On Mendelian Randomisation Mixed-Scale Treatment Effect Robust Identification (MR MiSTERI) and Estimation for Causal Inference

Zhonghua Liu[1], Ting Ye[2], Baoluo Sun[3], Mary Schooling[4,6], and Eric Tchetgen Tchetgen[5*]

[1]Department of Statistics and Actuarial Science, University of Hong Kong, Hong Kong
[2,5]Department of Statistics, The Wharton School of Business, University of Pennsylvania, Philadelphia, PA, USA
[3]Department of Statistics and Applied Probability, National University of Singapore, Singapore
[4]CUNY Graduate School of Public Health and Health Policy, New York, NY, USA
[6]School of Public Health, Li Ka Shing Faculty of Medicine, University of Hong Kong, Hong Kong
[*]ett@wharton.upenn.edu

March 30, 2021

Abstract

Standard Mendelian randomization analysis can produce biased results if the genetic variant defining instrumental variable (IV) is confounded and/or has a horizontal pleiotropic effect on the outcome of interest not mediated by the treatment. We provide novel identification conditions for the causal effect of a treatment in the presence of unmeasured confounding by leveraging an invalid IV for which both the IV independence and exclusion restriction assumptions may be violated. The proposed Mendelian Randomization Mixed-Scale Treatment Effect Robust Identification (MR MiSTERI) approach relies on (i) an assumption that the treatment effect does not vary with the invalid IV on the additive scale; and (ii) that the selection bias due to confounding does not vary with the invalid IV on the odds ratio scale; and (iii) that the residual variance for the outcome is heteroscedastic and thus varies with the invalid IV. Although assumptions (i) and (ii) have, respectively appeared in the IV literature, assumption (iii) has not; we formally establish that their conjunction can identify a causal effect even with an invalid IV subject to pleiotropy. MR MiSTERI is shown to be particularly advantageous in the presence of pervasive heterogeneity of pleiotropic effects on the additive scale. For estimation, we propose a simple and consistent three-stage estimator that can be used as preliminary estimator to a carefully constructed one-step-update estimator, which is guaranteed to be more efficient under the assumed model. In order to incorporate multiple, possibly correlated and weak IVs, a common challenge in MR studies, we develop a MAny Weak Invalid Instruments (MR MaWII MiSTERI) approach for strengthened identification and improved estimation accuracy. Both simulation studies and UK Biobank data analysis results demonstrate the robustness of the proposed MR MiSTERI method.

Keywords: Treatment effect on the treated, Mendelian randomization, horizontal pleiotropy, invalid instrument, unmeasured confounding, UK Biobank, weak instrument.

1 Introduction

Mendelian randomization (MR) is an instrumental variable (IV) approach that uses genetic variants, for example, single-nucleotide polymorphisms (SNPs), to infer the causal effect of a modifiable risk treatment on a health outcome (Smith and Ebrahim 2003). MR has recently gained popularity in epidemiological studies because, under certain conditions, it can provide unbiased estimates of causal effects even in the presence of unmeasured exposure-outcome confounding. For example, findings from a recent MR analysis assessing the causal relationship between low-density lipoprotein cholesterol and coronary heart disease (Ference et al. 2017) in an observational study are consistent with the results of earlier randomized clinical trials (Scandinavian Simvastatin Survival Study Group 1994).

For a SNP to be a valid IV, it must satisfy the following three core assumptions (Didelez and Sheehan 2007; Lawlor et al. 2008):

(A1) IV relevance: the SNP must be associated (not necessarily causally) with the exposure;

(A2) IV independence: the SNP must be independent of any unmeasured confounder of the exposure-outcome relationship;

(A3) Exclusion restriction: the SNP cannot have a direct effect on the outcome variable not mediated by the treatment, that is, no horizontal pleiotropic effect can be present.

The causal diagram in Figure 1(a) graphically represents the three core assumptions for a valid IV. It is well-known that even when assumptions (A1)-(A3) hold for a given IV, the average causal effect of the treatment on the outcome cannot be point identified without an additional condition, the nature of which dictates the interpretation of the identified causal effect. Specifically, Angrist et al. (1996) proved that under (A1)-(A3) and a monotonicity assumption about the IV-treatment relationship, the so-called complier average treatment effect is nonparametrically identified. More recently, Wang and Tchetgen Tchetgen (2018) established identification of population average causal effect under (A1)-(A3) and an additional assumption of no unmeasured common effect modifier of the association between the IV and the endogenous variable, and the treatment causal effect on the outcome. A special case of this assumption is that the association between the IV and the treatment variable is constant on the additive scale across values of the unmeasured confounder (Tchetgen Tchetgen et al. (2020)). In a separate strand of work, Robins (1994) identified the effect of treatment on the treated under (A1)-(A3) and a so-called "no current-treatment value interaction" assumption (A.4a) that the effect of treatment on the treated is constant on the additive scale across values of the IV. In contrast, Liu et al. (2020) established identification of the effect of treatment on the treated (ETT) under (A1)-(A3), and an assumption (A.4b) that the selection bias function defined as the odds ratio association between the potential outcome under no treatment and the treatment variable, is constant across values of the IV. Note that under the IV DAG Figure 1(a), assumption (A1) is empirically testable while (A2) and (A3) cannot be refuted empirically without a different assumption being imposed (Glymour et al. 2012). Possible violation or near violation of assumption (A1) known as the weak IV problem poses an important challenge in MR studies as the associations between a single SNP IV and complex traits can be fairly weak (Staiger and Stock 1997; Stock et al. 2002). But massive genotyped datasets have provided many such weak IVs. This has motivated a rich body of work addressing the weak IV problem under a many weak instruments framework, from a generalized method of moments perspective given individual level data (Chao and Swanson 2005; Newey and Windmeijer 2009; Davies et al. 2015), and also from a summary-data perspective (Zhao et al. 2019a,b; Ye et al. 2019). Violation of assumption (A2) can occur due to population stratification, or when a selected SNP IV is in linkage disequilibrium (LD) with another genetic variant which has a direct effect on the outcome (Didelez and Sheehan 2007). Violations of (A3) can occur when the chosen SNP IV has a non-null direct effect on the outcome not mediated by the exposure, commonly referred to as horizontal pleiotropy and is found to be widespread (Solovieff et al. 2013; Verbanck et al. 2018). A standard MR analysis (i.e. based on standard IV methods such as 2SLS) with an invalid IV that violates any of those three core assumptions might yield biased causal effect estimates.

Methods to address possible violations of (A2) or (A3) given a single candidate IV are limited. Two methods have recently emerged as potentially robust against such violation under certain

Figure 1: Directed acyclic graph (DAG) with an instrument (Z), an outcome (Y), a treatment (A) and unmeasured confounders (U). The left panel shows a valid Mendelian randomization study and the right panel shows violations of IV independence and exclusion restriction assumptions.

conditions. The first, known as MR-GxE, assumes that one has observed an environmental factor (E), which interacts with the invalid IV in its additive effects on the treatment of interest, and that such interaction is both independent of any unmeasured confounder of the exposure-outcome relationship, and does not operate on the outcome in view (Spiller et al. 2019, 2020). In other words, MR-GxE essentially assumes that while the candidate SNP (the G variable) may not be a valid IV, its additive interaction with an observed covariate constitutes a valid IV which satisfies (A1)-(A3). In contrast, MR GENIUS relies on an assumption that the residual variance of the first stage regression of the treatment on the candidate IV is heteroscedastic with respect to the candidate IV, i.e., the variance of the treatment depends on the IV, an assumption that may be viewed as strengthening of the IV relevance assumption (Lewbel 2012; Tchetgen Tchetgen et al. 2020). Interestingly, as noted by Tchetgen Tchetgen et al. (2020), existence of a GxE interaction that satisfies conditions (A1)-(A3) required by MR GxE had such an E variable (independent of G) been observed, necessarily implies that the heteroscedasticity condition required by MR GENIUS holds even when the relevant E variable is not directly observed. Furthermore, as it is logically possible for heteroscedasticity of the variance of the treatment to operate even in absence of any GxE interaction, MR GENIUS can be valid in certain settings where MR GxE is not.

In this paper, we develop an alternative robust MR approach for estimating a causal effect of a treatment subject to unmeasured confounding, leveraging a potentially invalid IV which fails to fulfill either IV independence or exclusion restriction assumptions. The proposed "Mendelian Randomization Mixed-Scale Treatment Effect Robust Identification" (MR MiSTERI) approach relies on (i) an assumption that the treatment effect does not vary with the invalid IV on the additive scale; and (ii) that the selection bias due to confounding does not vary with the invalid IV on the odds ratio scale; and (iii) that the residual variance for the outcome is heteroscedastic and thus varies with the invalid IV. Note that assumption (iii) is empirically testable. Although assumptions (i) and (ii) have, respectively appeared in the IV literature, assumption (iii) has not; we formally establish that their conjunction can identify a causal effect even with an invalid IV subject to pleiotropy. MR MiSTERI is shown to be particularly advantageous in the presence of pervasive heterogeneity of pleiotropic effects on the additive scale, a setting in which both MR GxE and MR GENIUS can be severely biased whenever heteroscedastic first stage residuals can fully be attributed to latent heterogeneity in SNP-treatment association (Spiller et al. 2020). For estimation, we propose a simple and consistent three-stage estimator that can be used as a preliminary estimator to a carefully constructed one-step-update estimator, which is guaranteed to be more efficient under the assumed model. In order to incorporate multiple, possibly correlated and weak IVs, a common challenge in MR studies, we develop a MAny Weak Invalid Instruments (MaWII MR MiSTERI) approach for strengthened identification and improved accuracy. Simulation study results show that our proposed MR method gives consistent estimates of the causal parameter and the selection bias parameter with nominal confidence interval coverage with an invalid IV, and the accuracy is further improved with multiple weak invalid IVs. For illustration, we apply our method to the UK Biobank data set to estimate the causal effect of body mass index (BMI) on average glucose level. We also develop an R package freely available for public use at `https://github.com/zhonghualiu/MRMiSTERI`.

