

Winner's Curse Free Robust Mendelian Randomization with Summary Data

Zhongming Xie* Wanheng Zhang† Jingshen Wang‡ Chong Wu§

August 26, 2025

Abstract

In the past decade, the increased availability of genome-wide association studies summary data has popularized Mendelian Randomization (MR) for conducting causal inference. MR analyses, incorporating genetic variants as instrumental variables, are known for their robustness against reverse causation bias and unmeasured confounders. Nevertheless, classical MR analyses utilizing summary data may still produce biased causal effect estimates due to the winner's curse and pleiotropy issues. To address these two issues and establish valid causal conclusions, we propose a unified robust Mendelian Randomization framework with summary data, which systematically removes the winner's curse and screens out invalid genetic instruments with pleiotropic effects. Unlike existing robust MR literature, our framework delivers valid statistical inference on the causal effect without requiring the genetic pleiotropy effects to follow any parametric distribution or relying on perfect instrument screening property. Under appropriate conditions, we demonstrate that our proposed estimator converges to a normal distribution, and its variance can be well estimated. We demonstrate the performance of our proposed estimator through Monte Carlo simulations and two case studies.

Keywords: Bootstrap aggregation; GWAS; Post-selection inference.

1 Introduction

1.1 Background and motivation

Drawing inferences about cause and effect lies at the core of uncovering essential scientific principles. In biological and biomedical sciences, causal inference deepens our understanding of underlying etiology and advances developments in disease diagnosis, treatment, and prevention. While

*Division of Biostatistics, University of California Berkeley.

†Department of Biostatistics, The University of Texas MD Anderson Cancer Center.

‡Division of Biostatistics, University of California Berkeley. Corresponding author.

§Department of Biostatistics, The University of Texas MD Anderson Cancer Center. Corresponding author.

observational data present unique opportunities for causal inference by employing large and rich datasets, causal discoveries from observational studies are often susceptible to unmeasured confounding and reverse causation bias issues [26, 15, 17, 46]. As a remedy, Mendelian Randomization (MR) has become a popular research design. Its popularity is not only ascribed to the fact that MR mitigates unmeasured confounding bias by using genetic variants as instrumental variables (IVs) to assess the causal relationship between exposures and outcomes but also credited to the increasing availability of large-scale genome-wide association studies (GWAS) summary data on various complex traits [46, 11, 29, 45].

However, MR with GWAS summary may still produce biased estimates of causal effects due to several sources of bias. These include measurement error in exposure GWAS, winner’s curse bias resulting from using the same exposure GWAS for both IV selection and effect estimation, and most crucially, bias from including invalid IVs with pleiotropy [42]. Firstly, the effect of IV on exposure is measured by exposure GWAS, which inherently contains measurement error. Ignoring such measurement error can produce biased causal effect estimates, especially when the strength of IVs is weak [54, 33]. Secondly, the practice of selecting genetic instruments based on their estimated associations with the exposure variable from GWAS, and using the same data for both instrument selection and estimation, can lead to biased causal effect estimates due to the winner’s curse phenomenon [58, 57, 18]. Lastly, typical MR analyses inevitably involve some invalid IVs that either directly affect the outcome or through unmeasured confounding factors—a phenomenon known as pleiotropy [23, 52]. The nature of pleiotropy is widespread and usually unknown or complex [52]. Failure to fully account for pleiotropy will also lead to biased causal effect estimates.

A broad literature addresses the biases discussed above to improve the credibility of MR analyses, yet no single approach can simultaneously tackle all these biases. Some methods have made progress in addressing individual issues. For instance, [54] formally tackled the measurement error bias in the popular inverse variance weighted estimator, while [33] proposed a randomized instrument selection and Rao-Blackwellization procedure to address both measurement error bias and winner’s curse bias. However, the validity of these methods relies heavily on the assumption that all IVs either have no pleiotropic effects or exhibit balanced pleiotropic effects—an assumption unlikely to hold in practice due to the unknown and complex nature of pleiotropy [52], potentially

leading to biased causal effect estimates.

To account for widespread pleiotropy, many robust MR methods have been proposed. These methods primarily focus on addressing the issue raised by invalid IVs, but often at the expense of neglecting measurement error and winner’s curse biases. They can be broadly categorized into two strategies. The first strategy imposes normal mixture model assumptions on the pleiotropic effects. By modeling the observed GWAS summary data within a joint likelihood function, these methods simultaneously estimate the unknown parameters and the desired causal effect. Such methods include RAPS [56], ContMix [9], MR-APSS [25], MRMix [38]. However, as demonstrated in our simulation studies, when the normal mixture model assumption is violated, these approaches tend to produce false positive findings or have low detection power. Moreover, incorporating procedures to address winner’s curse bias, such as that proposed by [33], is challenging within this framework as it may violate parametric modeling assumptions and result in an incorrect likelihood function. The second strategy avoids imposing parametric modeling assumptions on the pleiotropic effects. Instead, it adopts penalization methods to screen out invalid instruments with pleiotropic effects, using only the selected valid instruments for causal effect estimation. Such methods include, for example, cML [53] and MR-Lasso [31]. However, these methods either lack rigorous statistical justifications or require that the selected IVs are valid and include all valid IVs (a condition we refer to as “perfect IV screening”). For example, [53] prove that their procedure can screen out all invalid IVs with a probability tending to one under the asymptotic regime where the number of IVs is fixed, and the sample size tends to infinity. When this is achieved, the resulting causal effect estimate is consistent and asymptotically normal. However, the theoretical results under this asymptotic regime do not account for how the magnitudes of the pleiotropic effects impact the validity of statistical inference. In fact, perfect IV screening is often unattainable when the pleiotropic effects are small, and the differences between valid and invalid IVs in MR studies are subtle. Notably, two-sample MR is a rapidly evolving field with numerous methodological advancements, such as [35, 30, 19]. For comprehensive reviews of statistical methods in MR, we refer readers to [43] and [1].

1.2 Contribution

To bridge the aforementioned gaps in the existing literature, we propose a unified MR framework with summary data that simultaneously addresses winner’s curse bias, bias from measurement error in exposure GWAS, and bias from invalid IVs with pleiotropy (Section 3). Specifically, we propose an l_0 constrained optimization framework that can simultaneously screen out invalid IVs, account for measurement error, and seamlessly integrate with the winner’s removal step from [33]. Moreover, we demonstrate that the proposed l_0 constrained optimization framework maintains computational efficiency due to the special form of our objective function. Furthermore, to improve statistical efficiency, we adopt a bootstrap aggregation procedure and use a non-parametric delta method to perform valid inference on the final causal effect.

On the theoretical side, we provide comprehensive theoretical investigations of the proposed method in Section 4. We prove that the final estimator in our proposed method is asymptotically unbiased and converges to a normal distribution even in the presence of directional pleiotropy. Moreover, different from existing theoretical analyses in robust MR, we show that our method can deliver consistent causal effect estimates without perfect invalid IV screening; see detailed discussion in Supplementary Material Section S.6. In brief, our theoretical investigation indicates that our proposed method can screen out IVs with large pleiotropic effects, and the resulting causal effect estimator remains consistent even if the selected IVs include some invalid ones with small pleiotropic effects. These theoretical investigations better characterize scenarios where our method performs well and demonstrate its robustness.

Benefiting from the above features in both methodological and theoretical aspects, we demonstrate that our proposed MR framework delivers robust causal effect estimates with improved statistical power in simulated Monte Carlo experiments (Section 5) and in two case studies (Section 6). From our simulated Monte Carlo experiments, we confirm that our proposed method outperforms benchmark methods in terms of type 1 error rates, power, absolute bias, mean squared error, and coverage probability in most scenarios. The results also highlight the importance of simultaneously correcting the winner’s curse bias and accounting for measurement error bias and generic pleiotropic effects. From our case study of negative control outcome analyses, in which the population causal effects are believed to be zero by design, we confirm that our approach yields well-controlled Type I

error rates (Section 6.1). From our case study to identify causal risk factors for COVID-19 severity, our approach identifies more causal risk factors than the existing approaches, and the identified causal exposures by our proposed method have more supporting evidence.

2 Framework and challenges

In this section, we review the classical two-sample Mendelian Randomization (MR) framework with summary data. We then revisit the pleiotropic effects, measurement error bias, and winner’s curse bias within this framework.

Referring to the causal diagram in Figure 1, we let X denote the exposure, Y the outcome, and U the unmeasured confounder between the exposure and the outcome. The goal of MR analysis is to estimate the causal effect (denoted by θ) of the exposure variable X on the outcome variable Y . However, in the presence of unmeasured confounder U , it is challenging to directly estimate θ solely using the information stored in X and Y . To overcome this, two-sample MR analyses incorporate p mutually independent SNPs G_1, \dots, G_p as instrumental variables (IVs) and estimate θ using the estimated association pairs $\{(\hat{\beta}_{X_j}, \hat{\beta}_{Y_j})\}_{j=1}^p$ collected from two independent GWAS datasets, where $\hat{\beta}_{X_j}$ and $\hat{\beta}_{Y_j}$ are the estimated effect sizes for IV j in exposure and outcome GWAS, respectively. Here, genetic variant $G_j \in \{0, 1, 2\}$ represents the number of effect alleles of a single-nucleotide polymorphism (SNP) j inherited by an individual. Following the two-sample summary-data MR literature [54, 56], we assume the following linear structural equation model:

$$\begin{aligned} U &= \sum_{j=1}^p \phi_j G_j + E_U, \\ X &= \sum_{j=1}^p \gamma_j G_j + \beta_{XU} U + E_X, \\ Y &= \sum_{j=1}^p \alpha_j G_j + \beta_{YU} U + \theta X + E_Y, \end{aligned} \tag{1}$$

where E_U , E_X , and E_Y are mutually independent random noises. E_U is independent of (G_1, \dots, G_p) , and E_X and E_Y are independent of (G_1, \dots, G_p, U) . To allow for the valid inference of the causal effect θ , we need G_j ($j = 1, \dots, p$) to be valid IVs in the sense that they satisfy the following three conditions: (1) $\gamma_j \neq 0$, meaning that G_j is associated with X (relevance assumption); (2)

$\phi_j = 0$, meaning that G_j has no correlated pleiotropic effect with Y (effective random assignment assumption); (3) $\alpha_j = 0$, meaning that G_j has no uncorrelated pleiotropic effect with Y (exclusion restriction assumption).

Provided that all included genetic IVs are valid, two-sample MR analyses can deliver valid inference on θ by appropriately using information stored in two independent GWAS datasets. To provide some justifications for this claim, we follow the causal model proposed in [37]. In particular, in the structural equation models given in Eq (1), the total effect of SNP G_j on Y and the total effect of G_j on X are given by:

$$\begin{aligned}\beta_{Y_j} &= \mathbb{E}[Y|do(G_j = g_j + 1)] - \mathbb{E}[Y|do(G_j = g_j)] = \alpha_j + \beta_{YU}\phi_j + \theta \cdot (\gamma_j + \beta_{XU}\phi_j), \\ \beta_{X_j} &= \mathbb{E}[X|do(G_j = g_j + 1)] - \mathbb{E}[X|do(G_j = g_j)] = \gamma_j + \beta_{XU}\phi_j.\end{aligned}$$

For a valid IV G_j , when G_j satisfies $\phi_j = 0$ (effective random assignment assumption) and $\alpha_j = 0$ (exclusion restriction assumption), the target causal effect θ will satisfy $\beta_{Y_j} = \theta\beta_{X_j}$, where $\beta_{X_j} = \gamma_j$ and $\beta_{Y_j} = \theta\gamma_j$. If the relevance assumption $\gamma_j \neq 0$ is also met, we are then able to use β_{Y_j} and β_{X_j} to assist valid inference on θ , as they can be well estimated through the estimated association pairs $\{(\hat{\beta}_{X_j}, \hat{\beta}_{Y_j})\}_{j=1}^p$ collected from two independent GWAS dataset in two-sample summary-data MR framework.

However, in practice, due to the widespread pleiotropy in human genetics [23, 52], the effective random assignment ($\phi_j = 0$) and exclusion restriction assumptions ($\alpha_j = 0$) are frequently violated, leading to invalid IVs. In the presence of invalid IVs, the total effect of G_j on Y can be expressed as:

$$\beta_{Y_j} = \underbrace{\theta \cdot \beta_{X_j}}_{\text{causal effect}} + \underbrace{\alpha_j}_{\text{uncorrelated pleiotropy}} + \underbrace{\beta_{YU} \cdot \phi_j}_{\text{correlated pleiotropy}} \equiv \theta \cdot \beta_{X_j} + r_j. \quad (2)$$

Here, α_j is the uncorrelated pleiotropic effect that captures the direct effect of G_j on Y , and $\beta_{YU} \cdot \phi_j$ is the correlated pleiotropic effect that captures the effect of G_j on Y through the pathway $G_j \rightarrow U \rightarrow Y$. Their combined effect, $r_j = \alpha_j + \beta_{YU} \cdot \phi_j$, represents the total effect of a genetic variant G_j on the outcome Y induced by pleiotropy. These violations make it challenging to accurately estimate θ using MR. If not appropriately accounted for, genetic pleiotropy can result

in biased causal effect estimates in MR analyses (see Section 5 for our simulation results).

