods. As the prerequisite for MR-Egger, RAPS and IVW methods is that the instrumental variables are independent of each other, we adopt a step-wise GSMR method to remove SNPs with LD structure. We repeat such experiment for 100 times.

The simulation results are shown in Figure 3.1. In all the three cases considered, we find that the proposed RBMR and MR-LDP methods are more stable than the other four methods: RAPS, GSMR MR-Egger and IVW. However, we found that our proposed RBMR method has smaller bias and root mean square error (RMSE) than the MR-LDP method. More detailed results are provided in the Supplementary Materials.

4. Real Data Analysis

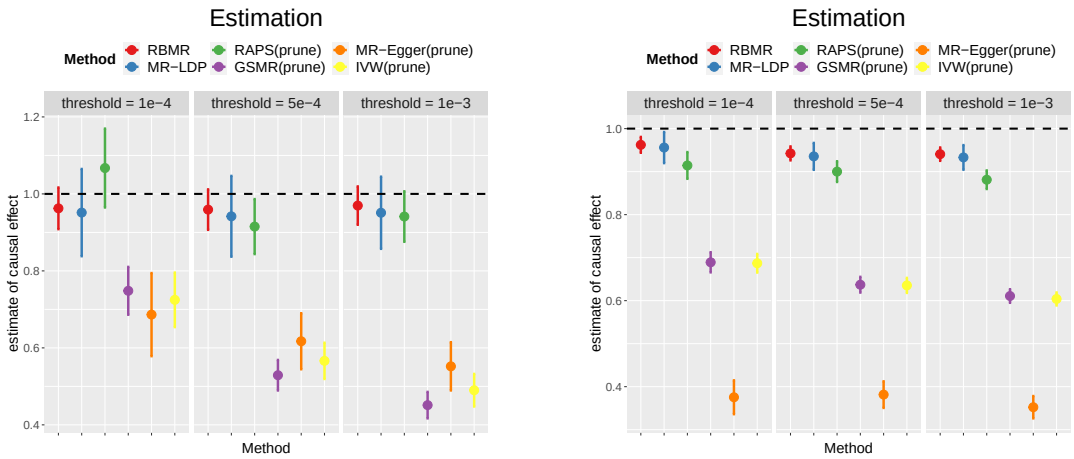
In this section, we analyzed three real data sets to demonstrate the performance of our proposed method. The 1000 Genome Project Phase 1 (1KGP) is used as the reference panel to compute the LD matrix (Consortium et al., 2012). We first analyze two benchmark data sets commonly used for method comparison purpose, then we will estimate the causal effect of coronary artery disease (CAD) on the risk of critically ill COVID-19 outcome defined as those who end up on respiratory support or die from COVID-19.

The first benchmark data analysis is based on the summary-level data sets from two non-overlapping GWAS studies for the coronary artery disease (CAD), usually referred to as the CAD-CAD data. The true causal effect should be exactly one. The selection data set is from the Myocardial Infarction Genetics in the UK Biobank, the exposure data is from the Coronary Artery Disease (C4D) Genetics Consortium (Consortium et al., 2011), and the outcome data is from the transatlantic Coronary Artery Disease Genome Wide Replication and Meta-analysis (CARDIoGRAM) (Schunkert et al., 2011). We first filter the genetic variants using the selection data under different association p-value thresholds (p-value $\leq 1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}$). Then we applied our proposed RBMR method and the MR-LDP to all the selected and possibly correlated SNPs by accounting for the LD structure explicitly. We applied the GSMR, IVW, MR-Egger and MR-RAPS methods using the independent SNPs after LD pruning because those methods require independent SNPs. We obtain causal effect point estimates and the corresponding 95% confidence intervals (CI) as shown in Figure 4.1(a). We found that our proposed RBMR method outperforms other methods because it has the smallest bias and shortest confidence intervals for a range of

medRxiv preprint doi: https://doi.org/10.1101/2021.03.02.21252801; this version posted March 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

p-value thresholds.

To further investigate the performance of our proposed RBMR method, we consider the case that both the exposure and outcome are body mass index (BMI). We select SNPs based on previous research (Locke et al., 2015). The exposure is the BMI for physically active men and the outcome is the BMI for physically active women, both are of European ancestry (https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files#2018_GIANT_and_UK_BioBank_Meta_Analysis_for_Public_Release). The point estimates and the corresponding 95% confidence intervals are shown in Figure 4.1(b). We found that our proposed RBMR method has smaller bias than other competing methods. More numerical results are provided in the Supplementary Materials.