2 Methods

2.1 Identification with a Binary Treatment and a Possibly Invalid Binary IV

Suppose that we observe data $O_i = (Z_i, A_i, Y_i)$ of sample size n drawn independently from a common population, where Z_i, A_i, Y_i denote a SNP IV, a treatment and a continuous outcome of interest for the ith subject ($1 \leq i \leq n$), respectively. In order to simplify the presentation, we drop the sample index i and do not consider observed covariates at this stage, although all the conclusions continue to hold within strata defined by observed covariates. Let z, a, y denote the possible values that Z, A, Y could take, respectively. Let Y_{az} denote the potential outcome, had possibly contrary to fact, A and Z been set to a and z respectively, and let Y_a denote the potential outcome had A been set to a. We are interested in estimating the ETT defined as

$$\beta(a) = E(Y_a - Y_0 \mid A = a).$$

To facilitate the exposition, consider the simple setting where both the treatment and the SNP IV are binary, then the ETT simplifies to $\beta = E(Y_1 - Y_0 \mid A = 1)$. By the consistency assumption (Hernan and Robins 2020), we know that $E(Y_1 \mid A = 1) = E(Y \mid A = 1)$. However, the expectation of the potential outcome Y_0 among the exposed subpopulation $E(Y_0 \mid A = 1)$ cannot empirically be observed due to possible unmeasured confounding for the exposure-outcome relationship. The following difference captures this confounding bias on the additive scale $Bias = E(Y_0 \mid A = 1) - E(Y_0 \mid A = 0)$, which is exactly zero only when exposed and unexposed groups are exchangeable on average (i.e. under no confounding) (Hernan and Robins 2020), and is otherwise not null. With this representation and the consistency assumption, we have

$$E(Y \mid A = 1) - E(Y \mid A = 0) = \beta + Bias.$$

This simple equation implies that one can only estimate the sum of the causal effect β and the confounding bias but cannot tease them apart using the data (A, Y) only. With the availability of a binary candidate SNP IV Z that is possibly invalid, we can further stratify the population by Z to obtain under consistency:

$$E(Y \mid A = 1, Z = z) - E(Y \mid A = 0, Z = z) = \beta(z) + Bias(z),$$

where z is equal to either 0 or 1, and $\beta(z) = E(Y_a - Y_0 \mid A = a, Z = z)$, $Bias(z) = E(Y_0 \mid A = 1, Z = z) - E(Y_0 \mid A = 0, Z = z)$ denote the causal effect and bias in the stratified population with the IV taking value z. Note that there are only two equations but four unknown parameters: $\beta(z)$, $Bias(z), z = 0, 1$. Therefore, the causal effect cannot be identified without imposing assumptions to reduce the total number of parameters to two.

Our first assumption extends the no current-treatment value interaction assumption (A.4a) originally proposed by Robins (1994), that the causal effect does not vary across the levels of the SNP IV so that $\beta(z)$ is a constant as function of z. Formally stated:

(B1) Homogeneous ETT assumption: $E(Y_{a=1,z} - Y_{a=0,z} \mid A = 1, Z = z) = \beta$.

It is important to note that this assumption does not imply the exclusion restriction assumption (A3); it is perfectly compatible with presence of a direct effect of Z on Y (the direct arrow from Z to Y is present in Figure 1(b)), i.e. $E(Y_{a=0,z=1} - Y_{a=0,z=0} \mid A = 1, Z = z) \neq 0$ as well as unmeasured confounding of the effects of Z on (A,Y), both of which we accommodate.

In order to state our second core assumption, consider the following data generating mechanism for the treatment, which encodes presence of unmeasured confounding by making dependence between A and Y_0 explicit under a logistic model:

$$\text{logit}\left\{Pr(A = 1 \mid Y_0 = y_0, Z = z)\right\} = \gamma_0 + \gamma y_0 + \gamma_z z + \gamma_{y_0 z} y_0 z.$$

This model can of course not be estimated directly from the observed data without any additional assumption because it would require the potential outcome Y_0 be observed both on the untreated (guaranteed by consistency) and the treated. Nevertheless, this model implies that the

conditional (on Z) association between A and Y_0 on the odds ratio scale is $OR(Y_0 = y_0, A = 1 \mid Z = z) = \exp(\gamma y_0 + \gamma_{y_0 z} y_0 z)$. Together, γ and $\gamma_{y_0 z}$ encode the selection bias due to unmeasured confounding on the log-odds ratio scale. If both γ and $\gamma_{y_0 z}$ are zeros, then A and Y_0 are conditionally (on Z) independent, or equivalently, there is no residual confounding bias upon conditioning on Z. Our second identifying assumption formally encodes assumption A4.b of Liu et al. (2020) of homogeneous odds ratio selection bias $\gamma_{y_0 z} = 0$:

(B2) Homogeneous selection bias: $OR(Y_0 = y_0, A = a \mid Z = z) = \exp(\gamma a y_0)$.

Assumption (B2) states that the selection bias on the odds ratio scale is homogeneous across levels of the IV. Thus, this assumption allows for the presence of unmeasured confounding which, upon setting $\gamma_{y_0 z} = 0$ is assumed to be on average balanced with respect to the SNP IV (on the odds ratio scale). Following Liu et al. (2020), we have that under (B2)

$$E(Y_0 \mid A = a, Z = z) = \frac{E\{Y \exp(\gamma a Y) \mid A = 0, Z = z\}}{E\{\exp(\gamma a Y) \mid A = 0, Z = z\}}$$

Define $\varepsilon = Y - E(Y \mid A, Z)$, and suppose that $\varepsilon \mid A, Z \sim N(0, \sigma^2(Z))$, then after some algebra we have,

$$E(Y_0 \mid A = 1, Z = z) = E(Y \mid A = 0, Z = z) + \gamma \sigma^2(Z).$$

The selection bias on the odds ratio scale does not vary with the levels of the IV, however in order to achieve identification, the bias term $\gamma\sigma^2(Z)$ must depend on the IV. This observation motivates our third assumption,

(B3) Outcome heteroscedasticity, that is $\sigma^2(Z)$ cannot be a constant.

Assumption (B3) is empirically testable using existing statistical methods for detecting non-constant variance in regression models. As shown by Paré et al. (2010), observed or unobserved gene–gene (GxG) and/or gene–environment (GxE) can result in changes in variance of a quantitative trait per genotype. Paré et al. (2010) found several SNPs associated with the variance of inflammation markers. Therefore, GxE interaction can be inferred from genetic variants associated with phenotypic variability without the need of measuring environmental factors, usually termed variance quatitative trait loci (vQTL) screening (Marderstein et al. 2021). In particular, Wang et al. (2019) performed a genome-wide vQTL analysis of about 5.6 million variants on 348,501 unrelated individuals of European ancestry in the UK Biobank and identified 75 significant vQTLs for 9 quantitative traits (out of 13 traits under study). Hence, one can generally expect this novel identification assumption (B3) to hold in the presence of pervasive heterogeneity of pleiotropic effects on the additive scale for the outcome, a setting in which both MR GxE and MR GENIUS can be severely biased. Under assumptions (B1)-(B3), we have

$$E(Y \mid A = 1, Z = z) - E(Y \mid A = 0, Z = z) = \beta + \gamma\sigma^2(z).$$

Note that $\sigma^2(z)$ is the variance of Y given z, and thus can be estimated with no bias. For the binary IV Z, $\sigma^2(Z = 0)$ and $\sigma^2(Z = 1)$ are the variance of Y within each IV stratum and can easily be estimated using the sample variance in each group. We will describe estimating procedures in the next section for more general settings where Z is not necessarily binary. Importantly, in the present case, we now have two equations and two unknown parameters β, γ. Denote $D(Z = z) = E(Y \mid A = 1, Z = z) - E(Y \mid A = 0, Z = z)$. We have the following main identification result.

Theorem 1. *Under assumptions (B1)-(B3), the selection bias parameter and the causal effect of interest are uniquely identified as follows:*

$$\begin{aligned}
\gamma &= \frac{D(Z = 1) - D(Z = 0)}{\sigma^2(Z = 1) - \sigma^2(Z = 0)} \\
\beta &= E\{D(Z) - \frac{D(Z = 1) - D(Z = 0)}{\sigma^2(Z = 1) - \sigma^2(Z = 0)}\sigma^2(Z)\} \\
&= D(z) - \frac{D(Z = 1) - D(Z = 0)}{\sigma^2(Z = 1) - \sigma^2(Z = 0)}\sigma^2(z); z = 0, 1.
\end{aligned}$$

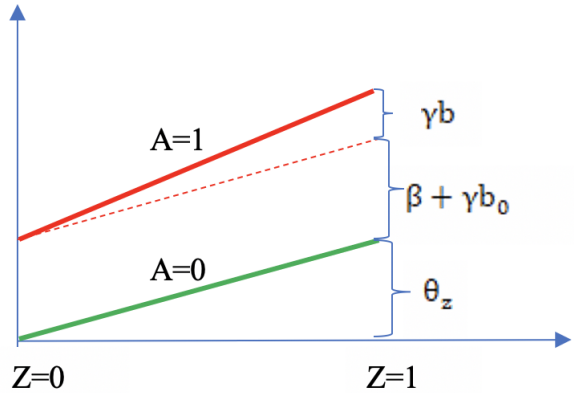

Figure 2: An illustration of our identification strategy for a binary treatment A and binary IV Z. The green line represents $E(Y \mid A = 0, Z = z)$ and its difference between $Z = 1$ and $Z = 0$ is θ_z. The red dashed line (parallel to the green line) represents the hypothetical setting when $\sigma^2(Z = z)$ is constant, and the red solid line represents the setting when $\sigma^2(Z = z)$ is not a constant function in Z. Note that b_0 and b are treated as known.

In an observed sample, one can simply use the sample versions of unknown quantities to obtain estimates for β and γ respectively. Standard errors can be deduced by a direct application of the multivariate delta method, or by using resampling techniques such as the jackknife or the bootstrap.