On top of the potential bias induced by pleiotropic effects, two additional sources of bias in MR analyses are measurement error bias and winner’s curse bias. Measurement error bias arises from the fact that the true effect of an IV on the exposure, β_{X_j} , is unobserved. Instead, we rely on $\hat{\beta}_{X_j}$, an estimate derived from exposure GWAS (I), which inherently contains measurement error, to conduct MR. The winner’s curse bias, on the other hand, is induced by pre-selecting IVs that are strongly associated with the exposure variable to meet the relevance assumption (that is, $\gamma_j \neq 0$). This selection exercise is often based on hard-thresholding measured SNP z -scores obtained from GWAS (I): SNP j is selected if $|\hat{\beta}_{X_j}/\sigma_{X_j}| > \lambda$, where λ is a pre-specified cut-off value, and $\hat{\beta}_{X_j}$ and σ_{X_j} are estimated effect size and its standard error from exposure GWAS dataset, respectively. The selected IVs are then used to construct downstream causal effect estimators. The selected IV-exposure associations tend to overestimate the underlying true association effects β_{X_j} , as the distribution of any $\hat{\beta}_{X_j}$ that survives the selection is a truncated Gaussian and the post-selection mean is no longer β_{X_j} when commonly used Gaussian assumption on $\hat{\beta}_{X_j}$ is adopted. Subsequently, by doubly using the data in GWAS (I) for IV selection and estimation, classical MR estimators are expected to be biased and have an intractable limiting distribution, making statistical inference problematic.

In the rest of this manuscript, we employ the following model frequently adopted in the Mendelian Randomization literature [56, 38, 53]:

Assumption 1 (Measurement error model) (i)For any $j \neq j'$, $(\hat{\beta}_{Y_j}, \hat{\beta}_{X_j})$ and $(\hat{\beta}_{Y_{j'}}, \hat{\beta}_{X_{j'}})$ are mutually independent. (ii)For each j , the association pair $(\hat{\beta}_{Y_j}, \hat{\beta}_{X_j})$ follows

$$\begin{bmatrix} \hat{\beta}_{X_j} \\ \hat{\beta}_{Y_j} \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \beta_{X_j} \\ \theta\beta_{X_j} + r_j \end{bmatrix}, \begin{bmatrix} \sigma_{X_j}^2 & 0 \\ 0 & \sigma_{Y_j}^2 \end{bmatrix} \right).$$

Furthermore, there exists a positive integer $n \rightarrow \infty$ and positive constants m and M such that $\frac{m}{n} \leq \sigma_{X_j}^2 \leq \frac{M}{n}$, $\frac{m}{n} \leq \sigma_{Y_j}^2 \leq \frac{M}{n}$ for $j = 1, \dots, p$.

The assumption of independent SNPs, while seemingly stringent, is grounded in established practice in two-sample MR analyses [54, 56, 33]. This approach helps ensure that each selected SNP represents a signal from a unique genetic locus, thereby mitigating potential confounding effects

from LD and facilitating clearer interpretation of causal effect estimates. We acknowledge that alternative cis-MR methods such as Transcriptome-Wide Association Studies (TWAS) [21, 50] and Proteome-Wide Association Studies (PWAS), effectively utilize correlated SNPs, particularly for investigating relationship between omics and complex traits. However, as the reviewer suggested, when inferring causal relationships between complex traits/diseases (such as the two case studies in Section 6), using independent IVs from the whole genome is typically efficient enough and simple to implement. This strategy is also widely adopted in the literature. Therefore, in line with this common practice, we adopt the independence assumption. To ensure independent IVs, we apply a sigma-based LD pruning method [33].

3 Methodology

3.1 Measurement error correction and invalid IV screening

To estimate the causal effect θ , a straightforward approach is to replace the population association effects with their empirical estimates from GWAS in the causal structure equation in (2). Given that all population associations are measured with error in GWAS, the sample analogue of the structure equations can be represented as the following two-stage regression model with measurement errors:

$$\underbrace{\widehat{\beta}_{Y_j}}_{\text{response}} = \underbrace{\theta}_{\text{target parameter}} \cdot \underbrace{\beta_{X_j}}_{\text{true covariate}} + \underbrace{r_j}_{\text{unknown parameter}} + \underbrace{\nu_j}_{\text{noise}}, \quad \underbrace{\widehat{\beta}_{X_j} = \beta_{X_j} + u_j}_{\substack{\text{covariates are} \\ \text{measured with error}}},$$

where ν_j and u_j are centered noises.

To operationalize an accurate estimate of θ using the above two-stage least squares model, we first consider a situation where a set of IVs with $\beta_{X_j} \neq 0$ (denoted as \mathcal{S}) is known. Our method does not require \mathcal{S} to be known, and we will discuss the selection of \mathcal{S} and the practical implementation of our algorithm in the next subsection. With a known \mathcal{S} , we propose estimating θ by solving the

following constrained optimization problem:

$$\begin{aligned} \min_{\theta, r_j} l(\theta, \{r_j\}_{j \in \mathcal{S}}) &= \sum_{j \in \mathcal{S}} l_j(\theta, r_j) \triangleq \sum_{j \in \mathcal{S}} \frac{(\hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_j} - r_j)^2}{\sigma_{Y_j}^2} - \sum_{j \in \mathcal{S}} \frac{\theta^2 \cdot \sigma_{X_j}^2}{\sigma_{Y_j}^2} \mathbb{1}_{(r_j=0)}, \\ \text{s.t. } \sum_{j \in \mathcal{S}} \mathbb{1}_{(r_j=0)} &= v. \end{aligned} \quad (3)$$

Intuitively, the objective function above is a bias-corrected least squares function designed to account for measurement error, subject to the constraint that the adopted IVs for estimating θ are valid. In the following, we will show that the optimization problem above not only accounts for the measurement errors in $\hat{\beta}_{X_j}$ but also accurately identifies invalid IVs with $r_j \neq 0$. This is achieved with computational efficiency, even when an l_0 -type constraint is adopted. As a result, the solution of this optimization problem provides an accurate estimate of θ .

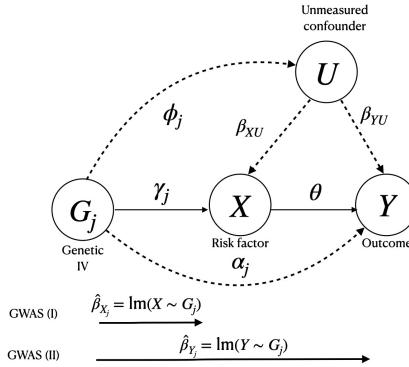


Figure 1: The causal diagram and GWAS (I) and (II) summary data adopted in the two-sample MR. The corresponding causal effect for each pathway is labeled near the directed edge.

To start with, when the set of IVs with $r_j = 0$ is known, the solution of the above optimization problem provides an unbiased estimate of θ . As in this case, we have

$$L(\theta) \triangleq \min_{r_j} l(\theta, \{r_j\}_{j \in \mathcal{S}}) = \frac{1}{2} \sum_{j \in \mathcal{V}} \frac{(\hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_j})^2}{\sigma_{Y_j}^2} - \frac{1}{2} \sum_{j \in \mathcal{V}} \frac{\theta^2 \cdot \sigma_{X_j}^2}{\sigma_{Y_j}^2}.$$

We can verify that $L(\theta)$ is unbiased for the weighted least squares loss function in the sense that $\mathbb{E}[L(\theta)] = \mathbb{E}[\sum_{j \in \mathcal{V}} (\hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_j})^2 / (2\sigma_{Y_j}^2)]$. This suggests that its minimizer is unbiased for the causal effect θ .

Next, as the set of IVs with $r_j = 0$ is unknown, Problem (3) incorporates an l_0 -type constraint

to screen out invalid IVs. While classical l_0 -type optimization problems are solved by their convex relaxations, this technique does not apply to our problem due to the inclusion of a measurement error bias correction term in our objective function (that is, the term $\sum_{j \in \mathcal{S}_\lambda} \theta^2 \cdot \sigma_{X_j}^2 / \sigma_{Y_j}^2 \mathbf{1}_{(r_j=0)}$). To address this issue, we propose an iterative algorithm that mimics block coordinate descent and guarantees the decay of our objective function in Algorithm 3; see justification in the Supplementary Material Section S.1.

Lastly, the number of valid IVs v is unknown and requires tuning. To choose the final set of valid IVs, we propose a generalized Bayesian Information Criteria (GBIC), that is:

$$\text{GBIC}(v) = -2\hat{l}(\hat{\theta}(v), \{\hat{r}_j(v)\}_{j \in \hat{\mathcal{V}}}) + \kappa_n \cdot (s - v), \quad s = |\mathcal{S}|,$$

where $\kappa_n = \log(n)$, and choose the final set of valid IVs by minimizing the GBIC. The proposed GBIC with $\kappa_n = \log(n)$ is different from the classical BIC criteria that adopts $\kappa_n = \log(s_\lambda)$. The reason for this choice is that the classical model selection consistency result of the BIC is established in the asymptotic regime with fixed s_λ . As we are in an asymptotic regime with $s_\lambda \rightarrow \infty$, our proposed GBIC criteria adjusts κ_n accordingly to ensure invalid IV screening consistency. In particular, in Section S.6 of the Supplemental Material, we demonstrate that our procedure provides a consistent causal effect estimator without requiring the perfect IV screening property under a simplified scenario and Conditions 1-2 and 8-9. One of these conditions imposes a constraint on the penalization coefficient κ_n : $\kappa_n \gg \log(s_\lambda)$. We argue that $\kappa_n = \log(n)$ is a feasible choice to satisfy this condition, as the order of the sample size is typically larger than the order of the number of selected relevant IVs in a two-sample MR study.

3.2 Unknown \mathcal{S} and practical implementation

We now consider the realistic scenario where the set \mathcal{S} is unknown. Because the collection of relevant IVs is not known, practitioners typically perform a pre-selection procedure to identify IVs strongly associated with the exposure. These selected IVs are then used to estimate the causal effect. As discussed in Section 2, selecting genetic instruments based on their estimated associations with the exposure variable from GWAS and using the same data for both instrument selection and estimation can lead to biased causal effect estimates due to the winner's curse phenomenon. To

address the issue of winner's curse bias when \mathcal{S} is unknown, we integrate the proposed method from the previous section with the approach described in [33] to perform Rao-Blackwellized randomized instrument selection.

For each SNP $j = 1, 2, \dots, p$, we generate a pseudo SNP-exposure association effect $Z_j \sim \mathcal{N}(0, \eta^2)$, and select SNP j if $\left| \frac{\widehat{\beta}_{X_j}}{\sigma_{X_j}} + Z_j \right| > \lambda$. Define the set of selected SNPs as $\mathcal{S}_\lambda = \left\{ j : \left| \frac{\widehat{\beta}_{X_j}}{\sigma_{X_j}} + Z_j \right| > \lambda, j = 1, 2, \dots, p \right\}$ and its cardinality $|\mathcal{S}_\lambda| = s_\lambda$. For each selected SNP $j \in \mathcal{S}_\lambda$, we construct an unbiased estimator of β_{X_j} as

$$\widehat{\beta}_{X_j, \text{RB}} = \widehat{\beta}_{X_j} - \frac{\sigma_{X_j}}{\eta} \frac{\phi(A_{j,+}) - \phi(A_{j,-})}{1 - \Phi(A_{j,+}) + \Phi(A_{j,-})}, \text{ where } A_{j,\pm} = -\frac{\widehat{\beta}_{X_j}}{\sigma_{X_j} \eta} \pm \frac{\lambda}{\eta},$$

Algorithm 1: Algorithm to solve the optimization problem in (4)

Input: Data inputs and initial parameters

Output: Estimated parameters $\widehat{\theta}$ and \widehat{r}_j

Initialization Set $k = 0$, generate $\theta^{(0)} \sim \text{Uniform} \left(\min_{1 \leq j \leq s_\lambda} \frac{\widehat{\beta}_{Y_j}}{\widehat{\beta}_{X_j}}, \max_{1 \leq j \leq s_\lambda} \frac{\widehat{\beta}_{Y_j}}{\widehat{\beta}_{X_j}} \right)$;

Block Coordinate Descent

repeat

 Fix $\theta^{(k)}$, update $r_j^{(k+1)}$;

 Order $\frac{(\widehat{\beta}_{Y_j} - \theta^{(k)} \cdot \widehat{\beta}_{X_j, \text{RB}})^2}{\sigma_{Y_j}^2} - \frac{\theta^2 \cdot \widehat{\sigma}_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2}$, $j = 1, 2, \dots, s_\lambda - v$ in decreasing order;

 Set $r_j^{(k+1)} = \widehat{\beta}_{Y_j} - \theta^{(k)} \widehat{\beta}_{X_j, \text{RB}}$ for the largest $s_\lambda - v$ components, $j = 1, \dots, s_\lambda - v$, and

$r_j^{(k+1)} = 0$ for $j = s_\lambda - v + 1, \dots, s_\lambda$;

 Fix $r_j^{(k+1)}$, update $\theta^{(k)}$ by minimizing the following objective function:

$$\theta^{(k+1)} = \arg \min_{\theta \in \mathbb{R}} \sum_{j \in \mathcal{S}_\lambda} \frac{\left(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}} - r_j^{(k+1)} \right)^2 - \theta^2 \cdot \widehat{\sigma}_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2} \mathbb{1}_{(r_j^{(k+1)} = 0)}.$$