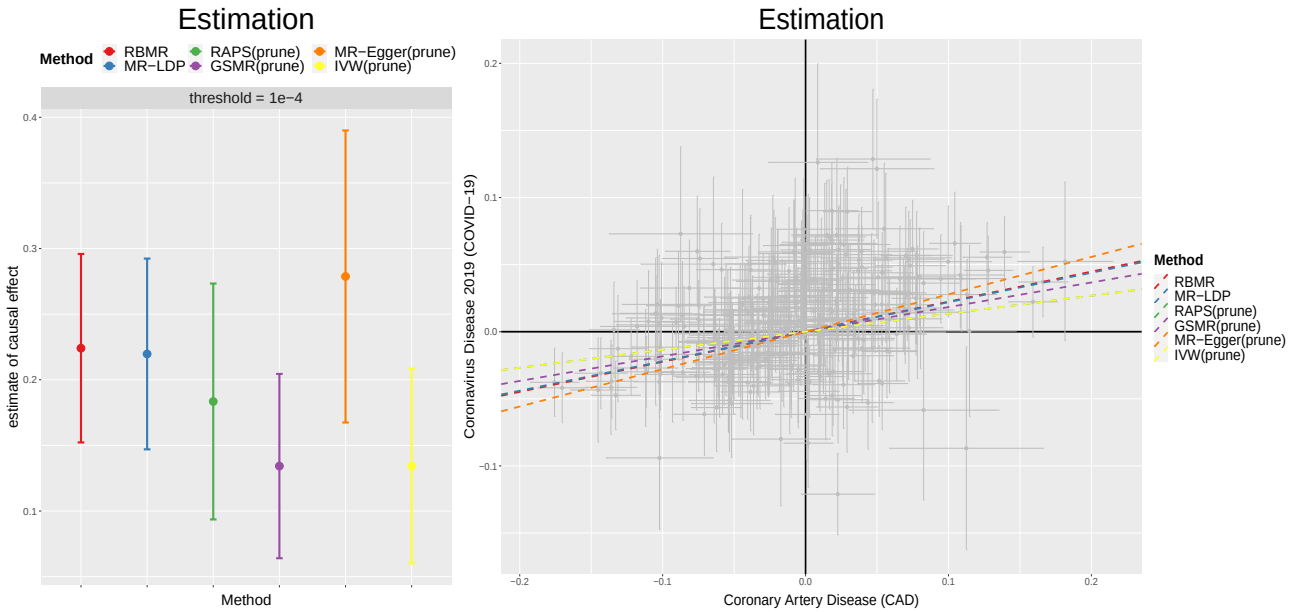


(a) *CAD-CAD* (b) *BMI-BMI*

Figure 4.1: *The results of CAD-CAD and BMI-BMI using 1KGP as the reference panel with shrinkage parameter* $\lambda = 0.15$ *and screening the corresponding SNPs under three thresholds* (p*-value* $\leq 1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}$)*, , RBMR, MR-LDP, RAPS, GSMR, MR-Egger and IVW methods use SNPs selected to calculate the casual effect estimate* $\widehat{\beta}_0$*.*

We apply our proposed RBMR method together with other competing methods to estimate the causal effect of CAD on the risk of critically ill coronavirus disease 2019 (COVID-19) defined as those who end up on respiratory support or die from COVID-19. Specifically, the selection data set is the Myocardial Infraction Genetics in the UK Biobank and the exposure data set is from Consortium et al. (2011). The outcome is obtained from Freeze 5 (January 2021) of the COVID-19 Host Genetics Initiative (COVID-19 HGI) Genome-Wide Association Study (Initiative et al., 2020) (https://www.covid19hg.org/results/). The

medRxiv preprint doi: https://doi.org/10.1101/2021.03.02.21252801; this version posted March 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.



(a) *CAD-COVID-19* (b) *Scatter plot of* $\widehat{\Gamma}_j$ *versus* $\widehat{\gamma}_j$ *for CAD-COVID-19 data*