Remark 1. Note that as stated in the theorem, the causal parameter β can be estimated based on data among either $Z = 0$ or $Z = 1$ group. In practical settings, in order to improve efficiency, one may take either unweighted average of the two estimates as given in the theorem, or their inverse variance weighted average. In fact, one may fit a saturated linear model for the variance $\sigma^2(Z = z) = b_0 + bz$; We give a graphical illustration of our identification strategy as shown in Figure 2. By using the outcome heteroscedasticity assumption (B3), we can identify the selection bias parameter γ and then the causal effect parameter β. Without the assumption (B3) as shown by the red dashed line, we can only obtain $\beta + \gamma b_0$ and thus cannot tease apart the causal effect and the selection bias.

2.2 Identification with a Continuous Treatment and a Possibly Invalid Discrete IV

The identification strategy in the previous section extends to the more general setting of a continuous treatment and a more general discrete IV, for example, the SNP IV taking values 0, 1, 2 corresponding to the number of minor alleles. To do so, we introduce the notion of generalized conditional odds ratio function as a measure of the conditional association between a continuous treatment and a continuous outcome. To ground idea, consider the following model for the treatment free potential outcome:

$$Y_0 = E(Y_0|A, Z) + \sigma(Z)\epsilon$$

where ϵ is independent standard normal; then one can show that the generalized conditional odds ratio function associated with Y_0 and A given Z with $(Y_0 = 0, A = 0)$ taken as reference values (Chen 2007) is given by

$$OR(y_0, a \mid z) = \frac{f(y_0 \mid a, z) f(y_0 = 0 \mid a = 0, z)}{f(y_0 \mid a = 0, z) f(y_0 = 0 \mid a, z)} = \exp\{(E(Y_0|A = a, z) - E(Y_0|A = 0, z))y_0/\sigma^2(z)\}.$$

By Assumption (B2), we have

$$E(Y_0|A, Z) - E(Y_0|A = 0, Z) = \gamma\sigma^2(Z)A.$$

Therefore,

$$\mu(a,z) = E(Y \mid A=a, Z=z) = \beta a + \gamma\sigma^2(Z=z)a + E(Y \mid A=0, Z=z).$$

For practical interpretation, we assume A has been centered so that $A=0$ actually represents average treatment value in the study population. Assume the following model $E(Y \mid A=0, Z=z) = \theta_0 + \theta_z z$ and a log linear model for $\sigma^2(Z)$ such that

$$\begin{aligned} \mu(a,z) &= \beta A + \gamma A\sigma^2(z) + \theta_0 + \theta_z z, \\ log(\sigma^2(z)) &= \eta_0 + \eta_z z. \end{aligned}$$

The conditional distribution of Y given A and Z is

$$Y \mid A=a, Z=z \sim N\left(\mu(a,z), \sigma^2(z)\right).$$

Then we can simply use standard maximum likelihood estimation (MLE) to obtain consistent and fully efficient estimates for all parameters under the assumed model. However, the likelihood may not be log-concave and direct maximization of the likelihood function with respect to all parameters might not always converge. Instead, we propose a novel three-stage estimation procedure which can provide a consistent but inefficient preliminary estimator of unknown parameters. We then propose a carefully constructed one-step update estimator to obtain a consistent and fully efficient estimator. Our three-stage estimation method works as follows.

Stage 1: Fit the following linear regression using standard (weighted) least-squares

$$Y = \theta_0 + \theta_a A + \theta_z Z + \theta_{az} AZ + \varepsilon.$$

We obtain parameter estimates $\hat{\theta}_0, \hat{\theta}_a, \hat{\theta}_z, \hat{\theta}_{az}$, $\hat{E}(Y \mid A=0, Z) = \hat{\theta}_0 + \hat{\theta}_z Z$, with corresponding estimated residuals $\hat{\varepsilon} = Y - \hat{\theta}_0 - \hat{\theta}_a A - \hat{\theta}_z Z - \hat{\theta}_{az} AZ$.

Stage 2: Regress the squared residuals $\hat{\varepsilon}^2$ on $(1, Z)$ in a log-linear model and obtain parameter estimates $\hat{\eta}_0, \hat{\eta}_z$ and $\hat{\sigma}^2(Z) = \exp(\hat{\eta}_0 + \hat{\eta}_z Z)$.

Stage 3: Regress $Y - \hat{E}(Y \mid A=0, Z)$ on A and $A\hat{\sigma}^2(Z)$ without an intercept term and obtain the two corresponding regression coefficients estimates $\hat{\beta}$ and $\hat{\gamma}$ respectively. Alternatively, we may also fit $E(Y \mid A, Z) = \beta A + \gamma A\hat{\sigma}^2(Z) + \theta_0 + \theta_z Z$.

The proposed three-stage estimation procedure is computationally convenient, appears to always converge and provides a consistent estimator for all parameters under the assumed model. Next we propose the following one-step update to obtain a fully efficient estimator. Consider the following normal model

$$\frac{Y - \{\beta A + \gamma A \exp(\eta_0 + \eta_z Z) + \theta_0 + \theta_z Z\}}{\sqrt{\exp(\eta_0 + \eta_z Z)}} \sim N(0,1). \tag{1}$$

Denote $\Theta = (\beta, \gamma, \eta_0, \eta_z, \theta_0, \theta_z)$ and denote the log-likelihood function as $l(\Theta; Y, A, Z)$ for an individual. With a sample of size n, the log-likelihood function is

$$l_n(\Theta) = \sum_{i=1}^{n} l_i(\Theta; Y_i, A_i, Z_i).$$

The score function $S_n(\Theta) = \partial l_n(\Theta)/\partial\Theta$ is defined as the first order derivative of the log-likelihood function with respect to Θ, and use $i_n(\Theta) = \partial^2 l_n(\Theta)/\partial\Theta\partial\Theta^T$ to denote the second order derivative of the log-likelihood function. Suppose we have obtained a consistent estimator $\tilde{\Theta}^{(0)}$ of Θ using the three-stage estimating procedure, then one can show that a consistent and fully efficient estimator is given by the one-step update estimator $\tilde{\Theta}^*$

$$\tilde{\Theta}^* = \tilde{\Theta}^{(0)} - \{i_n(\tilde{\Theta}^{(0)})\}^{-1} S_n(\tilde{\Theta}^{(0)})). \tag{2}$$

Remark 2. In biomedical studies, continuous outcomes of interest are often approximately normally distributed, and the normal linear regression is also routinely used by applied researchers.

When the normal assumption is violated, a typical remedy is to apply a transformation to the outcome prior to modeling it, such as the log-transformation or the more general Box-Cox transformation to achieve approximate normality. Alternatively, in section 2.4 we propose to model the error distribution as a more flexible Gaussian mixture model which is more robust than the Gaussian model. Alternatively, as described in the Supplemental Materials, one may consider a semiparametric location-scale model which allows the distribution of standardized residuals to remain unrestricted, a potential more robust approach although computationally more demanding.

2.3 Estimation and Inference under Many Weak Invalid IVs

Weak identification bias is a salient issue that needs special attention when using genetic data to strengthen causal inference, as most genetic variants typically have weak association signals. When many genetic variants are available, we recommend using the conditional maximum likelihood estimator (CMLE) based on (1), where Z is replaced with a multi-dimensional vector. Let $\hat{\Theta}$ be the solution to the corresponding score functions, i.e., $S_n(\hat{\Theta}) = 0$; let

$$I_1(\Theta) = -E\left\{\frac{\partial^2}{\partial\Theta\partial\Theta^T} l_1(\Theta, Y_1, A_1, Z_1)\right\}$$

be the Fisher information matrix based on the conditional likelihood function for one observation. Let k be the total number of parameters in Θ, which is equal to $4 + 2p$ when Z is p dimensional. When $\lambda_{\min}\{nI_1(\Theta)\}/k \to \infty$ with $\lambda_{\min}\{nI_1(\Theta)\}$ being the minimum eigenvalue of $nI_1(\Theta)$, we show in the Supplementary Materials (Theorem 2) that $\hat{\Theta}$ is consistent and asymptotically normal as the sample size goes to infinity, and satisfies

$$\sqrt{n}(\hat{\Theta} - \Theta) \xrightarrow{d} N\big(0, \{I_1(\Theta)\}^{-1}\big). \tag{3}$$

In other words, the CMLE is robust to weak identification bias and the usual inference procedure can be directly applied.

The key condition $\lambda_{\min}\{nI_1(\Theta)\}/k \to \infty$ warrants more discussion. Note that the quantity $\lambda_{\min}\{nI_1(\Theta)\}/k$ measures the ratio between the amount of information that a sample of size n carries about the unknown parameter and the number of parameters. In classical strong identification scenarios where the minimum eigenvalue of $I_1(\Theta)$ is assumed to be lower bounded by a constant, then the condition simplifies to that the number of parameter is small compared with sample size, i.e., $n/k \to \infty$. However, when identification is weak, the minimum eigenvalue of $I_1(\Theta)$ can be small. In practice, the condition $\lambda_{\min}\{nI_1(\Theta)\}/k \to \infty$ can be evaluated by

$$\hat{\kappa} = -\lambda_{\min}\{i_n(\hat{\Theta})\}/k,$$

which is the ratio between the minimum eigenvalue of the negative Hessian matrix and the number of parameters. We remark that the condition stated in the Supplementary Materials (Theorem 2) is more general and applies to general likelihood-based methods, which, for example, implies the condition for consistency for the profile likelihood estimator (MR.raps) in Zhao et al. (2019b). In practice, we recommend checking that the $\hat{\kappa}$ is at least greater than 10 based on our simulation studies. For valid statistical inference, a consistent estimator for the variance covariance matrix for $\hat{\Theta}$ is simply the negative Hessian matrix $-i_n(\hat{\Theta})$, whose corresponding diagonal element estimates the variance of the treatment effect estimator $\hat{\beta}$. Other variants of the CMLE can also be used, for example, the one-step iteration estimator $\tilde{\Theta}^*$ in (2) is asymptotically equivalent to the CMLE $\hat{\Theta}$ as long as the initial estimator $\tilde{\Theta}^{(0)}$ is $\sqrt{n}$-consistent (Shao 2003).