If $\left| \frac{\theta^{(k+1)} - \theta^{(k)}}{\theta^{(k)}} \right| < 10^{-7}$ **then** Stop and output $\widehat{\theta}(v) = \theta^{(k+1)}$ and $\widehat{r}_j(v) = r_j^{(k+1)}$;

else Set $k = k + 1$;

until $\left| \frac{\theta^{(k+1)} - \theta^{(k)}}{\theta^{(k)}} \right| < 10^{-7}$;

end

Valid IV Selection via GBIC

for $v = 2, \dots, s_\lambda$ **do**

 Calculate

$$\text{GBIC}(v) = -2\widehat{l}\left(\widehat{\theta}(v), \{\widehat{r}_j(v)\}_{j \in \widehat{\mathcal{V}}}\right) + \log(n) \cdot (s_\lambda - v);$$

end

 Select $\widehat{\mathcal{V}}$ with the smallest $\text{GBIC}(v)$;

end

$\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal density and cumulative distribution functions. Here,

η is a pre-specified constant that reflects the noise level of the pseudo SNPs. We recommend using $\eta = 0.5$ as a default value [33]. This choice balances the need for sufficient randomization to address the winner's curse bias while maintaining the stability of the selection process. The above procedure only randomizes the IV selection near the cut-off value λ , which implies that the strong IVs with large β_{X_j} are invariably selected. Here, the choice of the significance cutoff (λ) for selecting IVs presents a trade-off between including a sufficient number of informative IVs and maintaining the overall strength of the selected IV set. While lowering the cutoff may improve statistical power by incorporating more IVs with moderate effects, setting it too low can introduce weak or null IVs that potentially violate the relevance assumption and compromise the validity of the MR analysis. In our proposed method, we provide a sufficient condition to ensure the asymptotic normality of the estimator, which depends on the average strength of the selected IVs relative to the cutoff value. Specifically, we choose a cutoff of 5×10^{-5} , commonly used as a threshold for suggestive significance in GWAS, to strike a balance between including informative IVs and maintaining the validity of the selected IV set. We note that Rao-Blackwellization has also been applied in [4] to efficiently combine information from an initial GWAS and a replication study to obtain unbiased estimates of SNP effect sizes. Our approach differs as we do not require a replication study to construct an unbiased estimation for β_{X_j} (see Supplement Materials Section 5 for details). Benefiting from such randomized IV selection, $\widehat{\beta}_{X_j, \text{RB}}$ is free of winner's curse bias, implying that $\mathbb{E}[\widehat{\beta}_{X_j, \text{RB}} | j \in \mathcal{S}_\lambda] = \beta_{X_j}$. Therefore, our proposed bias-corrected least squares objective function and l_0 constraint optimization framework in the previous section can be applied:

$$\min_{\theta \in \mathbb{R}, r_j \in \mathbb{R}} \widehat{l}(\theta, \{r_j\}_{j \in \mathcal{S}_\lambda}), \text{ s.t. } \sum_{j \in \mathcal{S}_\lambda} \mathbb{1}_{(r_j=0)} = v. \quad (4)$$

As one reviewer suggested, we also implemented two l_1 -type methods and make comparison with our l_0 based method through simulations. Our results demonstrate that while both approaches maintain comparable Type I error control, absolute bias, mean squared error (MSE), and coverage probability across various scenarios, the l_0 -based CARE method achieves higher statistical power. We have added relevant descriptions, methods, and results in Supplemental Material Section S.2-S.3

and Section S.8.12, where the loss function is defined as

$$\begin{aligned}\widehat{l}(\theta, \{r_j\}_{j \in \mathcal{S}_\lambda}) &= \sum_{j \in \mathcal{S}_\lambda} \widehat{l}_j(\theta, r_j) = \sum_{j \in \mathcal{S}_\lambda} \frac{(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}} - r_j)^2}{\sigma_{Y_j}^2} - \frac{\theta^2 \cdot \widehat{\sigma}_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2} \mathbf{1}_{(r_j=0)}, \\ \widehat{\sigma}_{X_j, \text{RB}}^2 &= \sigma_{X_j}^2 \left(1 - \frac{1}{\eta^2} \frac{A_{j,+}\phi(A_{j,+}) - A_{j,-}\phi(A_{j,-})}{1 - \Phi(A_{j,+}) + \Phi(A_{j,-})} + \frac{1}{\eta^2} \left(\frac{\phi(A_{j,+}) - \phi(A_{j,-})}{1 - \Phi(A_{j,+}) + \Phi(A_{j,-})} \right)^2 \right).\end{aligned}$$

3.3 Bootstrap aggregation and statistical inference

Since the IV screening step can be rather noisy and we do not expect to perfectly screen out all invalid IVs, we next incorporate bagging (or bootstrap aggregation) [6] to reduce IV screening variability and to further improve statistical efficiency. Then, we adopt the non-parametric delta method [13] to construct a confidence interval for our bagged estimator.

To be specific, we draw bootstrap sample B times from \mathcal{S}_λ . For the b -th bootstrap sample (Denoted by $\mathcal{S}_{\lambda,b}^*$), we adjust the loss function as $\widehat{l}_b^*(\theta, \{r_j\}_{j \in \mathcal{S}_\lambda}) = \sum_{j \in \mathcal{S}_\lambda} w_{jb}^* \widehat{l}_j(\theta, r_j)$, where w_{jb}^* is the number of occurrences in $\mathcal{S}_{\lambda,b}^*$ for j -th IVs in \mathcal{S}_λ . Then, we conduct the invalid IV screening step for each bootstrap sample $\mathcal{S}_{\lambda,b}^*$ and select $\widehat{\mathcal{V}}_b = \{j : \widehat{r}_{jb} = 0 \text{ and } j \in \mathcal{S}_{\lambda,b}^*\}$. The downstream causal estimator is derived by aggregating the estimated effects from all bootstrap samples, that is:

$$\widehat{\theta}_b = \frac{\sum_{j \in \widehat{\mathcal{V}}_b} \widehat{\beta}_{Y_j} \widehat{\beta}_{X_j, \text{RB}} / \sigma_{Y_j}^2}{\sum_{j \in \widehat{\mathcal{V}}_b} (\widehat{\beta}_{X_j, \text{RB}}^2 - \widehat{\sigma}_{X_j, \text{RB}}^2) / \sigma_{Y_j}^2}, \quad \widetilde{\theta} = \frac{1}{B} \sum_{b=1}^B \widehat{\theta}_b, \quad (5)$$

where $\widehat{\theta}_b$ is obtained by refitting the loss function $\widehat{l}(\theta, \{r_j\}_{j \in \widehat{\mathcal{V}}_b})$.

To provide valid statistical inference on the true causal effect θ , we use the non-parametric delta method [14] to estimate the variance of the bagged estimator with $\widehat{\sigma}_n^2 = \sum_{j \in \mathcal{S}_\lambda} \widehat{S}_j^2$, where $\widehat{S}_j = B^{-1} \sum_{b=1}^B (w_{ib}^* - B^{-1} \sum_{k=1}^B w_{ik}^*) (\widehat{\theta}_b - \widetilde{\theta})$. Then we construct a $(1 - \alpha)$ -level confidence interval for θ with $[\widetilde{\theta} - z_{\alpha/2} \cdot \widehat{\sigma}_n, \widetilde{\theta} + z_{\alpha/2} \cdot \widehat{\sigma}_n]$. Here α is the upper $\alpha/2$ -quantile of the standard normal distribution.

In the remainder of this manuscript, we refer to the proposed method as Causal Analysis with Randomized Estimators (CARE). The formalization of our proposed algorithm can be found in Algorithm 2. We also provide the discussion on the time complexity of this algorithm in Section S.1 in Supplemental Material.

4 Theoretical investigations

To discuss our theoretical investigations in detail, we begin by revisiting and introducing notations and assumptions. Recall that the set of selected IVs after rerandomization is defined as $\mathcal{S}_\lambda = \{j : |\frac{\widehat{\beta}_{X_j}}{\sigma_{X_j}} + Z_j| > \lambda, j = 1, \dots, p\}$ and its cardinality is denoted as $|\mathcal{S}_\lambda| = s_\lambda$. We next define κ_λ as the average of squared standardized IV effects to measure the selected IV strength in \mathcal{S}_λ , that is $\kappa_\lambda = \frac{1}{s_\lambda} \sum_{j \in \mathcal{S}_\lambda} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}$. Among the selected IVs after rerandomization, we denote $\mathcal{V}_\lambda = \{j : j \in \mathcal{S}_\lambda \text{ and } r_j = 0\}$ as the set of valid IVs in \mathcal{S}_λ and denote its cardinality as $|\mathcal{V}_\lambda| = v_\lambda$.

Considering the dual sources of randomness in our proposed estimator (one from the original GWAS sample, and the other from the bootstrap resampling), we separate these two sources of randomness by denoting the conditional expectation taken with respect to bootstrap resampling as $\mathbb{E}^*[\cdot] = \mathbb{E}\left[\cdot \mid \mathcal{S}_\lambda, \{(\widehat{\beta}_{Y_j}, \widehat{\beta}_{X_{j,\text{RB}}})\}_{j \in \mathcal{S}_\lambda}\right]$. Next, we introduce three additional assumptions for our theoretical investigations:

Algorithm 2: CARE

```

for  $j \leftarrow 1$  to  $p$  do
  Generate a pseudo SNP-exposure association effect  $Z_j \sim \mathcal{N}(0, \eta^2)$ ,
  If  $\left| \frac{\hat{\beta}_{X_j}}{\sigma_{X_j}} + Z_j \right| > \lambda$ , Then select SNP  $j$ .
end
Define the set of selected SNPs as  $\mathcal{S}_\lambda = \{j : \left| \frac{\hat{\beta}_{X_j}}{\sigma_{X_j}} + Z_j \right| > \lambda, j = 1, 2, \dots, p\}$  and  $|\mathcal{S}_\lambda| = s_\lambda$ ,
for  $j \in \mathcal{S}_\lambda$  do
  Construct an unbiased estimator of  $\hat{\beta}_{X_j, \text{RB}}$  as
  
$$\hat{\beta}_{j, \text{RB}} = \hat{\beta}_{X_j} - \frac{\sigma_{X_j}}{\eta} \frac{\phi(A_{j,+}) - \phi(A_{j,-})}{1 - \Phi(A_{j,+}) + \Phi(A_{j,-})}, \text{ where } A_{j,\pm} = -\frac{\hat{\beta}_{X_j}}{\sigma_{X_j} \eta} \pm \frac{\lambda}{\eta}$$

  and  $\phi(\cdot)$  and  $\Phi(\cdot)$  denote the standard normal density and cumulative distribution functions.
end
for  $b = 1$  to  $B$  do
  Draw bootstrap sample  $\mathcal{S}_{\lambda,b}^*$  from  $\mathcal{S}_\lambda$ ,
  Conduct the invalid IV screening procedure for  $\mathcal{S}_{\lambda,b}^*$ 
  
$$\min_{\theta \in \mathbb{R}, r_j \in \mathbb{R}} \left\{ \hat{l}_b^*(\theta, \{r_j\}_{j \in \mathcal{S}_\lambda}) : \sum_{j \in \mathcal{S}_{\lambda,b}^*} \mathbb{1}_{r_j=0} = v \right\} \Rightarrow \hat{\mathcal{V}}_b^*(v) = \{j : \hat{r}_j = 0, j \in \mathcal{S}_{\lambda,b}^*\},$$

  where  $\hat{l}_b^*(\theta, \{r_j\}_{j \in \mathcal{S}_\lambda}) = \sum_{j \in \mathcal{S}_\lambda} w_{jb}^* \hat{l}_j(\theta, \{r_j\}_{j \in \mathcal{S}_\lambda})$ .
  Select the final estimated set of Valid IVs  $\hat{\mathcal{V}}_b^*$  by GBIC,
  Derive the causal estimator for the  $b$ -th bootstrap
  
$$\hat{\theta}_b = A_b^{-1} \sum_{j \in \hat{\mathcal{V}}_b} \frac{\hat{\beta}_{Y_j} \hat{\beta}_{X_j, \text{RB}}}{\sigma_{Y_j}^2}, \quad A_b = \sum_{j \in \hat{\mathcal{V}}_b} \frac{\hat{\beta}_{X_j, \text{RB}}^2 - \hat{\sigma}_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2}.$$

end
Obtain the final estimator by bootstrap aggregation  $\tilde{\theta} = \frac{1}{B} \sum_{b=1}^B \hat{\theta}_b$ ,
Adopt the non-parametric delta method to estimate the variance of the bagged estimator with

$$\hat{\sigma}_n^2 = \sum_{j \in \mathcal{S}_\lambda} \hat{S}_j^2, \quad \hat{S}_j = \frac{1}{B} \sum_{b=1}^B \left( w_{jb}^* - \frac{1}{B} \sum_{k=1}^B w_{ik}^* \right) (\hat{\theta}_b - \tilde{\theta}),$$

Construct a  $(1 - \alpha)$ -level confidence interval for  $\theta$  with  $[\tilde{\theta} - z_{\alpha/2} \cdot \hat{\sigma}_n, \tilde{\theta} + z_{\alpha/2} \cdot \hat{\sigma}_n]$ , here  $z_{\alpha/2}$  is the upper  $\alpha/2$ -quantile of the standard normal distribution.