Figure 4.2: *The results of CAD-COVID-19 using 1KGP as the reference panel with shrinkage parameter* $\lambda = 0.15$ *and screening the corresponding SNPs under thresholds (p-value* $\leq 1 \times 10^{-4}$*), RBMR, MR-LDP, RAPS, GSMR, MR-Egger and IVW methods use SNPs selected to calculate the casual effect estimate* $\widehat{\beta}_0$*. Each point of scatter plot in Figure (b) is augmented by the standard error of* $\widehat{\Gamma}_j$ *and* $\widehat{\gamma}_j$ *on the vertical and horizontal sides. Dashed lines are the slopes fitted by six methods.*

data combines the genetic data of 49562 patients and two million controls from 46 studies across 19 countries (Initiative et al., 2021). We mainly consider the GWAS data on the 6179 cases with critical illness due to COVID-19 and 1483780 controls from the general populations in our analysis. We use the selection data with p-value $\leq 1 \times 10^{-4}$ threshold to select genetic variants as IVs. As shown in Figure 4.2(a), we found a significant effect of CAD on the risk of critically ill COVID-19 using our RBMR method ($\widehat{\beta} = 0.2242$, p-value $= 0.0273$, 95% CI = [0.0834, 0.3648]), MR-LDP ($\widehat{\beta} = 0.2197$, p-value $= 0.0325$, 95% CI = [0.0773, 0.3621]), RAPS ($\widehat{\beta} = 0.1835$, p-value $= 0.0412$, 95% CI = [0.0074, 0.3596]) and MR-Egger ($\widehat{\beta} = 0.2787$, p-value $= 0.012$, 95% CI = [0.0607, 0.4967]). However, the results of GSMR ($\widehat{\beta} = 0.1343$, p-value $= 0.0558$, 95% CI = [-0.0034, 0.2719]) and IVW ($\widehat{\beta} = 0.1343$, p-value $= 0.0700$, 95% CI = [-0.0111, 0.2796]) are not significant (p-value > 0.05). Although MR-LDP and RBMR give similar point estimate, however, our RBMR is more accurate as its confidence interval is slightly shorter and its p-value is more significant.

medRxiv preprint doi: https://doi.org/10.1101/2021.03.02.21252801; this version posted March 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

5. Discussion

In this paper, we propose a novel two-sample robust MR method RBMR by accounting for the LD structure, systematic pleiotropy and idiosyncratic pleiotropy simultaneously in a unified framework. Specifically, we propose to use the more robust multivariate generalized t-distribution rather the less robust Gaussian distribution to model the direct effects of the IV on the outcome not mediated by the exposure. Moreover, the multivariate generalized t-distribution can be reformulated as Gaussian scaled mixtures to facilitate the estimation of the model parameters using the parameter expanded variational Bayesian expectation-maximum algorithm (PX-VBEM). Through extensive simulations and analysis of two real benchmark data sets, we found that our method outperforms the other competing methods. We also found that CAD is associated with increased risk of critically ill COVID-19 outcome using our RBMR method.

We make the following two major contributions. First, our method can account for the LD structure explicitly and thus can include more possibly correlated SNPs to reduce bias and increase estimation efficiency. Second, our RBMR method is more robust to the presence of idiosyncratic pleiotropy. One limitation of our proposed method is that it cannot handle correlated pleiotropy where the direct effect of the IV on the outcome might be correlated with the IV strength. We leave it as our future work.

Acknowledgements

We would like to thank the GIANT Consortium and COVID-19 host genetics initiative for providing publicly available summary statistics to support our analysis.

References

Ala-Luhtala, J. and Piché, R. (2016). Gaussian scale mixture models for robust linear multivariate regression with missing data. *Communications in Statistics-Simulation and Computation*, 45(3).

Arellano-Valle, R. B. and Bolfarine, H. (1995). On some characterizations of the t-distribution. *Statistics & Probability Letters*, 25(1):79–85.

Beal, M. J. et al. (2003). *Variational algorithms for approximate Bayesian inference*. University of London London.

medRxiv preprint doi: https://doi.org/10.1101/2021.03.02.21252801; this version posted March 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

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

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(528):859–877.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430):443–450.

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.

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.

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.

Consortium, . G. P. et al. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56.

Consortium, C. A. D. C. G. et al. (2011). A genome-wide association study in europeans and south asians identifies five new loci for coronary artery disease. *Nature Genetics*, 43(4):339.

Dai, M., Ming, J., Cai, M., Liu, J., Yang, C., Wan, X., and Xu, Z. (2017). IGESS: a statistical approach to integrating individual-level genotype data and summary statistics in genome-wide association studies. *Bioinformatics*, 33(18):2882–2889.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

medRxiv preprint doi: https://doi.org/10.1101/2021.03.02.21252801; this version posted March 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Ebrahim, S. and Smith, G. D. (2008). Mendelian randomization: can genetic epidemiology help redress the failures of observational epidemiology? *Human Genetics*, 123(1):15–33.