2.4 More Robust Gaussian Mixture Model for the Outcome

As mentioned in Remark 2, although many continuous outcomes can be well modelled using normal linear model, however sometimes, a more robust model for the outcome error term is desired. As shown by Verbeke and Lesaffre (1996), a Gaussian mixture model is more general and thus more robust than the normal model for modelling an outcome error distribution. Consider the general location-scale model

$$Y = E(Y|A, Z) + \sigma(Z)\epsilon,$$

where $\epsilon \perp A, Z$. Under assumptions (B1)-(B3), the conditional mean function is given by

$$E\left(Y|A=a,Z=z\right)=\beta a+E\left(Y|A=0,Z=z\right)+\frac{\partial}{\partial(\gamma a)}\ln\int\exp\left[\gamma a\sigma\left(Z=z\right)\epsilon\right]dF\left(\epsilon\right),$$

where $F(\cdot)$ denotes the cumulative distribution function of ϵ. We assume that the outcome error distribution can be reasonably approximated by a Gaussian mixture model with enough components (Verbeke and Lesaffre 1996). Specifically, let $f(\epsilon)=\sum_{k=1}^{K}\pi_k\phi_k(\epsilon)$ where $\phi_k(\cdot)$ is the normal density with mean μ_k and variance δ_k^2 satisfying the constraints

$$\sum_{k=1}^{K}\pi_k=1,\quad E(\epsilon)=\sum_{k=1}^{K}\pi_k\mu_k=0,\quad \mathrm{Var}\left(\epsilon\right)=\sum_{k=1}^{K}\pi_k\left(\delta_k^2+\mu_k^2\right)=1.$$

The conditional mean function with Gaussian mixture error is given by

$$\beta a+E\left(Y|A=0,Z=z\right)+\frac{\sigma\left(Z=z\right)}{\sum_{k=1}^{K}\pi_k\omega_k}\left\{\left[\sum_{k=1}^{K}\pi_k\omega_k\mu_k\right]\right\}+\gamma a\sum_{k=1}^{K}\frac{\pi_k\omega_k\delta_k^2\sigma^2\left(Z=z\right)}{\sum_{k=1}^{K}\pi_k\omega_k},$$

where $\omega_k=\exp\left(\gamma a\sigma\left(Z=z\right)\mu_k+\delta_k^2\left[\gamma a\sigma\left(Z=z\right)\right]^2/2\right)$. The conditional distribution is $\{Y-E\left(Y|A=a,Z=z\right)\}/\sigma\left(Z=z\right)\sim\sum_{k=1}^{K}\pi_kN(\mu_k,\delta_k^2)$. In practice, estimation of (β,γ), which are of primary interest, may proceed under a user-specified integer value $K<\infty$ as well as parametric models for $E(Y|A=0,Z=z)$ and $\sigma(Z=z)$ via an alternating optimization algorithm which we describe in the Supplementary Materials.

3 Simulation Studies

In this section, we conduct extensive simulation studies to evaluate the finite-sample performance of MR-MiSTERI compared to its main competitors, standard two stage least squares (TSLS) and MR GENIUS. We considered sample sizes of $n=10000, 30000, 100000$. The IV Z is generated from a Binomial distribution with probability equal to 0.3 (minor allele frequency). The treatment A is generated from a standard normal distribution $N(0,1)$. The outcome Y is also generated from a normal distribution with mean and variance compatible with assumptions B1-B3:

$$\begin{aligned}E(Y\mid A,Z)&=0.8A+0.2A\sigma^2(Z)+1+0.3Z\\ \sigma^2(Z)&=\exp(0.1+\eta_zZ)\end{aligned}$$

where η_z is set to be 0.2, 0.15, 0.1, 0.05. Under this model, we have that the ETT is equal to $\beta=0.8$ and $\gamma=0.2$. We generated in total 1000 such data sets and applied our proposed method to each of them along with TSLS and MR GENIUS. We found that TSLS gives severely biased estimates of the causal effect with estimates as extreme as -195.72 and standard errors as large as 381.57. MR GENIUS is also severely biased as the average point estimate for the causal effect is -14.99. Numerical results for TSLS and MR GENIUS are expected as their corresponding identification conditions are not met in the simulation setup. Specifically, both exclusion restriction and heteroscedasticity of the treatment with respect to Z assumptions do not hold. Simulation results for MR MiSTERI are summarized in Table 1. We found that when sample size is 10000, the bias and standard error of the causal effect and selection bias parameter estimates become larger when the IV strength η_z decreases from 0.2 to 0.1. The causal effect is less sensitive to the IV strength compared to the selection bias parameter γ. As we increase sample size, the bias decreases and becomes negligible at n=100000. Confidence intervals achieved the nominal 95% coverage.

The second simulation study is designed to evaluate the finite sample performance of the proposed three-stage estimator and the CMLE in the presence of many weak invalid IVs. The sample size is set to be 100000, each IV $Z_j, j=1,\ldots,p$, is still generated to take values 0,1,2 with minor allele frequency equal to 0.3. The treatment A is generated from standard normal distribution.

Table 1: Simulation results for estimating the confounding bias parameter γ and the ETT β using the one-step estimator approach (without covariates) based on 1000 Monte Carlo experiments with a range of sample sizes n. $\hat{\gamma}$ is the averaged point estimates of γ, and bias is calculated in percentage format. Likewise for $\hat{\beta}$. SE is the averaged standard error. SD stands for the sample standard deviation of the 1000 point estimates for γ and β. Coverage is calculated as the proportion of 95% confidence intervals that contains the true parameter value among those 1000 experiments. We vary the value of η_z in $\sigma^2(Z) = exp(\eta_0 + \eta_z Z))$ to assess the impact of decreasing IV strength on the estimation results.

n	η_z	(β, γ)	$\hat{\beta}$	Bias	SE	SD	Cover	$\hat{\gamma}$	Bias	SE	SD	Cover
1e4	0.2	(0.8,0.2)	0.806	0.74 %	0.092	0.094	93.8%	0.196	-2.12%	0.074	0.075	94.6%
1e4	0.15	(0.8,0.2)	0.788	-1.44%	0.124	0.126	95.6%	0.210	4.77 %	0.103	0.104	95.6%
1e4	0.1	(0.8,0.2)	0.778	-2.76%	0.197	0.210	96.2%	0.219	9.32%	0.168	0.181	96.0%
3e4	0.1	(0.8,0.2)	0.789	-1.35%	0.104	0.106	96.0%	0.209	4.58 %	0.089	0.090	95.8%
3e4	0.05	(0.8,0.2)	0.788	-1.46%	0.226	0.222	97.0%	0.210	5.17%	0.199	0.195	97.4%
1e5	0.05	(0.8,0.2)	0.797	-0.36%	0.113	0.120	95.2%	0.203	1.29%	0.099	0.105	95.2%

The outcome Y is generated from a normal distribution with mean and variance under B1-B3:

$$E(Y \mid A, Z) = 0.8A + 0.2A\sigma^2(Z) - 0.5 + 0.5\sum_{j=1}^{p} Z_j,$$

$$\sigma^2(Z) = \exp\left(0.1 + 0.05\sum_{j=1}^{p} Z_j\right).$$

Simulation results are presented in Table 2, where the standard error (SE) for the three-stage estimator is the bootstrap estimator approximated using 100 Monte Carlo simulations, the SE for the CMLE is obtained from the inverse Hessian matrix discussed in Section 2.3. From results in Table 2, the three-stage estimator has negligible bias when $p = 20$, but shows evidently larger bias when p grows to 50, which is also reflected by the fact that coverage of 95% CI is substantially lower than its nominal level. In contrast, the CMLE shows negligible bias in both scenarios. The standard errors for the CMLE are close to Monte Carlo standard deviation, and the CMLE estimators show nominal coverage in both simulation scenarios. These observations agree with our theoretical assessment that the CMLE is efficient and is robust to many weak invalid IVs. In particular, it is consistent and asymptotically normal as long as the condition $\hat{\kappa}$ is reasonably large. Based on empirical evaluations (in the Supplementary Materials), we recommend checking that the $\hat{\kappa}$ value is at least greater than 10.

We similarly generate the treatment A from the standard normal distribution, the IV Z which takes values in $\{0, 1, 2\}$ with minor allele frequency equal to 0.3. However we generate $Y = E(Y|A, Z) + \sigma(Z)\epsilon$ where $E(Y \mid A, Z) = 0.8A + (1 + 0.3Z) + \sigma(Z)\left\{\frac{\sum_{k=1}^{2}\pi_k\omega_k\left[\mu_k+\delta_k^2\gamma A\sigma(Z)\right]}{\sum_{k=1}^{2}\pi_k\omega_k}\right\}$, $\sigma^2(Z) = \exp(0.1 + \eta_z Z)$, and the error ϵ is generated from a Gaussian mixture distribution with two components,

$$\varepsilon_1 \sim N(\mu_1, \delta_1^2), \quad \varepsilon_2 \sim N(\mu_2, \delta_2^2), \quad K \sim \text{Bernoulli}(p = \pi_1), \quad \epsilon = K\varepsilon_1 + (1 - K)\varepsilon_2,$$

with the parameter values $(\pi_1, \pi_2) = (0.4, 0.6)$, $(\mu_1, \mu_2) = (-0.6, 0.4)$ and $(\delta_1, \delta_2) = (0.5, 1.049)$. We vary the value of η_z in the set $\{0.1, 0.25, 0.5\}$ to assess the impact of IV strength. The results using the proposed Gaussian mixture method with $K = 2$ are summarized in Table 3. There is noticeable finite-sample bias and under-coverage when $\eta_z = 0.1$, especially for estimation of the selection bias parameter γ, but the performance improves with larger η_z or sample size.