```

Assumption 2 (Variance stabilization) *There exists a variance stabilizing quantity a_λ and a vector $\tau \in \mathbb{R}^{s_\lambda}$ in which each component is independent of $\{(u_j, \nu_j)\}_{j \in S_\lambda}$ and uniformly bounded away from infinity in probability in the sense that*

$$\sup_{j \in S_\lambda} \left| a_\lambda \cdot \mathbb{E}^* \left[A_b^{-1} \cdot \hat{w}_{jb} \right] - \tau_j \right| = o_p(1),$$

where $A_b = \sum_{k \in S_\lambda} \hat{w}_{kb} \cdot (\hat{\beta}_{X_k, \text{RB}}^2 - \hat{\sigma}_{X_k, \text{RB}}^2) / \sigma_{Y_k}^2$, and $\hat{w}_{jb} = w_{jb}^* \cdot \mathbb{I}(\hat{r}_{jb} = 0) \cdot \mathbb{I}(w_{jb}^* \geq 1)$. In addition,

there is no dominating IV in the sense that $\frac{\max_{j \in \mathcal{S}_\lambda} \beta_{X_j}^2}{\sum_{j \in \mathcal{S}_\lambda} \beta_{X_j}^2} \xrightarrow{p} 0$.

The first part of the above assumption, intuitively, ensures that our estimator $\tilde{\theta}$ converges to a non-degenerative distribution asymptotically when appropriately scaled by $a_\lambda/\sqrt{s_\lambda \cdot \kappa_\lambda}$. This scaling factor accounts for the number of selected instruments and their average strength, enabling valid statistical inference. The second part of the condition requires that, after selection, no single IV exerts a “dominating effect” on exposure, which aligns with the biological understanding that complex traits are influenced by many genetic variants with small effects (i.e., the omnigenic model [5]). To cast more insight into Assumption 2, in Section S.4.3 of the Supplemental Material, we consider a special case where perfect IV screening is achieved. We show that in this case, Assumption 2 holds for both valid and invalid IVs in \mathcal{S}_λ .

Assumption 3 (Negligible invalid IV induced bias) *There is negligible bias induced by potential imperfect screening of invalid IVs after bootstrap aggregation in the sense that*

$$\frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \mathbb{E}^* \left[A_b^{-1} \sum_{j \in \mathcal{S}_\lambda} \widehat{\beta}_{X_{j,\text{RB}}} \cdot r_j \cdot \widehat{w}_{jb} / \sigma_{Y_j}^2 \right] = o_p(1).$$

Our theoretical investigations reveal two sets of sufficient conditions under which Assumption 3 holds (See Section S.5 and S.6 in the Supplemental Material). The first set of sufficient conditions ensures that the selected IVs are “nearly perfect,” meaning they are valid but do not include all possible valid IVs. We show that this nearly perfect IV screening property can be satisfied when there is strong prior knowledge about the trait’s genetic architecture or where valid and invalid IVs are easily distinguishable. The second set of sufficient conditions ensures Assumption 3 holds even if our proposed IV screening procedure does not screen all invalid IVs. In particular, our analysis indicates that when IVs with large r_j values (strong pleiotropic effects) are effectively screened out, our estimator maintains consistency even if the selected set includes some invalid IVs with small r_j values (weak pleiotropic effects). Together, these theoretical investigations suggest that perfect IV screening is not a prerequisite for valid inference in our proposed method.

Assumption 4 (Instrument Selection) *Define $\underline{\eta} = \min_{1 \leq j \leq p} \eta_j$ and $\bar{\eta} = \max_{1 \leq j \leq p} \eta_j$, then both $\underline{\eta}$ and $\bar{\eta}$ are bounded and bounded away from zero.*

The above assumption requires that the parameter η should not be too small or too large, as it

impacts the concentration behavior and asymptotic normality of our estimator. This assumption can be satisfied by design in our method. We recommend using a default value of $\eta_j = 0.5$ for all j (where $1 \leq j \leq p$), which ensures that both $\underline{\eta}$ and $\bar{\eta}$ are bounded and bounded away from zero. This choice simplifies the implementation while maintaining the theoretical guarantees of our method. Our simulation study also suggests that our method is not sensitive to the choice of η .

We are now in a position to describe the asymptotic behavior of our bootstrap aggregated estimator. Without loss of generality, we consider a particular form of our estimator in an ideal case where $\tilde{\theta} = \mathbb{E}^*[\hat{\theta}_b]$.

Theorem 1 *Under Assumptions S1-4, as $s_\lambda \xrightarrow{p} \infty$ and $\frac{\kappa_\lambda}{\lambda^2} \xrightarrow{p} \infty$, our proposed estimator satisfies the following representation*

$$\frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \cdot (\tilde{\theta} - \theta) = \frac{1}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \sum_{j \in S_\lambda} \tau_j \cdot \tilde{u}_j + o_p(1).$$

where $\tilde{u}_j = \hat{\beta}_{X_{j,\text{RB}}}(\theta \cdot \beta_{X_j} + \nu_j) - \theta(\hat{\beta}_{X_{j,\text{RB}}}^2 - \hat{\sigma}_{X_{j,\text{RB}}}^2)$. Therefore, conditional on the selection event \mathcal{S}_λ , our estimator converges to a Gaussian distribution, that is

$$\tilde{\sigma}^{-1}(\tilde{\theta} - \theta) \sim N(0, 1), \text{ where } \tilde{\sigma}^2 = \frac{\sum_{j \in S_\lambda} \tau_j^2 \mathbb{V}[\tilde{u}_j | \mathcal{S}_\lambda]}{a_\lambda^2}.$$

In the theorem above, we consider the asymptotic regime in which both $s_\lambda \xrightarrow{p} \infty$ and $\frac{\kappa_\lambda}{\lambda^2} \xrightarrow{p} \infty$ tend towards infinity. This asymptotic regime is quite natural in the context of MR. On the one hand, $s_\lambda \xrightarrow{p} \infty$ requires the number of IVs selected through re-randomization to be large enough, so that our inverse variance weighting-based estimator exhibits concentrated behavior. On the other hand, the condition $\frac{\kappa_\lambda}{\lambda^2} \xrightarrow{p} \infty$ does not involve the bootstrapping procedure; instead, it pertains to the strength of the selected IVs relative to the threshold λ used in the re-randomization step (Step 1). This assumption ensures that, on average, the selected IVs are sufficiently strong compared to the threshold, thereby satisfying the relevance assumption. It is also likely to hold, as it is of the same order as the GWAS sample size n after IV selection through re-randomization. From a theoretical standpoint, both conditions have been rigorously verified in [33] under appropriate conditions.

5 Simulations studies

We generate different simulation settings to evaluate the methods performance. To save space, the simulation settings are put into Supplementary Section S.8.1. Figure 2 summarizes the performance of various MR methods under the setting of 50% of the IVs are invalid, which we discuss below.

First, both cML (Type 1 error rate: 0.136) and MR-Lasso (0.112) produce inflated Type 1 error rates. This is because cML and MR-Lasso ignore the randomness in the valid IV selection procedure and assume all invalid IVs have been screened out, which is not the case under this simulation setting. In contrast, cML-DP (0.042) and CARE (0.042), which explicitly consider the randomness in valid IV selection, yield well-calibrated Type 1 error rates. Furthermore, other benchmark methods, including (random effects) IVW (0.056), MR-Egger (0.050), MRmix (0.020), MR-Median (0.032), MR-mode (0.004), MR-APSS (0.054) and RAPS (0.038) also yield well-controlled Type 1 error rates, though MRmix, MR-Median, MR-mode, and RAPS yield slightly conservative Type 1 error rates. Notably, the winner’s curse bias itself does not cause an inflated Type 1 error rate issue [33], partially explaining the robust performance of many MR methods under the null.

Second, CARE achieves considerably higher statistical power than benchmark methods (Figure 2a). Notably, CARE corrects the winner’s curse bias and measurement error bias, which allows for a more liberal threshold (say, $p < 5 \times 10^{-5}$) for instrument selection, resulting in higher power than other methods that typically use the genome-wide significance level ($p < 5 \times 10^{-8}$) as the threshold. Even though MR-APSS, like CARE, allows a liberal threshold ($p < 5 \times 10^{-5}$) due to its direct winner’s curse bias correction without theoretical guarantee, CARE outperforms MR-APSS, because of its full correction of the winner’s curse bias and meticulous consideration of measurement errors and invalid IVs.

Third, CARE yields smaller absolute bias compared to benchmark methods, attributable to its comprehensive approach to simultaneously addressing multiple sources of bias (measurement error bias, pleiotropic effects, and winner’s curse bias). In comparison, benchmark methods focus on addressing some biases specifically, leading to biased results. For instance, while MR-APSS directly corrects for the winner’s curse bias and considers potential invalid IVs, it still presents a larger absolute bias compared to CARE, possibly due to its more limited scope in bias correction and incomplete correction of the winner’s curse bias. However, while CARE significantly reduces bias,

its estimates are not entirely bias-free. This residual bias likely stems from the subtle differences between valid and invalid IVs. Consequently, the estimates are inevitably influenced by some invalid IVs, albeit to a lesser extent than in other methods. Furthermore, we confirm that ignoring the winner’s curse bias and directly applying the measurement error model with $\hat{\beta}_{X_j}$ in CARE generally results in worse performance, particularly concerning the absolute bias (Supplementary Figure S1). As expected, CARE yields much smaller MSE compared to benchmark methods as CARE has higher power and smaller absolute bias than any benchmark methods.

Fourth, the confidence intervals provided by CARE have coverage probabilities close to the nominal 95% level. When the absolute causal effect $|\theta|$ is large (say, 0.1), the absolute bias is relatively large, resulting in slight undercoverage of the true causal effect.

We conduct several additional simulations, including varying proportions of invalid IVs (Supplementary Section S.8.2), uniform-distributed effects in correlated pleiotropy (Supplementary Section S.8.3), balanced horizontal pleiotropy with InSIDE assumption satisfied (Supplementary Section S.8.4) and directional pleiotropy with InSIDE assumption violated (Supplementary Section S.8.5). The results patterns are similar.

Furthermore, to validate the results are not sensitive to the specific value of η within a reasonable range, we conducted sensitivity analyses using different values of η (0.1, 0.3, 0.5, 0.7, 0.9) in our main setting. The results demonstrate that the performance of our method remains stable and consistent for η values between 0.3 and 0.9 (Section S.8.6 in Supplementary Material). As expected, a very small η (0.1) led to worse results, likely due to insufficient rerandomization to fully account for the winner’s curse bias. Based on these findings, we recommend that practitioners use the default value of $\eta = 0.5$ in most cases without the need for dataset-specific fine-tuning.

While CARE demonstrates robust performance across various scenarios, it is important to note its limitations. As one reviewer suggested, we consider a simulation scenario that the parameter assumptions of other methods are true (where a three-sample MR design is used and the first GWAS is reserved solely for IV selection based on association strength so that the normality of $\hat{\beta}_{X_j}$ is not distorted). In this case, some alternative robust MR methods may outperform CARE, indicating that other robust MR methods may outperform CARE in a three-sample MR design (Supplementary Section S.8.13). Further simulations revealed two situations CARE is suboptimal. Firstly, in settings with non-linear relationships between genetic variants and exposures, CARE

showed slightly inflated Type 1 error rates, larger bias, and worse coverage (Section S.8.8 in Supplementary Material). This limitation stems from the method’s underlying assumption of linear relationships, which is common in MR studies and often justified by the predominantly linear or additive nature of genetic effects on complex traits [51]. Unlike our current approach, which exclusively utilizes GWAS summary data to estimate causal effects, recent advancements have addressed the non-linearity issue through methods like DeepMR [34], a deep learning-based approach applicable when individual-level DNA sequence data are available. Secondly, CARE’s performance may be compromised when the sample size of the exposure GWAS is small, resulting in a limited number of selected candidate IVs (Section S.8.9 in Supplementary Material). This issue may also arise due to a relatively small number of independent IVs (Section S.8.10 in Supplementary Material). Such scenarios can lead to increased sensitivity to violations of IV assumptions and challenge our asymptotic normality results, which require the number of candidate IVs to approach infinity. Users should exercise caution when applying CARE and other MR methods in these scenarios and consider alternative methods or larger sample sizes when possible.

In the end, it is worth mentioning that the core algorithm in CARE is written in C++ using the R package RcppArmadillo, and each step within the algorithm has a closed-form solution. Consequently, CARE has similar computational efficiency to many other methods, such as cML-DP and MRmix (Supplementary Figure S4), despite utilizing a larger number of IVs and a relatively high number of bootstrap iterations (2,000). Under the main simulation setting (12,000 simulations across 30%, 50%, and 70% invalid IVs), the average computational time of CARE is 12.6 seconds. Notably, the computational time for all methods is less than a minute in most situations when using one single core in a server. Thus, computational time should not be the primary consideration when deciding the method to be used.

6 Case studies

In this section, we investigate the performance of proposed CARE in two case studies. We put the data harmonization details in Supplementary Section S.9.1.

6.1 Negative control outcomes

To evaluate the Type 1 error rates in real data, we employ negative control outcome analyses, applying CARE and benchmark methods to investigate the causal effect of exposures on outcomes known a priori to have no causal relationship with the exposures. Briefly, in these negative control outcome analyses, the causal effect size is expected to be $\theta = 0$ [44] because negative control outcomes are determined prior to the exposures. However, unmeasured confounding factors may affect the estimates of θ . In particular, following others [44], we use ease of skin tanning to sun exposures and natural hair color before greying (six outcomes: Ease of skin tanning, Hair color black, Hair color red, Hair color blonde, Hair color light brown, and Hair color dark brown) as negative control outcomes. These data were downloaded from the IEU OpenGWAS Project [32] with GWAS ID: ukb-b-533 and ukb-d-1747. Notably, both tanning ability and natural hair color before greying are primarily determined at birth (thus, prior to considered exposures) but could be affected by unmeasured confounders [44]. In this setting, the inclusion of invalid IVs due to widespread pleiotropic effects or unmeasured confounding factors (e.g., population stratification) may result in incorrect rejections of the null hypothesis ($\theta = 0$) for MR analyses, leading to inflated Type 1 error rates.