Ehret, G. B., Munroe, P. B., Rice, K. M., Bochud, M., Johnson, A. D., Chasman, D. I., Smith, A. V., Tobin, M. D., Verwoert, G. C., Hwang, S.-J., et al. (2011). Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, 478(7367):103.

Evans, D. M., Brion, M. J. A., Paternoster, L., Kemp, J. P., McMahon, G., Munafò, M., Whitfield, J. B., Medland, S. E., Montgomery, G. W., Timpson, N. J., et al. (2013). Mining the human phenome using allelic scores that index biological intermediates. *PLoS Genet*, 9(10):e1003919.

Evans, D. M. and Davey Smith, G. (2015). Mendelian randomization: new applications in the coming age of hypothesis-free causality. *Annual Review of Genomics and Human Genetics*, 16:327–350.

Frahm, G. (2004). *Generalized elliptical distributions: theory and applications*. PhD thesis, Universitätsbibliothek.

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

Hemani, G., Zheng, J., Wade, K. H., Laurin, C., Elsworth, B., Burgess, S., Bowden, J., Langdon, R., Tan, V., Yarmolinsky, J., et al. (2016). MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations. *BioRxiv*, page 078972.

Initiative, C.-. H. G. et al. (2020). The covid-19 host genetics initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the sars-cov-2 virus pandemic. *European Journal of Human Genetics*, 28(6):715.

Initiative, C.-. H. G. et al. (2021). Mapping the human genetic architecture of covid-19 by worldwide meta-analysis. *MedRxiv*.

Kotz, S. and Nadarajah, S. (2004). *Multivariate t-distributions and their applications*. Cambridge University Press.

medRxiv preprint doi: https://doi.org/10.1101/2021.03.02.21252801; this version posted March 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

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.

Liu, C., Rubin, D. B., and Wu, Y. N. (1998). Parameter expansion to accelerate em: the px-em algorithm. *Biometrika*, 85(4):755–770.

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.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (gwas catalog). *Nucleic Acids Research*, 45(D1):D896–D901.

Martens, E. P., Pestman, W. R., de Boer, A., Belitser, S. V., and Klungel, O. H. (2006). Instrumental variables: application and limitations. *Epidemiology*, pages 260–267.

Pickrell, J. K., Berisa, T., Liu, J. Z., Ségurel, L., Tung, J. Y., and Hinds, D. A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. *Nature Genetics*, 48(7):709.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575.

Rothman, A. J. (2012). Positive definite estimators of large covariance matrices. *Biometrika*, 99(3):733–740.

Schunkert, H., König, I. R., Kathiresan, S., Reilly, M. P., Assimes, T. L., Holm, H., Preuss, M., Stewart, A. F., Barbalic, M., Gieger, C., et al. (2011). Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genetics*, 43(4):333–338.

Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M., and Smoller, J. W. (2013). Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*, 14(7):483–495.

medRxiv preprint doi: https://doi.org/10.1101/2021.03.02.21252801; this version posted March 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*, volume 3. Cambridge University Press.

Verbanck, M., Chen, C.-y., Neale, B., and Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from mendelian randomization between complex traits and diseases. *Nature Genetics*, 50(5):693–698.

Wang, B. and Titterington, D. (2005). Inadequacy of interval estimates corresponding to variational bayesian approximations. In *AISTATS*. Citeseer.

Yang, Y., Dai, M., Huang, J., Lin, X., Yang, C., Chen, M., and Liu, J. (2018). LPG: A four-group probabilistic approach to leveraging pleiotropy in genome-wide association studies. *BMC Genomics*, 19(1):503.

Yang, Y., Shi, X., Jiao, Y., Huang, J., Chen, M., Zhou, X., Sun, L., Lin, X., Yang, C., and Liu, J. (2020). CoMM-S2: a collaborative mixed model using summary statistics in transcriptome-wide association studies. *Bioinformatics*, 36(7):2009–2016.

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

Zhao, Q., Wang, J., Hemani, G., Bowden, J., and Small, D. S. (2020b). Statistical inference in two-sample summary-data mendelian randomization using robust adjusted profile score. *Annals of Statistics*, 48(3):1742–1769.

Zhu, X. and Stephens, M. (2017). Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *The Annals of Applied Statistics*, 11(3):1561.

Zhu, Z., Zheng, Z., Zhang, F., Wu, Y., Trzaskowski, M., Maier, R., Robinson, M. R., McGrath, J. J., Visscher, P. M., Wray, N. R., et al. (2018). Causal associations between risk factors and common diseases inferred from gwas summary data. *Nature Communications*, 9(1):1–12.