Table 2: Simulation results for estimating the confounding bias parameter $\gamma = 0.2$ and the ETT $\beta = 0.8$ using the three-stage estimator and the CMLE based on 1000 Monte Carlo experiments, with $n = 100000$ and varying p. The third column is the averaged $\hat{\kappa} = -\lambda_{\min}\{i_n(\hat{\Theta})\}/k$, which is the consistency and asymptotic normality condition for the CMLE and is preferably to be large. $\hat{\gamma}$ is the averaged point estimates of γ, and bias is calculated in percentage format. Likewise for $\hat{\beta}$. SE is the averaged standard error. SD stands for the sample standard deviation of the 1000 point estimates for γ and β. Coverage is calculated as the proportion of 95% confidence intervals that contains the true parameter value among those 1000 experiments.

n	p	$\hat{\kappa}$	Method	$\hat{\beta}$	Bias	SE	SD	Cover	$\hat{\gamma}$	Bias	SE	SD	Cover
1e5	20	15.55	3-stage	0.806	0.76%	0.034	0.034	94.1%	0.197	-1.49%	0.017	0.017	94.2%
			CMLE	0.799	-0.14%	0.033	0.034	94.9%	0.201	0.28%	0.017	0.017	95.2%
1e5	50	4.51	3-stage	0.818	2.21%	0.038	0.037	91.9%	0.197	-1.62%	0.008	0.008	93.2%
			CMLE	0.801	0.11%	0.036	0.036	95.3%	0.200	0.02%	0.007	0.007	94.6%

Table 3: Simulation results for estimating the confounding bias parameter $\gamma = 0.2$ and the ETT $\beta = 0.8$ under Gaussian mixture error based on 1000 replicates. See the caption of Table 1 for description of the summary statistics.

n	η_z	$\hat{\beta}$	Bias	SE	SD	Cover	$\hat{\gamma}$	Bias	SE	SD	Cover
	0.10	0.771	-3.64%	0.373	0.195	75.6%	0.256	27.78%	0.314	0.211	66.7%
1e4	0.25	0.798	-0.29%	0.056	0.059	93.9%	0.205	2.39%	0.044	0.049	91.4%
	0.50	0.799	-0.12%	0.039	0.039	95.9%	0.201	0.69%	0.026	0.026	96.0%
	0.10	0.780	-2.54%	0.106	0.084	77.0%	0.230	14.95%	0.090	0.101	69.2%
3e4	0.25	0.799	-0.09%	0.032	0.033	95.1%	0.202	0.80%	0.025	0.027	92.9%
	0.50	0.799	-0.07%	0.023	0.022	95.4%	0.201	0.36%	0.015	0.015	95.8%
	0.10	0.793	-0.86%	0.031	0.041	78.0%	0.210	4.96%	0.026	0.052	69.5%
1e5	0.25	0.800	0.03%	0.018	0.018	94.7%	0.200	0.24%	0.014	0.015	93.5%
	0.50	0.800	0.03%	0.012	0.012	95.3%	0.200	0.14%	0.008	0.008	95.6%

4 An Application to the Large-Scale UK Biobank Study Data

UK Biobank is a large-scale ongoing prospective cohort study that recruited around 500,000 participants aged 40-69 in 2006-2010. Participants provided biological samples, completed questionnaires, underwent assessments, and had nurse led interviews. Follow up is chiefly through cohort-wide linkages to National Health Service data, including electronic, coded death certificate, hospital and primary care data (Sudlow et al. 2015). To control for population stratification, we restricted our analysis to participants with self-reported and genetically validated white British ancestry. For quality control, we also excluded participants with (1) excess relatedness (more than 10 putative third-degree relatives) or (2) mismatched information on sex between genotyping and self-report, or (3) sex-chromosomes not XX or XY, or (4) poor-quality genotyping based on heterozygosity and missing rates $> 2\%$.

To illustrate our methods, we extracted a total of 289 010 white British subjects from the UK Biobank data with complete measurements in body mass index (BMI) and glucose levels. We further selected BMI associated SNP potential IVs among the list provided by Sun et al. (2019). We found seven SNPs (in Table 4) that are predictive of the residual variance of glucose levels in our Stage 2 regression (with p-values $P < 0.01$), that is, those seven SNPs might satisfy assumption (B3). The average BMI was 27.39 kg/m^2 (SD: 4.75 kg/m^2), and the average glucose level was 5.12

$mmol/L$ (SD: 1.21 $mmol/L$). We applied MR MiSTERI to this data with the goal of evaluating the causal relationship between BMI as treatment and glucose level as outcome; analysis results are summarized in Table 4.

For comparison purposes, we also include analysis results for standard two-stage least square (TSLS) implemented in the R package *ivreg* (Fox et al. 2020). The allele score for the seven SNPs selected is defined as the signed sum of their minor alleles, where the sign is determined by the regression coefficient in our stage 2 regression. We also implemented MR-GENIUS method (Tchetgen Tchetgen et al. 2020) using the same seven SNPs, the causal effect of BMI on glucose was estimated as 0.046 (se: 0.01, 95% CI: (0.0264,0.0656), P-value: 2.24×10^{-6}). Crude regression analysis by simply regressing glucose on BMI gives an estimate of 0.041 (se: 4.65×10^{-4}, 95% CI: (0.0401, 0.0419), P-value: $< 2 \times 10^{-16}$). Further, we included all seven SNPs into a conventional regression together with BMI, and obtained a similar estimate 0.039 (se: 4.65×10^{-4}, 95% CI: (0.0381, 0.0399), P-value: $< 2 \times 10^{-16}$). It is interesting to consider results obtained by selecting each SNP as single candidate IV as summarized in the table. For instance, take the SNP rs2176040, the corresponding causal effect estimate is $\hat{\beta} = 0.041$ (se: 0.0139, 95% CI: (0.0138, 0.0682), P-value: 0.003), and the selection bias parameter is estimated as $\hat{\gamma}$: 0.059 (se: 0.0098, 95% CI: (0.0398, 0.0782), P-value: 2.93×10^{-9}). However, the TSLS method gives causal effect estimate -1.619 (se: 2.04, 95% CI: (-5.6174, 2.3794), P-value: 0.428). These conflicting findings may be due to SNP rs2176040 being an invalid IV, in which case TSLS likely yields biased results. We further combine 20 SNPs, 13 of which are weakly associated with the residual variance of glucose levels, we obtain using CMLE a causal effect estimate of BMI on glucose $\hat{\beta} = 0.028$ (se: 0.0025, 95% CI: (0.0231, 0.0329), P-value: 6.87×10^{-31}), and a selection bias estimate of $\hat{\gamma} = 0.009$ (se: 0.0017, 95% CI: (0.0057, 0.0123), P-value: 2.82×10^{-7}). The consistency and asymptotic normality condition $\hat{\kappa} = 23.72$, greater than our recommended value 10. This final analysis suggests a somewhat smaller treatment effect size than all the other MR methods, while providing strong statistical evidence of a positive effect with a relatively small but statistically significant amount of confounding bias detected.

5 Discussion

In this paper, we have proposed a novel MR MiSTERI method for identification and inference about causal effects in observational studies. MR MiSTERI leverages a possible association between candidate IV SNPs and the variance of the outcome in order to disentangle confounding bias from the causal effect of interest. Importantly, MR MiSTERI can recover unbiased inferences about the causal effect in view even when none of the candidate IV SNPs is a valid IV. Key assumptions which the approach relies on include no additive interaction involving a candidate IV SNP and the treatment causal effect in the outcome model, and a requirement that the amount of selection bias (measured on the odds ratio scale) remains constant as a function of SNP values. Violation of either assumption may result in incorrect inferences about treatment effect. Although not pursued in this paper, robustness to such violation may be assessed via a sensitivity analysis in which the impact of various departures from the assumption may be explored by varying sensitivity parameters. Alternatively, as illustrated in our application, providing inferences using a variety of MR methods each of which relying on a different set of assumptions provides a robustness check of empirical findings relative to underlying assumptions needed for a causal interpretation of results. We did not include the MR GxE method for numerical comparison because it requires users to specify an observed covariate interacting with the SNP IV. For a particular data analysis problem as we did in Section **??**, we typically do not have sufficient prior knowledge about which covariate (out of thousands) might interact with a given SNP. In contrast, MR GENIUS method requires that the treatment variable admits a residual which is heteroscedastic with respect to the IV, while MR MiSTERI requires that the outcome admits a residual that is heteroscedastic with respect to the IV. In practice, one can in principle perform a formal statistical test to determine which method might be most suitable for a particular data analysis problem. Note that both MR MiSTERI and MR GENIUS methods do not require us to specify any covariate a priori interacting with the candidate SNP IV, and hence are more convenient for routine data analyses.

A notable robustness property of CMLE estimator of MR MiSTERI is that as we formally establish, it is robust to many weak IV bias, thus providing certain theoretical guarantees of reliable

performance in many common MR settings where one might have available a relatively large number of weak candidate IVs, many of which may be subject to pleiotropy. While the proposed methods require a correctly specified Gaussian model for the conditional distribution of Y given A and Z, such an assumption can be relaxed. In fact, we also consider a Gaussian mixture model as a framework to assess and possibly correct for potential violation of this assumption. Furthermore, in the Supplementary Materials, we briefly describe a semiparametric three-stage estimation approach under the location-scale model which allows the distribution of standardized residuals to remain unrestricted. It is of interest to investigate robustness and efficiency properties of this more flexible estimator, which we plan to pursue in future work. It is likewise of interest to investigate whether the proposed methods can be extended to the important setting of a binary outcome, which is of common occurrence in epidemiology and other health and social sciences.

Acknowledgements

Dr. Zhonghua Liu's research is supported by the Start-up research fund (000250348) of the University of Hong Kong. Dr. Eric Tchetgen Tchetgen is supported by grants R01AG065276, R01CA222147 and R01AI27271. Dr. Baoluo Sun is supported by the National University of Singapore Start-up Grant (R-155-000-203-133). This research has been conducted using the UK Biobank resource (https://www.ukbiobank.ac.uk) under application number 44430.
Conflict of interest: None declared.

Table 4: UK Biobank data analysis results. Columns 2-5 contain the results based on the model (1), columns 6-9 contain results based on the Gaussian mixture model described in Section 2.4, columns 10-11 contain results using the standard two-stage least squares using the R package *ivreg*.