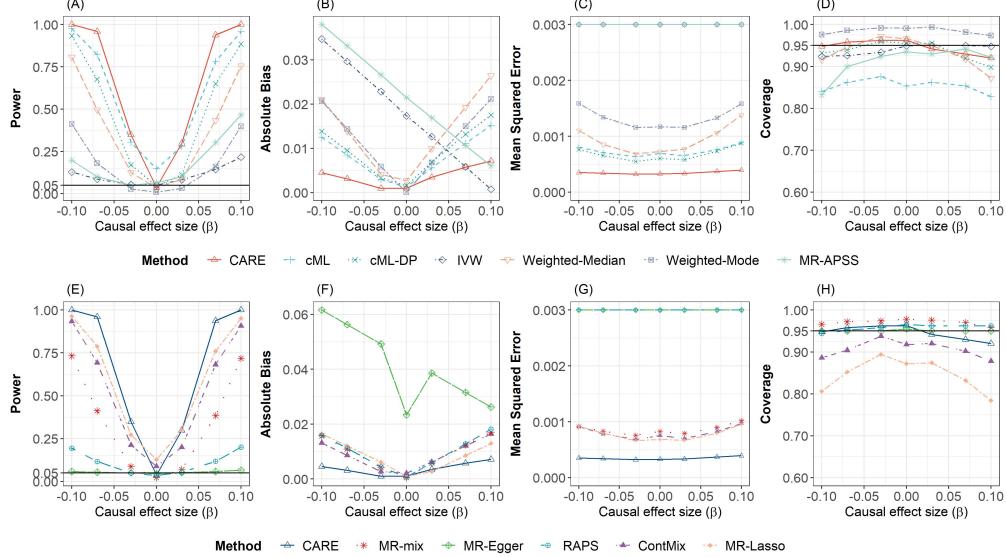


Figure 2: Power, absolute bias, mean squared error, and coverage of the CARE estimator and several robust MR methods under the main setting with 50% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold of 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

We consider 45 exposures, which include HDL cholesterol, body mass index (BMI), height, Alzheimer's disease, Lung cancer, Type 2 diabetes, stroke, asthma, and many others. All GWAS data are downloaded from the IEU OpenGWAS Project [32], and details of each exposure are relegated to the Supplementary Table 1. These exposures were selected based on their prevalence in existing literature and relevance to public health. Specifically, traits such as BMI, height, and HDL cholesterol have been extensively studied in genetic epidemiology and are known to be associated with various health outcomes. Disease outcomes like Alzheimer's disease, Type 2 diabetes, and cardiovascular diseases represent major public health concerns and have been the focus of numerous Mendelian randomization studies. This diverse set of exposures covers a wide range of physiological and pathological processes, allowing us to evaluate CARE's performance across various scenarios commonly encountered in Mendelian randomization studies. We apply CARE and benchmark methods to infer causal effects between these 45 exposures and six negative control outcomes (tanning ability and natural hair color before greying), resulting in 270 trait pairs. The corresponding p -values should follow a standard uniform distribution, given that the causal effect size $\theta = 0$ under the negative control outcomes analysis.

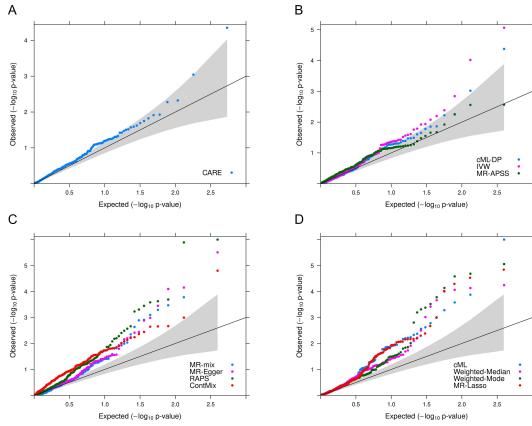


Figure 3: QQ plots of p -values in negative control outcome analysis. The gray-shaded part is 95% confidence interval.

Figure 3 summarizes the QQ-plots of $-\log_{10}(p)$ values for different methods. First, CARE yields well-calibrated p -values, indicating its reliability in controlling type 1 error rates under this negative control outcome analysis (Figure 3A). Similarly, IVW, cML-DP and MR-APSS also achieve good performance (Figure 3B). In contrast, MR-mix, MR-Egger, RAPS, ContMix, cML, Weighted-

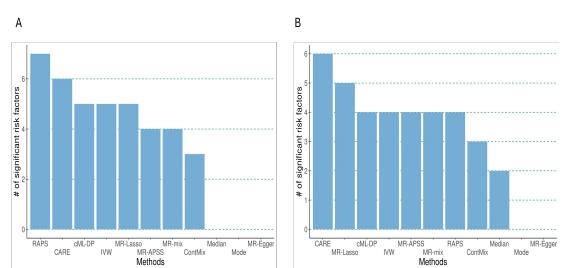


Figure 4: Number of significant causal pairs identified by different methods under Bonferroni-correction threshold $< 0.05/45 \simeq 10^{-3}$ using (A) 45 exposures used in negative control analysis and (B) 24 exposures that are reported by CDC and existing literature.

Median, Weighted-Mode, and MR-Lasso yield inflated p-values (Figures 3C and 3D). One may be surprised that widely used IVW achieves good performance. This is because we make every effort to make a fair comparison between different methods and use the (random effects) IVW to consider pleiotropic effects (i.e., invalid IVs) by allowing over-dispersion in the regression model. As expected, the fixed effects IVW that assumes all used IVs are valid leads to inflated p-values (Supplementary Figure S34A).

To understand why CARE performs well, we highlight two aspects. First, selecting valid IVs can be noisy in real data applications. That explains why cML and MR-Lasso, methods that ignore the screening variability in IV selection, produce inflated p-values (Figure 3D). Applying bagging reduces the screening variability and thus helps achieve well-calibrated p-values in CARE. Similarly, as cML-DP uses a data perturbation method to account for the screening variability, it also achieves relatively good performance. Second, CARE adopts a rerandomization step to select candidate IVs, accounting for the impact of the winner’s curse bias. Breaking the winner’s curse bias helps CARE achieve well-calibrated p-values as CARE uses a measurement error model and relies on the unbiasedness estimation of exposure-SNP effect β_{X_j} . This rerandomization step is crucial for CARE, and we confirm that applying CARE without the rerandomization step leads to inflated p-values (Supplementary Figure S34B).

6.2 Risk factors identification for COVID-19 severity

To better understand the underlying causal risk factors for COVID-19 severity and demonstrate the performance of our proposed method CARE, we apply CARE and competing MR methods to systematically identify causal risk factors for COVID-19 severity. Specifically, we investigate the same 45 exposures used in the negative control outcome analysis and use COVID-19 severity (B2) from the covid-19hg (B2, version v7, European ancestry only; [27]) as our outcome data. The dataset includes data from 32,519 hospitalized COVID-19 patients and 2,062,805 population controls.

First, we compare the number of significant causal exposures identified by CARE and competing methods under the Bonferroni correction ($< 0.05/45 \simeq 10^{-3}$) (Figure 4A). CARE identifies 6 causal exposures. In comparison, the competing methods RAPS, cML-DP, IVW, MR-Lasso, MR-APSS, MR-mix, ContMix, Weighted-Median, Weighted-Mode, MR-Egger identify 7, 5, 5, 5, 4, 4, 3, 0, 0

and 0 causal exposures, respectively. In terms of statistical power, CARE ranks second among all MR methods considered. RAPS achieves the highest power but also yields inflated p-values in our negative control outcome analysis and simulations, primarily due to neglecting variability in valid IV selection step.

Second, we compared the risk factors identified by different MR methods to known factors that meet two criteria: (1) they have been reported by the CDC or in peer-reviewed literature, and (2) they overlap with the 45 exposures used in our negative control outcome analyses. Through a comprehensive manual review by two researchers, we identified 24 well-established risk factors for COVID-19 severity (Supplementary Table 1). Notably, our new method, CARE, demonstrated superior performance by correctly identifying six of these 24 known risk factors: BMI, extreme BMI, HDL cholesterol, obesity class 1, obesity class 2, and overweight. In comparison, benchmark methods showed lower detection rates: MR-LASSO identified 5 risk factors, while cML-DP, IVW, MR-APSS, MR-Mix, and RAPS each identified 4. ContMix detected 3, and Median identified 2. Both Weighted-Mode and MR-Egger failed to identify any risk factors (Figure 4B). Importantly, CARE also avoided false positives, i.e., it did not incorrectly identify any factors lacking strong supporting evidence in the literature. In contrast, several benchmark methods produced potential false positives. For example, cML-DP incorrectly identified childhood obesity as a risk factor, while IVW erroneously identified both celiac disease and childhood obesity. Finally, when we focus on four methods with relatively good performance under our negative control outcome analysis, the result patterns are similar (Supplementary Section S.9.2).

In summary, CARE achieves high power in identifying likely causal risk factors for COVID-19 severity, and the identified risk factors can be largely validated by complementary analyses and literature.

7 Conclusion

We introduced a unified two-sample Mendelian randomization within the summary data framework, referred to as Causal Analysis with Randomized Estimators (CARE), that accounts for winner's curse, measurement error bias, and genetic pleiotropy simultaneously. Through simulations and biomedical applications, we demonstrate that CARE delivers robust causal effect estimates with

improved statistical power. More importantly, the CARE estimator enjoys rigorous theoretical guarantees under mild assumptions, which is often lacking for competing methods.

References

- [1] Boehm, F. J. and Zhou, X. (2022). Statistical methods for mendelian randomization in genome-wide association studies: A review. *Computational and Structural Biotechnology Journal*, 20:2338–2351.
- [2] 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.
- [3] Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016). Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*, 40(4):304–314.
- [4] Bowden, J. and Dudbridge, F. (2009). Unbiased estimation of odds ratios: combining genomewide association scans with replication studies. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 33(5):406–418.
- [5] Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An expanded view of complex traits: from polygenic to omnigenic. *Cell*, 169(7):1177–1186.
- [6] Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.
- [7] Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., and Neale, B. M. (2015). Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3):291–295.
- [8] 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.
- [9] Burgess, S., Foley, C. N., Allara, E., Staley, J. R., and Howson, J. M. (2020). A robust

and efficient method for mendelian randomization with hundreds of genetic variants. *Nature Communications*, 11(1):1–11.

[10] Burgess, S. and Thompson, S. G. (2017). Interpreting findings from mendelian randomization using the mr-egger method. *European Journal of Epidemiology*, 32(5):377–389.

[11] 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.

[12] Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the american statistical association*, 90(432):1200–1224.

[13] Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. SIAM.

[14] Efron, B. (2014). Estimation and accuracy after model selection. *Journal of the American Statistical Association*, 109(507):991–1007.

[15] Flegal, K. M., Graubard, B. I., Williamson, D. F., and Cooper, R. S. (2011). Reverse causation and illness-related weight loss in observational studies of body weight and mortality. *American Journal of Epidemiology*, 173(1):1–9.

[16] Gao, M., Wang, Q., Piernas, C., Astbury, N. M., Jebb, S. A., Holmes, M. V., and Aveyard, P. (2022). Associations between body composition, fat distribution and metabolic consequences of excess adiposity with severe covid-19 outcomes: observational study and mendelian randomisation analysis. *International Journal of Obesity*, 46(5):943–950.

[17] Gelman, A. and Imbens, G. (2013). Why ask why? Forward causal inference and reverse causal questions. Technical report, National Bureau of Economic Research.

[18] Gkatzionis, A. and Burgess, S. (2019). Contextualizing selection bias in mendelian randomization: how bad is it likely to be? *International Journal of Epidemiology*, 48(3):691–701.

[19] Grant, A. J. and Burgess, S. (2024). A bayesian approach to mendelian randomization using summary statistics in the univariable and multivariable settings with correlated pleiotropy. *The American Journal of Human Genetics*, 111(1):165–180.

[20] Guo, Z., Kang, H., Tony Cai, T., and Small, D. S. (2018). Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):793–815.

[21] Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., Jansen, R., De Geus, E. J., Boomsma, D. I., Wright, F. A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, 48(3):245–252.

[22] Hartwig, F. P., Davey Smith, G., and Bowden, J. (2017). Robust inference in summary data mendelian randomization via the zero modal pleiotropy assumption. *International Journal of Epidemiology*, 46(6):1985–1998.

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

[24] Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., Langdon, R., et al. (2018b). The mr-base platform supports systematic causal inference across the human genome. *Elife*, 7:e34408.

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

[26] Imai, K., Keele, L., Tingley, D., and Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, 105(4):765–789.

[27] Initiative, C.-. H. G. (2021). Mapping the human genetic architecture of covid-19. *Nature*, 600(7889):472–477.

[28] Kang, H., Zhang, A., Cai, T. T., and Small, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American statistical Association*, 111(513):132–144.

[29] 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.

[30] Liu, Z., Qin, Y., Wu, T., Tubbs, J. D., Baum, L., Mak, T. S. H., Li, M., Zhang, Y. D., and Sham, P. C. (2023). Reciprocal causation mixture model for robust mendelian randomization analysis using genome-scale summary data. *Nature Communications*, 14(1):1131.

[31] Luo, R., Wang, H., and Tsai, C.-L. (2008). On mixture regression shrinkage and selection via the mr-lasso. *International Journal of Pure and Applied Mathematics*, 46(3):403–414.

[32] Lyon, M. S., Andrews, S. J., Elsworth, B., Gaunt, T. R., Hemani, G., and Marcra, E. (2021). The variant call format provides efficient and robust storage of gwas summary statistics. *Genome Biology*, 22(1):1–10.

[33] Ma, X., Wang, J., and Wu, C. (2023). Breaking the winner’s curse in mendelian randomization: Rerandomized inverse variance weighted estimator. *The Annals of Statistics*, 51(1):211–232.

[34] Malina, S., Cizin, D., and Knowles, D. A. (2022). Deep mendelian randomization: Investigating the causal knowledge of genomic deep learning models. *PLOS Computational Biology*, 18(10):e1009880.

[35] Morrison, J., Knoblauch, N., Marcus, J. H., Stephens, M., and He, X. (2020). Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nature Genetics*, 52(7):740–747.

[36] Nain, M., Gupta, A., Malhotra, S., and Sharma, A. (2022). High-density lipoproteins may play a crucial role in covid-19. *Virology Journal*, 19(1):135.

[37] Pearl, J. (2009). *Causality*. Cambridge University Press.

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

[39] Qi, G. and Chatterjee, N. (2021). A comprehensive evaluation of methods for Mendelian

randomization using realistic simulations and an analysis of 38 biomarkers for risk of type 2 diabetes. *International Journal of Epidemiology*, 50(4):1335–1349.

[40] Rees, J. M., Wood, A. M., Dudbridge, F., and Burgess, S. (2019). Robust methods in mendelian randomization via penalization of heterogeneous causal estimates. *PloS One*, 14(9):e0222362.

[41] Robertson, D. S., Prevost, A. T., and Bowden, J. (2016). Accounting for selection and correlation in the analysis of two-stage genome-wide association studies. *Biostatistics*, 17(4):634–649.

[42] Sadreev, I. I., Elsworth, B. L., Mitchell, R. E., Paternoster, L., Sanderson, E., Davies, N. M., Millard, L. A., Smith, G. D., Haycock, P. C., Bowden, J., et al. (2021). Navigating sample overlap, winner’s curse and weak instrument bias in Mendelian randomization studies using the UK biobank. medRxiv.

[43] Sanderson, E., Glymour, M. M., Holmes, M. V., Kang, H., Morrison, J., Munafò, M. R., Palmer, T., Schooling, C. M., Wallace, C., Zhao, Q., et al. (2022). Mendelian randomization. *Nature Reviews Methods Primers*, 2(1):6.

[44] Sanderson, E., Richardson, T. G., Hemani, G., and Davey Smith, G. (2021). The use of negative control outcomes in Mendelian randomization to detect potential population stratification. *International Journal of Epidemiology*, 50(4):1350–1361.

[45] Skrivankova, V. W., Richmond, R. C., Woolf, B. A., Davies, N. M., Swanson, S. A., Van der Weele, T. J., Timpson, N. J., Higgins, J. P., Dimou, N., Langenberg, C., et al. (2021). Strengthening the reporting of observational studies in epidemiology using Mendelian randomisation (STROBE-MR): Explanation and elaboration. *BMJ*, 375.

[46] Smith, G. D. and Ebrahim, S. (2004). Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology*, 33(1):30–42.

[47] Stephens, M. (2017). False discovery rates: a new deal. *Biostatistics*, 18(2):275–294.

[48] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

[49] Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494.

[50] Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A. N., Knowles, D. A., Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., et al. (2019). Opportunities and challenges for transcriptome-wide association studies. *Nature genetics*, 51(4):592–599.

[51] Wainschtein, P., Jain, D., Zheng, Z., Cupples, L. A., Shadyab, A. H., McKnight, B., Shoemaker, B. M., Mitchell, B. D., et al. (2022). Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nature Genetics*, 54(3):263–273.

[52] Watanabe, K., Stringer, S., Frei, O., Mirkov, M. U., de Leeuw, C., Polderman, T. J., van der Sluis, S., Andreassen, O. A., Neale, B. M., and Posthuma, D. (2019). A global overview of pleiotropy and genetic architecture in complex traits. *Nature Genetics*, 51(9):1339–1348.

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

[54] Ye, T., Shao, J., and Kang, H. (2021). Debiased inverse-variance weighted estimator in two-sample summary-data mendelian randomization. *The Annals of Statistics*, 49(4):2079–2100.

[55] Zeng, J., De Vlaming, R., Wu, Y., Robinson, M. R., Lloyd-Jones, L. R., Yengo, L., Yap, C. X., Xue, A., Sidorenko, J., McRae, A. F., et al. (2018). Signatures of negative selection in the genetic architecture of human complex traits. *Nature Genetics*, 50(5):746–753.

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

[57] Zhong, H. and Prentice, R. L. (2010). Correcting “winner’s curse” in odds ratios from genomewide association findings for major complex human diseases. *Genetic Epidemiology*, 34(1):78–91.

[58] Zöllner, S. and Pritchard, J. K. (2007). Overcoming the winner’s curse: estimating penetrance parameters from case-control data. *The American Journal of Human Genetics*, 80(4):605–615.

[59] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

SUPPLEMENTARY MATERIAL

Contents

1	Introduction	1
1.1	Background and motivation	1
1.2	Contribution	4
2	Framework and challenges	5
3	Methodology	8
3.1	Measurement error correction and invalid IV screening	8
3.2	Unknown S and practical implementation	10
3.3	Bootstrap aggregation and statistical inference	13
4	Theoretical investigations	14
5	Simulations studies	18
6	Case studies	20
6.1	Negative control outcomes	21
6.2	Risk factors identification for COVID-19 severity	23
7	Conclusion	24
S.1	Algorithm to solve the optimization problem in (4)	3
S.1.1	Algorithm to solve the optimization problem in (4)	3
S.1.2	Justification of unique solution of Problem (4) under fixed θ	6
S.1.3	Adoption of l_0 penalty instead of using Lasso	8
S.2	Algorithm to solve the optimization problem using l_1 penalty	10
S.3	Theoretical justifications for two l_1 methods	14
S.3.1	Method 1	14
S.3.2	Method 2	17

S.4 Proof of Theorem 1	19
S.4.1 Notions and Assumptions	19
S.4.2 Proof	21
S.4.3 Verifying the Assumption S2 in the case with perfect screening property	22
S.4.4 The asymptotic analysis of A_b under perfect screening property	23
S.5 Invalid IV screening consistency	26
S.5.1 Notations	26
S.5.2 Sufficient conditions	28
S.5.3 Theoretical Results	31
S.5.4 Proof of Theorem S1	32
S.5.4.1 Case 1: $ \mathcal{V}^* = v^* \geq v_\lambda$ and $\mathcal{V}^* \neq \mathcal{V}_\lambda$	33
S.5.4.2 Case 2: When $c_1 \cdot v_\lambda \leq \mathcal{V}^* < v_\lambda$ but $\rho(\mathcal{V}^*) < 1$	38
S.5.4.3 Case 3: When $ \mathcal{V}^* = v^* < c_1 \cdot v_\lambda$	41
S.5.5 Proof of Perfect Screening Property	43
S.5.5.1 Case 1: $ \mathcal{V}^* = v^* \geq v_\lambda$ and $\mathcal{V}^* \neq \mathcal{V}_\lambda$	43
S.5.5.2 Case 2: When $ \mathcal{V}^* = v^* < v_\lambda$	44
S.5.6 Lemmas	45
S.5.7 Proof of Lemmas	47
S.5.7.1 Proof of Lemma 1	47
S.5.7.2 Proof of Lemma 2	51
S.5.7.3 Proof of Lemma 3	52
S.6 An example that Assumption S3 is satisfied without perfect screening	53
S.6.1 Main results	53
S.6.2 Proof of Theorem S3	56
S.6.3 Proof of Lemma 4	58
S.6.4 Proof of Lemma 5	65
S.6.5 Additional Lemmas	68
S.6.5.1 Proof of Lemma 6	68
S.6.6 Sufficient conditions for the Boundness condition	70

S.7 Connections and differences with [4]	73
S.8 Simulation settings and additional simulation results	75
S.8.1 Simulation settings	75
S.8.2 Different proportions of invalid IVs, CARE without winner’s curse, and running time	77
S.8.3 Uniform distributed effects in correlated pleiotropy	80
S.8.4 Balanced horizontal pleiotropy with InSIDE assumption satisfied	80
S.8.5 Directional horizontal pleiotropy with InSIDE assumption violated	87
S.8.6 Sensitivity analysis using different values of η	90
S.8.7 Consistency of using GBIC with different choices of κ_n as model selection methods .	90
S.8.8 Nonlinear settings	92
S.8.9 Sample size variation of GWAS	96
S.8.10 Variations in number of SNPs	100
S.8.11 Using the same liberal threshold	104
S.8.12 Comparison of l_0 and l_1 algorithms	105
S.8.13 Third sample for selecting IVs	105
S.9 Additional Real Data Results	108
S.9.1 Data harmonization	108
S.9.2 Comparative analysis of four MR methods for assessing COVID-19 severity	108
S.9.3 Supplementary tables and figures for real data results	109

S.1 Algorithm to solve the optimization problem in (4)

S.1.1 Algorithm to solve the optimization problem in (4)

In the section, we provide an algorithm borrowing ideas from coordinate descent [49] to solve the optimization problem in (4), that is

$$\min_{\theta \in \mathbb{R}, r_j \in \mathbb{R}} \left\{ \hat{l}(\theta, \{r_j\}_{j \in \mathcal{S}_\lambda}) : \sum_{j \in \mathcal{S}_\lambda} \mathbf{1}(r_j = 0) = v \right\} \Rightarrow \hat{\mathcal{V}}(v) = \{j : \hat{r}_j = 0, j \in \mathcal{S}_\lambda\},$$

This step allows us to screen out invalid IVs and select \mathcal{V} .

We note that the proposed algorithm borrows strength from the classical coordinate descent algorithm by iteratively minimizing the objective function by fixing either θ or r_j 's. As our algorithm aims to screen out invalid IVs with $r_j \neq 0$, one difference is that we iteratively search for IVs with large “residuals” (i.e., $\widehat{\beta}_{Yj} - \theta \widehat{\beta}_{X_j, \text{RB}}$) in Step 2. (i) so that the objective function can be further minimized. Furthermore, as our optimization problem involves l_0 penalty, instead of choosing the model size v based on cross-validation frequently adopted in Lasso-type problems [48, 59], we adopt the Bayesian Information Criterion to select the final set of valid IVs.

Our proposed algorithm consists of three steps as follows:

Algorithm 3: Algorithm to solve the optimization problem in (4)

Input: Data inputs and initial parameters

Output: Estimated parameters $\hat{\theta}$ and \hat{r}_j

Initialization Set $k = 0$, generate $\theta^{(0)} \sim \text{Uniform} \left(\min_{1 \leq j \leq s_\lambda} \frac{\hat{\beta}_{Y_j}}{\hat{\beta}_{X_j}}, \max_{1 \leq j \leq s_\lambda} \frac{\hat{\beta}_{Y_j}}{\hat{\beta}_{X_j}} \right)$;

Block Coordinate Descent

repeat

 Fix $\theta^{(k)}$, update $r_j^{(k+1)}$;

 Order $\frac{(\hat{\beta}_{Y_j} - \theta^{(k)} \cdot \hat{\beta}_{X_j, \text{RB}})^2}{\sigma_{Y_j}^2} - \frac{\theta^2 \cdot \hat{\sigma}_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2}$, $j = 1, 2, \dots, s_\lambda - v$ in decreasing order;

 Set $r_j^{(k+1)} = \hat{\beta}_{Y_j} - \theta^{(k)} \hat{\beta}_{X_j, \text{RB}}$ for the largest $s_\lambda - v$ components, $j = 1, \dots, s_\lambda - v$,
 and $r_j^{(k+1)} = 0$ for $j = s_\lambda - v + 1, \dots, s_\lambda$;

 Fix $r_j^{(k+1)}$, update $\theta^{(k)}$ by minimizing the following objective function:

$$\theta^{(k+1)} = \arg \min_{\theta \in \mathbb{R}} \sum_{j \in \mathcal{S}_\lambda} \frac{(\hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_j, \text{RB}} - r_j^{(k+1)})^2 - \theta^2 \cdot \hat{\sigma}_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2} \mathbb{1}_{(r_j^{(k+1)} = 0)}.$$

If $\left| \frac{\theta^{(k+1)} - \theta^{(k)}}{\theta^{(k)}} \right| < 10^{-7}$ **then** Stop and output $\hat{\theta}(v) = \theta^{(k+1)}$ and $\hat{r}_j(v) = r_j^{(k+1)}$;

else Set $k = k + 1$;

until $\left| \frac{\theta^{(k+1)} - \theta^{(k)}}{\theta^{(k)}} \right| < 10^{-7}$;

end

Valid IV Selection via BIC

for $v = 2, \dots, s_\lambda$ **do**

 Calculate

$$\text{BIC}(v) = -2\hat{l}(\hat{\theta}(v), \{\hat{r}_j(v)\}_{j \in \hat{\mathcal{V}}}) + \log(n) \cdot (s_\lambda - v);$$

end

 Select $\hat{\mathcal{V}}$ with the smallest $\text{BIC}(v)$;

end

S.1.2 Justification of unique solution of Problem (4) under fixed θ

To cast some insights into the proposed Algorithm 3 for solving Problem (3), we note that in each iteration, our algorithm breaks the optimization into two sub-problems and provides a closed-form global optimal solution for these sub-problems.