SNP	$\hat{\beta}$	$se(\hat{\beta})$	$\hat{\gamma}$	$se(\hat{\gamma})$	$\hat{\beta}^*$	$se(\hat{\beta}^*)$	$\hat{\gamma}^*$	$se(\hat{\gamma}^*)$	$\hat{\beta}^{**}$	$se(\hat{\beta}^{**})$
rs3736485	0.041	0.013	0.0001	0.0094	0.027	0.001	0.0031	0.0002	0.200	0.068
rs1558902	0.023	0.006	0.0128	0.0046	0.040	0.001	0.0025	0.0002	0.064	0.009
rs1808579	0.041	0.014	0.0001	0.0098	0.026	0.001	0.0042	0.0002	0.107	0.038
rs13021737	0.051	0.012	-0.0065	0.0084	0.036	0.001	-0.0043	0.0005	0.030	0.016
rs2176040	0.041	0.014	0.0585	0.0099	0.026	0.001	0.0009	0.0003	-1.619	2.041
rs10968576	0.047	0.016	-0.0040	0.0111	0.046	0.001	-0.0005	0.0004	0.079	0.025
rs10132280	0.030	0.013	0.0081	0.0094	0.062	0.004	-0.0228	0.0005	0.049	0.029
Allele score	0.030	0.003	0.0078	0.0021	0.033	0.005	0.0105	0.0002	0.089	0.009

References

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.

Chao, J. C. and Swanson, N. R. (2005). Consistent estimation with a large number of weak instruments. *Econometrica*, 73(5):1673–1692.

Chen, H. Y. (2007). A semiparametric odds ratio model for measuring association. *Biometrics*, 63(2):413–421.

Davies, N. M., von Hinke Kessler Scholder, S., Farbmacher, H., Burgess, S., Windmeijer, F., and Smith, G. D. (2015). The many weak instruments problem and mendelian randomization. *Statistics in Medicine*, 34(3):454–468.

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

Ference, B., Kastelein, J., Ginsberg, H., Chapman, M., Nicholls, S., Ray, K., Packard, C., Laufs, U., Brook, R., Oliver-Williams, C., Butterworth, A., Danesh, J., Smith, G., Catapano, A., and Sabatine, M. (2017). Association of genetic variants related to cetp inhibitors and statins with lipoprotein levels and cardiovascular risk. *JAMA - Journal of the American Medical Association*, 318(10):947–956.

Fox, J., Kleiber, C., and Zeileis, A. (2020). *ivreg: Two-Stage Least-Squares Regression with Diagnostics*. R package version 0.5-0.

Glymour, M. M., Tchetgen Tchetgen, E. J., and Robins, J. M. (2012). Credible Mendelian Randomization Studies: Approaches for Evaluating the Instrumental Variable Assumptions. *American Journal of Epidemiology*, 175(4):332–339.

Hernan, M. A. and Robins, J. M. (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., 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.

Lewbel, A. (2012). Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. *Journal of Business & Economic Statistics*, 30(1):67–80.

Liu, L., Miao, W., Sun, B., Robins, J., and Tchetgen Tchetgen, E. (2020). Identification and inference for marginal average treatment effect on the treated with an instrumental variable. *Statistica Sinica*.

Marderstein, A. R., Davenport, E. R., Kulm, S., Van Hout, C. V., Elemento, O., and Clark, A. G. (2021). Leveraging phenotypic variability to identify genetic interactions in human phenotypes. *American journal of human genetics*, 108:49–67.

Newey, W. K. and Windmeijer, F. (2009). Generalized method of moments with many weak moment conditions. *Econometrica*, 77(3):687–719.

Paré, G., Cook, N. R., Ridker, P. M., and Chasman, D. I. (2010). On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the women's genome health study. *PLoS Genet*, 6(6):e1000981.

Robins, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics - Theory and Methods*, 23(8):2379–2412.

Scandinavian Simvastatin Survival Study Group (1994). Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the scandinavian simvastatin survival study (4s). *The Lancet*, 344(8934):1383–1389.

Shao, J. (2003). *Mathematical Statistics*. Springer-Verlag New York.

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

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*, 14:483 – 495.

Spiller, W., Hartwig, F. P., Sanderson, E., Smith, G. D., and Bowden, J. (2020). Interaction-based mendelian randomization with measured and unmeasured gene-by-covariate interactions. *medRxiv*.

Spiller, W., Slichter, D., Bowden, J., and Davey Smith, G. (2019). Detecting and correcting for bias in mendelian randomization analyses using gene-by-environment interactions. *Int J Epidemiol*, 48(3):702–712.

Staiger, D. and Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586.

Stock, J. H., Wright, J. H., and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20(4):518–529.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *Plos Medicine*, 12(3):e1001779.

Sun, D., Zhou, T., Heianza, Y., Li, X., Fan, M., Fonseca, V. A., and Qi, L. (2019). Type 2 diabetes and hypertension: a study on bidirectional causality. *Circulation research*, 124(6):930–937.

Tchetgen Tchetgen, E. J., Sun, B., and Walter, S. (2020). The genius approach to robust mendelian randomization inference. *Statistical Science*.

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.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91(433):217–221.

Wang, H., Zhang, F., Zeng, J., Wu, Y., Kemper, K. E., Xue, A., Zhang, M., Powell, J. E., Goddard, M. E., Wray, N. R., Visscher, P. M., McRae, A. F., and Yang, J. (2019). Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the uk biobank. *Science Advances*, 5:eaaw3538.

Wang, L. and Tchetgen Tchetgen, E. (2018). Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):531–550.

Ye, T., Shao, J., and Kang, H. (2019). Debiased inverse-variance weighted estimator in two-sample summary-data mendelian randomization. arXiv:1911.09802.

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

Zhao, Q., Wang, J., Hemani, G., Bowden, J., and Small, D. S. (2019b). Statistical inference in two-sample summary-data mendelian randomization using robust adjusted profile score. *Annals of Statistics*.

Supplementary Materials

Proof of Theorem 1

Proof. Under assumptions (B1) - (B3), we have

$$D(Z=z) = E(Y \mid A=1, Z=z) - E(Y \mid A=0, Z=z) = \beta + \gamma\sigma^2(z). \tag{S1}$$

Hence, we have the following two equations

$$\begin{aligned} D(Z=1) &= \beta + \gamma\sigma^2(Z=1) \\ D(Z=0) &= \beta + \gamma\sigma^2(Z=0). \end{aligned}$$

Solving these two equations simultaneously gives the result in Theorem 1. □

Estimation and Inference under Weak Identification

Consider the following normal model also given in the main text

$$\frac{Y - \{\beta A + \gamma A \exp(\eta_0 + \eta_z Z) + \theta_0 + \theta_z Z\}}{\sqrt{\exp(\eta_0 + \eta_z Z)}} \sim N(0,1). \tag{S2}$$

Let

$$I_n(\Theta) = -E\left\{\frac{\partial^2}{\partial\Theta\partial\Theta^T} l_n(\Theta)\right\}$$

be the Fisher information matrix for a sample of size n, $\lambda_{\min}\{I_n(\Theta)\}$ be the minimum eigenvalue of $I_n(\Theta)$, $tr(A)$ be the trace of a square matrix A.

Condition 2 summaries the regularity conditions.

Condition 2 (Regularity Conditions)**.** *For every $o = (y, a, z)$ in the support of $O = (Y, A, Z)$, the likelihood function denoted by $f(\Theta; o)$ is twice continuously differentiable in Θ and satisfies*

$$\frac{\partial}{\partial\Theta}\int \psi_\Theta(o)d\nu = \int \frac{\partial}{\partial\Theta}\psi_\Theta(o)d\nu$$

for $\psi_\Theta(o) = f_\Theta(o)$ and $= \partial f_\Theta(o)/\partial\Theta$, where ν is a suitable measure; there is a positive ϵ such that the Fisher information matrix $I_n(\Gamma)$ is positive definite for $\Gamma : \|\Gamma - \Theta\| < \epsilon$; and for any given Θ, there exists a positive number ϵ and a positive integrable function h, such that

$$\sup_{\Gamma:\|\Gamma-\Theta\|<\epsilon} \left\| \frac{\partial^2 \log f_\Gamma(o)}{\partial\Gamma\partial\Gamma^T} \right\| \le h(o)$$

for all o in the support of O, where $\|A\| = \sqrt{tr(A^T A)}$ for any matrix A.

Theorem 3 establishes the consistency and asymptotic normality of solutions to the score function.

Theorem 3. *(a) Under Condition 2 and assume that the observed data $O_i, i = 1, \ldots, n$ are independent and identically distributed, if $k/[\lambda_{\min}\{nI_1(\Theta)\}] \to 0$, then the sequence of conditional maximum likelihood estimators $\{\hat{\Theta}_n\}$ is consistent, i.e., $\hat{\Theta}_n \xrightarrow{p} \Theta$.*
(b) Any consistent sequence $\tilde{\Theta}_n$ such that $S_n(\tilde{\Theta}) = 0$ is asymptotically normal, and as $n \to \infty$,

$$\sqrt{n}(\tilde{\Theta} - \Theta) \xrightarrow{d} N(0, \{I_1(\Theta)\}^{-1}). \tag{S3}$$

In fact, the general consistency condition in Theorem 3 can be written as $\lambda_{\min}\{nI_1(\Theta)\} \to \infty$ and

$$\frac{tr\left[I_n^{-1}(\Theta)E\{S_n(\Theta)S_n(\Theta)^T\}\right]}{\lambda_{\min}\{nI_1(\Theta)\}} \to 0 \tag{S4}$$

for general likelihood-based methods, which, for example, applies to the profile-likelihood estimator in Zhao et al. (2019b). Specifically, in Zhao et al. (2019b), in their notations, $k = 1$, $E\{S_n(\Theta)S_n(\Theta)^T\} = V_1$, $I_n(\Theta) = V_2$, and $V_1 = V_2 + O(p), V_2 \asymp n\|\gamma\|^2$, and thus, condition (S4) becomes $V_1/V_2^2 = O\{p/(n^2\|\gamma\|^4)\} \to 0$, which is the condition for consistency in Zhao et al. (2019b, Theorem 3.1).