In the first sub-problem, we fix θ and treat Problem (3) as an optimization problem with respect to $\{r_j\}_{j \in \mathcal{S}}$:

$$\min_{\theta, r_j} l(\theta, \{r_j\}_{j \in \mathcal{S}}), \text{ s.t. } \sum_{j \in \mathcal{S}} \mathbb{1}_{(r_j=0)} = v.$$

Unlike classical l_0 constrained linear regression with an arbitrary design matrix, solving this problem is computationally efficient as we can decompose the original loss function into the sum of $l_j(\theta, r_j)$. Each $l_j(\theta, r_j)$ only depends on a single r_j . In this case, a closed-form solution to this optimization problem can be given.

As we can see that, for invalid IVs with $r_j \neq 0$, $l_j(\theta, r_j)$ reaches its minimum 0 by setting $r_j = \hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_j}$ (See justifications below). While for valid IVs with $r_j = 0$, $l_j(\theta, r_j)$ takes a constant value of $\frac{1}{2}(\hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_j})^2 / \sigma_{Y_j}^2 - \frac{1}{2}\theta^2 \cdot \sigma_{X_j}^2 / \sigma_{Y_j}^2$.

Therefore, to minimize the the loss function $l(\theta, \{r_j\}_{j \in \mathcal{S}})$ for given θ , we only need to find v IVs with the smallest $\frac{1}{2}(\hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_j})^2 / \sigma_{Y_j}^2 - \frac{1}{2}\theta^2 \cdot \sigma_{X_j}^2 / \sigma_{Y_j}^2$ and set their $r_j = 0$ and the rest of r_j to $\hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_j}$. The Block Coordinate Descent Step of our algorithm is indeed providing such a closed-form global optimal solution of the above combinatorial optimization problem. After deriving $\{r_j\}_{j \in \mathcal{S}}$, we then solve our second sub-problem by solving Problem (3) with $\{r_j\}_{j \in \mathcal{S}}$ fixed. The alternative minimization of θ and $\{r_j\}_{j \in \mathcal{S}}$ together can ensure the objective function decay.

To justify any given θ , we can give a closed-form solution of the optimization problem

$$\min_{\{r_j\}_{j \in \mathcal{S}_\lambda}} l(\theta, \{r_j\}_{j \in \mathcal{S}_\lambda}) \text{ s.t. } \sum_{j \in \mathcal{S}_\lambda} \mathbb{1}_{r_j=0} = v,$$

we further investigate $l_j(\theta, r_j)$ and discuss the solution to this optimization problem in three different situations.

$$\frac{\partial l(\theta, \{r_j\}_{j \in \mathcal{S}_\lambda})}{\partial r_j} = \frac{\partial l_j(\theta, r_j)}{\partial r_j} = -\frac{\hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_j, \text{RB}} - r_j}{\sigma_{Y_j}^2} \text{ When } r_j \neq 0.$$

$$l_j(\theta, r_j) \triangleq \frac{1}{2} \frac{(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}})^2}{\sigma_{Y_j}^2} - \frac{1}{2} \frac{\theta^2 \cdot \sigma_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2} \text{ When } r_j = 0.$$

- In the case that $\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}} > 0$, we have $l_j(\theta, r_j)$ reach its local minimum 0 when $r_j = \widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}} > 0$.

When $r_j = 0$,

$$l_j(\theta, r_j) = \frac{1}{2} \frac{(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}})^2}{\sigma_{Y_j}^2} - \frac{1}{2} \frac{\theta^2 \cdot \sigma_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2}$$

And when $r_j < 0$, we have $\frac{\partial l(\theta, \{r_j\}_{j \in \mathcal{S}_\lambda})}{\partial r_j} = \frac{\partial l_j(\theta, r_j)}{\partial r_j} < 0$, and therefore

$$l_j(\theta, r_j) \geq \lim_{r_j \rightarrow 0^-} \frac{1}{2} \frac{(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}} - r_j)^2}{\sigma_{Y_j}^2} > 0.$$

Therefore we have

$$\min_{r_j \neq 0} l_j(\theta, r_j) = 0 \text{ and } l_j(\theta, r_j = 0) = \frac{1}{2} \frac{(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}})^2}{\sigma_{Y_j}^2} - \frac{1}{2} \frac{\theta^2 \cdot \sigma_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2}$$

- In the case that $\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}} < 0$, we have $l_j(\theta, r_j)$ reach its local minimum 0 when $r_j = \widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}} < 0$.

When $r_j = 0$,

$$l_j(\theta, r_j) = \frac{1}{2} \frac{(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}})^2}{\sigma_{Y_j}^2} - \frac{1}{2} \frac{\theta^2 \cdot \sigma_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2}$$

And when $r_j > 0$, we have $\frac{\partial l(\theta, \{r_j\}_{j \in \mathcal{S}_\lambda})}{\partial r_j} = \frac{\partial l_j(\theta, r_j)}{\partial r_j} > 0$, and therefore

$$l_j(\theta, r_j) \geq \lim_{r_j \rightarrow 0^+} \frac{1}{2} \frac{(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}} - r_j)^2}{\sigma_{Y_j}^2} > 0.$$

Therefore we have

$$\min_{r_j \neq 0} l_j(\theta, r_j) = 0 \text{ and } l_j(\theta, r_j = 0) = \frac{1}{2} \frac{(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}})^2}{\sigma_{Y_j}^2} - \frac{1}{2} \frac{\theta^2 \cdot \sigma_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2}$$

- In the case when $\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}} = 0$.

When $r_j > 0$, we have $\frac{\partial l(\theta, \{r_j\}_{j \in \mathcal{S}_\lambda})}{\partial r_j} = \frac{\partial l_j(\theta, r_j)}{\partial r_j} \geq 0$, and therefore

$$l_j(\theta, r_j) \geq \lim_{r_j \rightarrow 0^+} \frac{1}{2} \frac{(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}} - r_j)^2}{\sigma_{Y_j}^2} \geq 0.$$

When $r_j < 0$, we have $\frac{\partial l(\theta, \{r_j\}_{j \in \mathcal{S}_\lambda})}{\partial r_j} = \frac{\partial l_j(\theta, r_j)}{\partial r_j} < 0$, and therefore

$$l_j(\theta, r_j) \geq \lim_{r_j \rightarrow 0^-} \frac{1}{2} \frac{(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}} - r_j)^2}{\sigma_{Y_j}^2} \geq 0.$$

When $r_j = 0$, we have

$$l_j(\theta, r_j) = -\frac{1}{2} \frac{\theta^2 \cdot \sigma_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2}.$$

Therefore we have

$$\min_{r_j \neq 0} l_j(\theta, r_j) = 0 \text{ and } l_j(\theta, r_j = 0) = \frac{1}{2} \frac{(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}})^2}{\sigma_{Y_j}^2} - \frac{1}{2} \frac{\theta^2 \cdot \sigma_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2}.$$

S.1.3 Adoption of l_0 penalty instead of using Lasso

We adopted the l_0 penalty for three reasons:

- Unlike the classical l_0 constrained linear regression, our considered l_0 constrained optimization problem is computationally efficient to solve as closed form solutions of $\{r_j\}_{j \in \mathcal{S}_\lambda}$ can be derived when θ is fixed (See previous discussions in Section 1.2).
- Due to the inclusion of a measurement error bias correction term, $\frac{1}{2} \sum_{j \in \mathcal{S}_\lambda} \frac{\theta^2 \cdot \sigma_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2} I(r_j = 0)$, in our objective function, adopting a Lasso-type penalty results in an optimization problem with non-differentiable gradients, making the algorithm remains time-consuming to solve.
- Empirically, we have actually tested the use of the l_1 penalty, which was our original idea. There, to enable efficient optimization, we removed the bias correction term for the measurement error. Our preliminary investigations with the l_1 penalty revealed several limitations: i) The number of selected IVs exhibited high sensitivity to small changes in the tuning parameter λ . ii) The l_1 penalty's simultaneous penalization of valid and invalid IVs is suboptimal,

given the often subtle differences between these IVs in MR contexts. iii) The convex nature of the l_1 penalty resulted in discontinuous jumps in the number of selected IVs as λ varied, leading to suboptimal performance. In contrast, the l_0 penalty offers several advantages in our specific context: i) It provides a comprehensive set of potential solutions across varying numbers of potential (valid) IVs. ii) It better accommodates the nuanced differences between valid and invalid IVs typically encountered in MR studies.

These considerations collectively support the use of the l_0 penalty as a more suitable approach for our specific optimization problem in the MR framework.

S.2 Algorithm to solve the optimization problem using l_1 penalty

- **The first approach:** We replace the l_0 constraint with an l_1 penalty in the following objective function:

$$\begin{aligned}\widehat{l}(\theta, \{r_j\}_{j \in \mathcal{S}}, \gamma) &\triangleq \sum_{j \in \mathcal{S}} l_j(\theta, r_j) \\ l_j(\theta, r_j, \gamma) &= \frac{1}{2} \frac{(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j} - r_j)^2}{\sigma_{Y_j}^2} - \frac{1}{2} \frac{\theta^2 \cdot \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \mathbb{1}_{(r_j=0)} + \gamma |r_j|.\end{aligned}$$

For a fixed tuning parameter γ , we estimate the parameters by minimizing:

$$\min_{\theta \in \mathbb{R}, r_j \in \mathbb{R}} \widehat{l}(\theta, \{r_j\}_{j \in \mathcal{S}}, \gamma).$$

To solve this optimization problem, we alternate between minimizing with respect to θ and $\{r_j\}_{j \in \mathcal{S}}$. The optimal solution for $\{r_j\}_{j \in \mathcal{S}}$ given a fixed θ is:

$$r_j = \begin{cases} \text{sign}(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_{j,\text{RB}}}) \cdot (|\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_{j,\text{RB}}}| - \gamma \cdot \sigma_{Y_j}^2) & \text{if } |\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_{j,\text{RB}}}| > \gamma \cdot \sigma_{Y_j}^2 + |\theta| \cdot \sigma_{X_{j,\text{RB}}}, \\ 0 & \text{otherwise.} \end{cases}$$

Theoretical justifications for this result can be found in the Supplemental Material Section S.3. The optimal solution for θ , given fixed $\{r_j\}_{j \in \mathcal{S}}$, is:

$$\arg \min_{\theta \in \mathbb{R}} \sum_{j \in \mathcal{S}} \frac{(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_{j,\text{RB}}} - r_j)^2 - \theta^2 \cdot \widehat{\sigma}_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \mathbb{1}_{(r_j=0)}.$$

We iteratively update θ and $\{r_j\}_{j \in \mathcal{S}}$ until convergence. The tuning parameter γ is selected via BIC, and the corresponding estimator $\widehat{\theta}(\gamma)$ is used for inference. The full optimization procedure is detailed in **Algorithm 4**.

- **The second approach:** We further replace $\mathbb{1}_{(r_j=0)}$ in the measurement error term with $1 - |r_j|$ and derive the following objective function,

$$l(\theta, \{r_j\}_{j \in \mathcal{S}}) \triangleq \sum_{j \in \mathcal{S}} l_j(\theta, r_j)$$

$$l_j(\theta, r_j) = \frac{1}{2} \frac{(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j} - r_j)^2}{\sigma_{Y_j}^2} - \frac{1}{2} \frac{\theta^2 \cdot \sigma_{X_j}^2}{\sigma_{Y_j}^2} + (\gamma + \frac{1}{2} \frac{\theta^2 \cdot \sigma_{X_j}^2}{\sigma_{Y_j}^2}) |r_j|.$$

For a fixed tuning parameter γ , we also estimate the parameters by minimizing:

$$\min_{\theta \in \mathbb{R}, r_j \in \mathbb{R}} \widehat{l}(\theta, \{r_j\}_{j \in \mathcal{S}}, \gamma).$$

To solve this optimization problem, we alternate between minimizing with respect to θ and $\{r_j\}_{j \in \mathcal{S}}$. The optimal solution for $\{r_j\}_{j \in \mathcal{S}}$ given a fixed θ is:

$$r_j = \begin{cases} \text{sign}(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}}) \cdot (|\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}}| - \tilde{\lambda}_j \cdot \sigma_{Y_j}^2) & \text{if } |\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}}| > \tilde{\lambda}_j \cdot \sigma_{Y_j}^2, \\ 0 & \text{otherwise.} \end{cases}$$

where $\tilde{\lambda}_j = \gamma + \frac{1}{2} \frac{\theta^2 \cdot \sigma_{X_j}^2}{\sigma_{Y_j}^2}$ for all $j \in \mathcal{S}$. Theoretical justifications for this result can be found in the Supplemental Material Section S.3. The optimal solution for θ , given fixed $\{r_j\}_{j \in \mathcal{S}}$, is:

$$\arg \min_{\theta \in \mathbb{R}} \frac{1}{2} \sum_{j \in \mathcal{S}} \frac{(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j} - r_j)^2}{\sigma_{Y_j}^2} - \frac{1}{2} \sum_{j \in \mathcal{S}} \frac{\theta^2 \cdot \sigma_{X_j}^2}{\sigma_{Y_j}^2} + \sum_{j \in \mathcal{S}} (\gamma + \frac{1}{2} \frac{\theta^2 \cdot \sigma_{X_j}^2}{\sigma_{Y_j}^2}) |r_j|.$$

We also iteratively update θ and $\{r_j\}_{j \in \mathcal{S}}$ until convergence and use BIC to select the tuning parameter γ . The corresponding estimator $\widehat{\theta}(\gamma)$ is used for inference. The full optimization procedure is detailed in **Algorithm 5**.