Proof. First, notice that the condition $k/[\lambda_{\min}\{nI_1(\Theta)\}] \to 0$ implies $\lambda_{\min}\{nI_1(\Theta)\} \to \infty$. Let $B_n(c) = \{\Gamma : \|[I_n(\Theta)]^{1/2}(\Gamma - \Theta)\| \le c\}$ for $c > 0$, where $\|A\| = \sqrt{tr(A^TA)}$ for any matrix A. Since $B_n(c)$ shrinks to Θ as $n \to \infty$, the consistency result in Theorem 3(a) is implied by the fact that for any $\epsilon > 0$, there exists $c > 0$ and $n_0 > 1$ such that

$$P(l_n(\Gamma) - l_n(\Theta) < 0 \text{ for all } \Gamma \in \partial B_n(c)) \ge 1 - \epsilon, \qquad n \ge n_0 \tag{S5}$$

and $\hat{\Theta}_n$ is unique for $n \ge n_0$, where $\partial B_n(c)$ is the boundary of $B_n(c)$. For $\Gamma \in \partial B_n(c)$, the Taylor expansion gives

$$\begin{aligned} l_n(\Gamma) - l_n(\Theta) = {} & c\lambda^T[I_n(\Theta)]^{-1/2}S_n(\Theta) \\ & + (c^2/2)\lambda^T[I_n(\Theta)]^{-1/2}i_n(\Gamma^*)[I_n(\Theta)]^{-1/2}\lambda, \end{aligned}$$

where $\lambda = [I_n(\Theta)]^{1/2}(\Gamma - \Theta)/c$ satisfying $\|\lambda\| = 1$, $i_n(\Gamma) = \partial^2 l_n(\Gamma)/\partial\Gamma\partial\Gamma^T$, and Γ^* lies between Γ and Θ. Note that

$$\begin{aligned} E\frac{\|i_n(\Gamma^*) - i_n(\Theta)\|}{n} & \le E \max_{\Gamma \in B_n(c)} \frac{\|i_n(\Gamma) - i_n(\Theta)\|}{n} \\ & \le E \max_{\Gamma \in B_n(c)} \|i_1(\Gamma) - i_1(\Theta)\| \to 0 \end{aligned} \tag{S6}$$

which follows from the dominated convergence theorem combined with the facts that (a) $i_1(\Gamma) = \partial l_1(\Gamma)/\partial\Gamma\partial\Gamma^T$ is continuous in a neighborhood of Θ for any fixed observation; (b) $B_n(c)$ shrinks to $\{\Theta\}$ from $\lambda_{\min}(I_n(\Theta)) \to \infty$; and (c) for sufficiently large n, $\max_{\Gamma \in B_n(c)} \|i_1(\Gamma) - i_1(\Theta)\|$ is bounded by an integrate function under Condition 2. By the strong law of large numbers, $\|n^{-1}i_n(\Theta) + I_1(\Theta)\| \xrightarrow{a.s.} 0$. These results imply that

$$l_n(\Gamma) - l_n(\Theta) = c\lambda^T[I_n(\Theta)]^{-1/2}S_n(\Theta) - (1 + o_p(1))c^2/2. \tag{S7}$$

Note that $\max_\lambda\{\lambda^T[I_n(\Theta)]^{-1/2}S_n(\Theta)\} = \|I_n(\Theta)^{-1/2}S_n(\Theta)\|$ with

$$\lambda = I_n(\Theta)^{-1/2}S_n(\Theta)/\|I_n(\Theta)^{-1/2}S_n(\Theta)\|.$$

Therefore, (S5) follows from (S7) and

$$\begin{aligned} & P(c\lambda^T[I_n(\Theta)]^{-1/2}S_n(\Theta) - c^2/2 < 0 \text{ for all } \lambda \ s.t. \ \|\lambda\| = 1) \\ =& P\left(\max_\lambda \lambda^T[I_n(\Theta)]^{-1/2}S_n(\Theta) < c/2\right) \\ =& P\left(\|I_n(\Theta)^{-1/2}S_n(\Theta)\| < c/2\right) \\ =& 1 - P\left(\|I_n(\Theta)^{-1/2}S_n(\Theta)\| \ge c/2\right) \\ \ge& 1 - \frac{4E\|I_n(\Theta)^{-1/2}S_n(\Theta)\|^2}{c^2/4} \qquad \text{(from the Markov inequality)} \\ =& 1 - \frac{E\ tr\{S_n(\Theta)^TI_n(\Theta)^{-1}S_n(\Theta)\}}{c^2/4} \\ =& 1 - \frac{tr\{I_n(\Theta)^{-1}E[S_n(\Theta)S_n^T(\Theta)]\}}{c^2/4} \\ \ge& 1 - \epsilon, \end{aligned}$$

from choosing c such that $c^2/tr\{I_n(\Theta)^{-1}E[S_n(\Theta)S_n^T(\Theta)]\}$ is large enough, which is possible because

$$\begin{aligned}
&c^2/tr\{I_n(\Theta)^{-1}E[S_n(\Theta)S_n^T(\Theta)]\}\\
\geq&\|[I_n(\Theta)]^{1/2}(\Gamma-\Theta)\|^2/tr\{I_n(\Theta)^{-1}E[S_n(\Theta)S_n^T(\Theta)]\}\\
\geq&\lambda_{\min}[I_n(\Theta)]\|\Gamma-\Theta\|^2/tr\{I_n(\Theta)^{-1}E[S_n(\Theta)S_n^T(\Theta)]\}\\
=&\|\Gamma-\Theta\|^2/o(1).
\end{aligned}$$

For $n \geq n_0$, $\hat{\Theta}_n$ is unique because from Condition 2, the Fisher information $I_n(\Gamma)$ is positive definite in a neighborhood of Θ.

(b) Using the mean value theorem for vector-valued functions, we obtain that

$$-S_n(\Theta) = \left[\int_0^1 i_n\big(\Theta + t(\tilde{\Theta}-\Theta)\big)dt\right](\hat{\Theta}-\Theta).$$

Note that using a similar argument as in (S6), when $\hat{\Theta}-\Theta = o_p(1)$,

$$\frac{1}{n}\left\|i_n\big(\Theta + t(\tilde{\Theta}-\Theta)\big) - i_n(\Theta)\right\| \xrightarrow{p} 0.$$

Since $n^{-1}i_n(\Theta) \xrightarrow{p} -I_1(\Theta)$ and $I_n(\Theta) = nI_1(\Theta)$,

$$-S_n(\Theta) = -I_n(\Theta)(\tilde{\Theta}-\Theta) + o_p(\|I_n(\Theta)(\tilde{\Theta}-\Theta)\|).$$

This and Slutsky's theorem imply that $\sqrt{n}(\tilde{\Theta}-\Theta)$ is asymptotically equivalent with

$$\sqrt{n}[I_n(\Theta)]^{-1}S_n(\Theta) = n^{-1/2}[I_1(\Theta)]^{-1}S_n(\Theta) \xrightarrow{d} N_k(0, I[_1(\Theta)]^{-1})$$

by the central limit theorem for i.i.d. samples. □

Empirical guidelines for asymptotics

In practice, it is important to know what value of $\hat{\kappa}$ would be considered large enough for the CMLE to perform well, i.e., to have small bias and adequate coverage probability, in presence of many weak invalid IVs.

We follow the same setting as the second simulation study in Section 3 of the main article. We consider two scenarios: (i) $n = 10^5, p = 20$ in Figure 3; and (ii) $n = 10^6, p = 5$ in Figure 4. For each of the 1000 simulation runs, we plot $\hat{\beta}$ against $\hat{\kappa}$. We also plot the two standard error bands centered at β_0 (shaded area). From Figures 3-4, $\hat{\kappa}$ are all larger than 10, and $\hat{\beta}$ are inside the shaded area, close to the true value $\beta = 0.8$. Note that because $\hat{\kappa}$ (which relates to the Hessian matrix) is a function of $\hat{\beta}$, we see that $\hat{\beta}$ is not symmetric around the true value $\beta = 0.8$. But this usually does not affect using $\hat{\kappa}$ as a condition to evaluate whether the CMLE would perform well. We have also simulated extensively under other settings, we find that the CMLE has good performance when $\hat{\kappa} \geq 10$.

Estimation method under Gaussian mixture error

We describe the alternating optimization procedure with user-specified K and parametric models $\sigma(Z = z; \eta)$ and $E(Y|A = 0, Z = z; \theta) \equiv \mu_0(z; \theta)$ indexed by finite-dimensional parameters η and θ respectively.

(i) Initialization step: maximize the standard Gaussian location scale model

$$\frac{Y - \{\beta A + \gamma A\sigma^2(Z;\eta) + \mu_0(Z;\theta)\}}{\sigma(Z;\eta)} \sim N(0,1)$$

with respect to $(\beta, \theta, \eta, \gamma)$ by maximum likelihood estimation, to obtain $(\tilde{\beta}, \tilde{\theta}, \tilde{\eta}, \tilde{\gamma})$. Construct

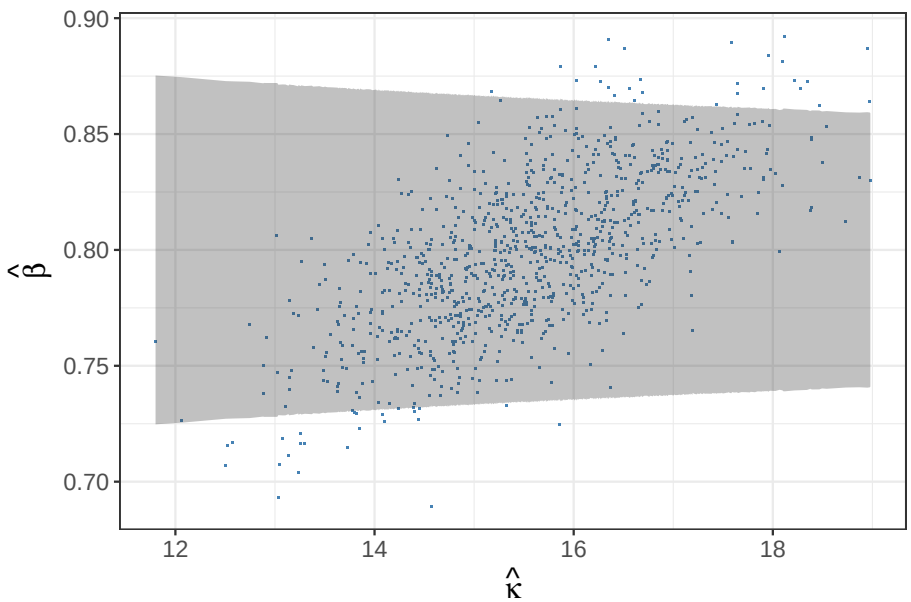

Figure 3: Evaluation of the consistency and asymptotic normality condition for the CMLE under the same setting as the second simulation study in Section 3 of the main article, with $n = 10^5, p = 20$. The x-axis plots the condition $\hat{\kappa}$, and the y-axis plots values of the CMLE in 1000 simulations. The shaded area represents the two-sided error bands centered at $\beta = 0.8$.

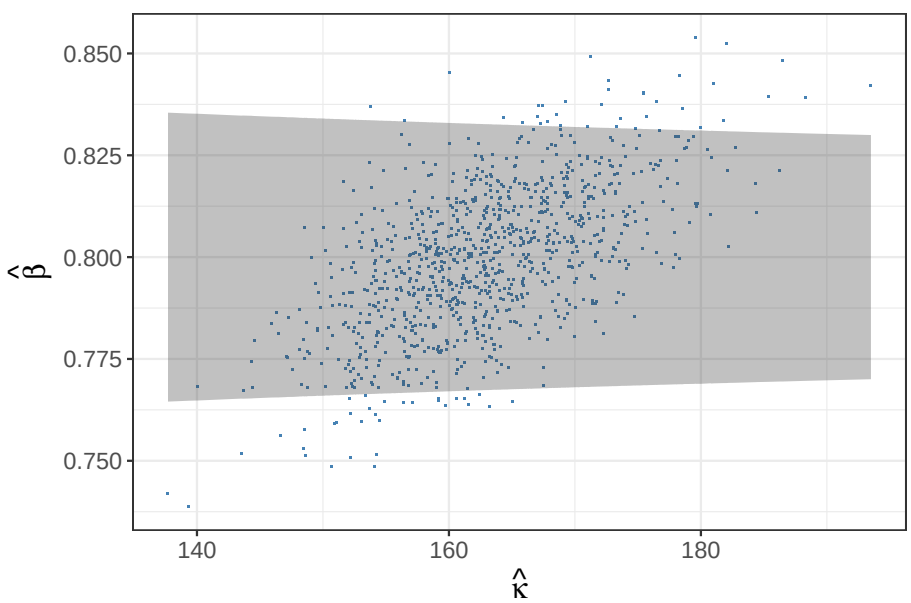

Figure 4: Evaluation of the consistency and asymptotic normality condition for the CMLE under the same setting as the second simulation study in Section 3 of the main article, with $n = 10^6, p = 5$. The x-axis plots the condition $\hat{\kappa}$, and the y-axis plots values of the CMLE in 1000 simulations. The shaded area represents the two-sided error bands centered at $\beta = 0.8$.

the standardized residuals

$$\tilde{\epsilon} = \frac{Y - \{\tilde{\beta}A + \tilde{\gamma}A\sigma^2(Z;\tilde{\eta}) + \mu_0(Z;\tilde{\theta})\}}{\sigma(Z;\tilde{\eta})}.$$

(ii) Maximize the Gaussian mixture location scale model

$$\tilde{\epsilon} \sim \sum_{k=1}^{K} \pi_k N\left(\mu_k, \delta_k^2\right)$$

with respect to $(\pi_1, ..., \pi_K, \mu_1, ..., \mu_K, \delta_1, ..., \delta_K)$ using constrained maximum likelihood estimation to obtain $(\tilde{\pi}_1, ..., \tilde{\pi}_K, \tilde{\mu}_1, ..., \tilde{\mu}_K, \tilde{\delta}_1, ..., \tilde{\delta}_K)$. Construct

$$\tilde{v} = A\sum_{k=1}^{K} \frac{\tilde{\pi}_k\tilde{\omega}_k\tilde{\delta}_k^2\sigma^2(Z;\tilde{\eta})}{\sum_{k=1}^{K}\tilde{\pi}_k\tilde{\omega}_k}, \quad \tilde{w} = \frac{\sigma(Z;\tilde{\eta})}{\sum_{k=1}^{K}\tilde{\pi}_k\tilde{\omega}_k}\left\{\left[\sum_{k=1}^{K}\tilde{\pi}_k\tilde{\omega}_k\tilde{\mu}_k\right]\right\},$$

where

$$\tilde{\omega}_k = \exp\{\tilde{\gamma}A\sigma(Z;\tilde{\eta})\tilde{\mu}_k + \tilde{\delta}_k^2\left[\tilde{\gamma}A\sigma(Z;\tilde{\eta})\right]^2/2\}.$$

(iii) Minimize $n^{-1}\sum_i\{Y_i - \beta A_i - \mu_0(Z_i;\theta) - \gamma\tilde{v}_i - \tilde{w}_i\}^2$ using the least squares method with offset $\tilde{w}$ to obtain new $(\tilde{\beta}, \tilde{\theta}, \tilde{\gamma})$, followed by regressing the squared residuals to obtain new $\tilde{\eta}$. Construct the standardized residuals

$$\tilde{\epsilon} = \frac{Y - \{\tilde{\beta}A + \mu_0(Z;\tilde{\theta}) + \tilde{\gamma}\tilde{v} + \tilde{w}\}}{\sigma(Z;\tilde{\eta})},$$

based on the new values $(\tilde{\beta}, \tilde{\theta}, \tilde{\gamma}, \tilde{\eta})$.

(iv) Iterate between steps (ii) and (iii) until change in log-likelihood derived at step (ii) falls below a user-specified tolerance level. In this simulation we iterate until $|\log LL_j - \log LL_{j-1}|/|\log LL_{j-1}| < 0.001$ at the j-th iteration.

Asymptotic variance is estimated based on standard M-estimation theory by stacking the estimating functions in steps (ii) and (iii), evaluated at the final parameter values at convergence.

Semiparametric three-stage estimation

Consider the general location-scale model

$$Y = E(Y|A, Z) + \sigma(Z)\epsilon \equiv \mu(A, Z) + \sigma(Z)\epsilon,$$

where $\epsilon \perp A, Z$. Under assumptions (B1)-(B3), the conditional mean function is given by

$$\begin{aligned}\mu(a, z) &= \beta a + \mu(0, z) + \frac{\partial}{\partial(\gamma a)}\ln E\{\exp(\gamma a\sigma(z)\epsilon)|A = 0, Z = z\} \\ &= \beta a + \mu(0, z) + \sigma(z)\frac{E\{\epsilon\exp(\gamma a\sigma(z)\epsilon)\}}{E\{\exp(\gamma a\sigma(z)\epsilon)\}}.\end{aligned}$$

A semiparametric three-stage estimator of (β, γ) may be implemented via the following steps:

(i) Fit the regression model $\mu(a, z) = m_a(a) + m_z(z) + m_{az}(a, z)$, where $0 = m_z(0) = m_{az}(a, 0) = m_{az}(0, z)$ for identification, using a nonparametric method such as generalized additive model. For example, if the support of Z is $\{0, 1, 2\}$, then the nonparametric model for the outcome is of the form $\mu(a, z) = m_a(a) + \{\gamma_1 z + \gamma_2 z^2\} + \{zm_1(a) + z^2 m_2(a)\}$ with $m_1(0) = m_2(0) = 0$; a saturated model may be considered if A also has finite support. Let $\widehat{\mu}(a, z)$ denote the resulting estimator of the mean $\mu(a, z)$.

(ii) Using the residuals from (i), fit a nonparametric model for the conditional variance $\sigma^2(z)$. If the support of Z is $\{0, 1, 2\}$, then we can specify the saturated model $\sigma^2(z) = \exp(\eta_0 + \eta_1 z +$

$\eta_2 z^2$). Let $\widehat{\sigma}^2(z)$ denote the resulting estimator of $\sigma^2(z)$.

(iii) Define $\widehat{\epsilon}_j = \{Y_j - \widehat{\mu}(A_j, Z_j)\}/\widehat{\sigma}(Z_j)$ and let

$$\widehat{g}_i(\gamma) = \frac{\sum_{j=1}^n \widehat{\epsilon}_j \exp(\gamma A_i \widehat{\sigma}(Z_i)\widehat{\epsilon}_j)}{\sum_{j=1}^n \exp(\gamma A_i \widehat{\sigma}(Z_i)\widehat{\epsilon}_j)}, \quad i = 1, ..., n.$$

The proposed semiparametric three-stage estimator of (β, γ) is

$$(\hat{\beta}, \hat{\gamma}) = \arg\min_{\beta,\gamma} \sum_i \{Y_i - \beta A_i - \widehat{\mu}(0, Z_i) - \widehat{\sigma}(Z_i)\widehat{g}_i(\gamma)\}^2.$$

We note that a potentially more efficient approach is to maximize the log-likelihood for the model using a kernel estimator for the density of standardized residuals from step (iii). Specifically, let

$$\widehat{\delta}_i(\beta, \gamma) = \{Y_j - \beta A_i - \widehat{\mu}(0, Z_i) - \widehat{\sigma}(Z_i)\widehat{g}_i(\gamma)\}/\widehat{\sigma}(Z_i)$$

and define

$$\ell_n(\beta, \gamma) = \sum_{i=1}^n \log \widehat{f}_h(\widehat{\delta}_i(\beta, \gamma)),$$

where $\widehat{f}_h(\epsilon) = \frac{1}{nh}\sum_{i=1}^n K\left(\frac{\epsilon - \widehat{\epsilon}_i}{h}\right)$ is a kernel density estimator with bandwidth $h > 0$. The alternative estimator of (β, γ) is given by

$$(\hat{\beta}, \hat{\gamma}) = \arg\max_{\beta,\gamma} \{\ell_n(\beta, \gamma)\}.$$