Algorithm 4: Algorithm to solve the optimization problem in (4)

Input: Data inputs and initial parameters

Output: Estimated parameters $\hat{\theta}$ and \hat{r}_j

Initialization Set $k = 0$, generate $\theta^{(0)} \sim \text{Uniform} \left(\min_{1 \leq j \leq s_\lambda} \frac{\hat{\beta}_{Y_j}}{\hat{\beta}_{X_j}}, \max_{1 \leq j \leq s_\lambda} \frac{\hat{\beta}_{Y_j}}{\hat{\beta}_{X_j}} \right)$;

Block Coordinate Descent

repeat

 Fix $\theta^{(k)}$, update $r_j^{(k+1)}$:

 For $\forall j \in \mathcal{S}$: If $|\hat{\beta}_{Y_j} - \theta^{(k)} \cdot \hat{\beta}_{X_j, \text{RB}}| > \gamma \cdot \sigma_{Y_j}^2 + |\theta^{(k)}| \cdot \sigma_{X_j, \text{RB}}$, we let

$$r_j^{(k+1)} = \text{sign}(\hat{\beta}_{Y_j} - \theta^{(k)} \cdot \hat{\beta}_{X_j, \text{RB}}) \cdot (|\hat{\beta}_{Y_j} - \theta^{(k)} \cdot \hat{\beta}_{X_j, \text{RB}}| - \gamma \cdot \sigma_{Y_j}^2).$$

 Otherwise, we set $r_j^{(k+1)} = 0$.

 Fix $r_j^{(k+1)}$, update $\theta^{(k)}$ by minimizing the following objective function:

$$\theta^{(k+1)} = \arg \min_{\theta \in \mathbb{R}} \sum_{j \in \mathcal{S}} \frac{(\hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_j, \text{RB}} - r_j^{(k+1)})^2 - \theta^2 \cdot \hat{\sigma}_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2} \mathbb{1}_{(r_j^{(k+1)} = 0)}.$$

$$\theta^{(k+1)} = \frac{\sum_{j \in \mathcal{S}} \frac{\hat{\beta}_{X_j, \text{RB}} \cdot \hat{\beta}_{Y_j}}{\hat{\sigma}_{X_j, \text{RB}}^2} \mathbb{1}_{(r_j^{(k+1)} = 0)}}{\sum_{j \in \mathcal{S}} \left(\frac{\hat{\beta}_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2} - \frac{\hat{\sigma}_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2} \right) \mathbb{1}_{(r_j^{(k+1)} = 0)}}.$$

If $\left| \frac{\theta^{(k+1)} - \theta^{(k)}}{\theta^{(k)}} \right| < 10^{-7}$ **then** Stop and output $\hat{\theta}(\gamma) = \theta^{(k+1)}$ and $\hat{r}_j(\gamma) = r_j^{(k+1)}$;

else Set $k = k + 1$;

until $\left| \frac{\theta^{(k+1)} - \theta^{(k)}}{\theta^{(k)}} \right| < 10^{-7}$;

end

Valid IV Selection via BIC

for all candidate γ **do**

 Calculate

$$\text{BIC}(\gamma) = -2\hat{l}(\hat{\theta}(\gamma), \{\hat{r}_j(\gamma)\}_{j \in \hat{\mathcal{V}}_\gamma}) + \log(n) \cdot (s_\lambda - \hat{v}_\gamma);$$

end

 Select $\hat{\mathcal{V}}_\gamma$ with the smallest $\text{BIC}(\gamma)$;

end

Algorithm 5: Algorithm to solve the optimization problem in (4)

Input: Data inputs and initial parameters

Output: Estimated parameters $\hat{\theta}$ and \hat{r}_j

Initialization Set $k = 0$, generate $\theta^{(0)} \sim \text{Uniform} \left(\min_{1 \leq j \leq s_\lambda} \frac{\hat{\beta}_{Y_j}}{\hat{\beta}_{X_j}}, \max_{1 \leq j \leq s_\lambda} \frac{\hat{\beta}_{Y_j}}{\hat{\beta}_{X_j}} \right)$;

Block Coordinate Descent

repeat

 Fix $\theta^{(k)}$, update $r_j^{(k+1)}$:

 For $\forall j \in \mathcal{S}$: If $|\hat{\beta}_{Y_j} - \theta^{(k)} \cdot \hat{\beta}_{X_j, \text{RB}}| > \tilde{\lambda}_j \cdot \sigma_{Y_j}^2$, we let

$$r_j^{(k+1)} = \text{sign}(\hat{\beta}_{Y_j} - \theta^{(k)} \cdot \hat{\beta}_{X_j, \text{RB}}) \cdot (|\hat{\beta}_{Y_j} - \theta^{(k)} \cdot \hat{\beta}_{X_j, \text{RB}}| - \tilde{\lambda}_j \cdot \sigma_{Y_j}^2) \text{ where } \tilde{\lambda}_j = \gamma + \frac{1}{2} \frac{\theta^{(k)} \cdot \hat{\sigma}_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2}.$$

 Otherwise, we set $r_j^{(k+1)} = 0$.

 Fix $r_j^{(k+1)}$, update $\theta^{(k)}$ by minimizing the following objective function:

$$\theta^{(k+1)} = \arg \min_{\theta \in \mathbb{R}} \frac{1}{2} \sum_{j \in \mathcal{S}} \frac{(\hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_j} - r_j^{(k+1)})^2}{\sigma_{Y_j}^2} - \frac{1}{2} \sum_{j \in \mathcal{S}} \frac{\theta^2 \cdot \hat{\sigma}_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2} + \sum_{j \in \mathcal{S}} \left(\gamma + \frac{1}{2} \frac{\theta^2 \cdot \hat{\sigma}_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2} \right) |r_j^{(k+1)}|.$$

$$\theta^{(k+1)} = \frac{\sum_{j \in \mathcal{S}} \frac{\hat{\beta}_{X_j, \text{RB}} (\hat{\beta}_{Y_j} - r_j^{(k+1)})}{\hat{\sigma}_{X_j, \text{RB}}^2}}{\sum_{j \in \mathcal{S}} \frac{\hat{\beta}_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2} - \frac{\hat{\sigma}_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2} \cdot (1 - |r_j^{(k+1)}|)}.$$

If $\left| \frac{\theta^{(k+1)} - \theta^{(k)}}{\theta^{(k)}} \right| < 10^{-7}$ **then** Stop and output $\hat{\theta}(\gamma) = \theta^{(k+1)}$ and $\hat{r}_j(\gamma) = r_j^{(k+1)}$;

else Set $k = k + 1$;

until $\left| \frac{\theta^{(k+1)} - \theta^{(k)}}{\theta^{(k)}} \right| < 10^{-7}$;

end

Valid IV Selection via BIC

for *all candidate* γ **do**

 Calculate

$$\text{BIC}(\gamma) = -2\hat{l}(\hat{\theta}(\gamma), \{\hat{r}_j(\gamma)\}_{j \in \hat{\mathcal{V}}_\gamma}) + \log(n) \cdot (s_\lambda - \hat{v}_\gamma);$$

end

 Select $\hat{\mathcal{V}}_\gamma$ with the smallest $\text{BIC}(\gamma)$;

end

S.3 Theorectical justifications for two l_1 methods

S.3.1 Method 1

For a fixed tuning parameter γ , we estimate the parameters by minimizing:

$$\min_{\theta \in \mathbb{R}, r_j \in \mathbb{R}} \hat{l}(\theta, \{r_j\}_{j \in \mathcal{S}}, \gamma).$$

To solve this optimization problem, we alternate between minimizing with respect to θ and $\{r_j\}_{j \in \mathcal{S}}$.

The optimal solution for $\{r_j\}_{j \in \mathcal{S}}$ given a fixed θ is:

$$r_j = \begin{cases} \text{sign}(\hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_j, \text{RB}}) \cdot (|\hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_j, \text{RB}}| - \gamma \cdot \sigma_{Y_j}^2) & \text{if } |\hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_j, \text{RB}}| > \gamma \cdot \sigma_{Y_j}^2 + |\theta| \cdot \sigma_{X_j, \text{RB}}, \\ 0 & \text{otherwise.} \end{cases}$$

To see this, we investigate $l_j(\theta, r_j)$ and discuss the solution of r_j when fixed θ . We consider the objective function:

$$l(\theta, \{r_j\}_{j \in \mathcal{S}}) = \sum_{j \in \mathcal{S}} l_j(\theta, r_j), \quad l_j(\theta, r_j) \triangleq \frac{1}{2} \frac{(\hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_j} - r_j)^2}{\sigma_{Y_j}^2} - \frac{1}{2} \frac{\theta^2 \cdot \sigma_{X_j}^2}{\sigma_{Y_j}^2} \mathbf{1}_{(r_j=0)} + \gamma \cdot |r_j|.$$

and we have

$$\frac{\partial l_j(\theta, r_j)}{\partial r_j} = -\frac{\hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_j, \text{RB}} - r_j}{\sigma_{Y_j}^2} + \gamma \quad \text{When } r_j > 0,$$

$$\frac{\partial l_j(\theta, r_j)}{\partial r_j} = -\frac{\hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_j, \text{RB}} - r_j}{\sigma_{Y_j}^2} - \gamma \quad \text{When } r_j < 0,$$

$$l_j(\theta, r_j) \triangleq \frac{1}{2} \frac{(\hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_j, \text{RB}})^2}{\sigma_{Y_j}^2} - \frac{1}{2} \frac{\theta^2 \cdot \sigma_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2} \quad \text{When } r_j = 0,$$

and consider three different scenarios:

- In the case that $\hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_j, \text{RB}} > \gamma \cdot \sigma_{Y_j}^2$,
when $r_j < 0$, we have $\frac{\partial l_j(\theta, r_j)}{\partial r_j} < 0$, and therefore

$$l_j(\theta, r_j) \geq \lim_{r_j \rightarrow 0^-} \frac{1}{2} \frac{(\hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_j, \text{RB}} - r_j)^2}{\sigma_{Y_j}^2} = \frac{1}{2} \frac{(\hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_j, \text{RB}})^2}{\sigma_{Y_j}^2}.$$

When $r_j = 0$,

$$l_j(\theta, r_j) = \frac{1}{2} \frac{(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}})^2}{\sigma_{Y_j}^2} - \frac{1}{2} \frac{\theta^2 \cdot \sigma_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2}.$$

When $r_j > 0$, we have $l_j(\theta, r_j)$ reach its local minimum when $r_j = \widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}} - \gamma \cdot \sigma_{Y_j}^2$.

$$l_j(\theta, r_j) = \frac{1}{2} \frac{\gamma^2 \cdot \sigma_{Y_j}^4}{\sigma_{Y_j}^2} + \gamma \cdot (\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}} - \gamma \cdot \sigma_{Y_j}^2).$$

and

$$\begin{aligned} & l_j(\theta, 0) - l_j(\theta, r_j) \\ &= \frac{1}{2} \frac{(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}})^2}{\sigma_{Y_j}^2} - \frac{1}{2} \frac{\theta^2 \cdot \sigma_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2} - \frac{1}{2} \frac{\gamma^2 \cdot \sigma_{Y_j}^4}{\sigma_{Y_j}^2} - \gamma \cdot (\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}} - \gamma \cdot \sigma_{Y_j}^2) \\ &= \frac{1}{2} \frac{(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}})^2}{\sigma_{Y_j}^2} + \frac{1}{2} \frac{\gamma^2 \cdot \sigma_{Y_j}^4}{\sigma_{Y_j}^2} - \gamma \cdot (\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}}) - \frac{1}{2} \frac{\theta^2 \cdot \sigma_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2} \\ &= \frac{1}{2} \frac{(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}} - \gamma \cdot \sigma_{Y_j}^2)^2}{\sigma_{Y_j}^2} - \frac{1}{2} \frac{\theta^2 \cdot \sigma_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2}. \end{aligned}$$

Thus when $\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}} - \gamma \cdot \sigma_{Y_j}^2 > |\theta| \cdot \sigma_{X_j, \text{RB}}$, $l_j(\theta, r_j)$ achieves the minimum when

$$r_j = \widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}} - \gamma \cdot \sigma_{Y_j}^2.$$

Otherwise, $l_j(\theta, r_j)$ achieves the minimum when $r_j = 0$.

- In the case that $\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}} < -\gamma \cdot \sigma_{Y_j}^2$,

when $r_j > 0$, we have $\frac{\partial l_j(\theta, r_j)}{\partial r_j} > 0$, and therefore

$$l_j(\theta, r_j) \geq \lim_{r_j \rightarrow 0^+} \frac{1}{2} \frac{(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}} - r_j)^2}{\sigma_{Y_j}^2} = \frac{1}{2} \frac{(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}})^2}{\sigma_{Y_j}^2}.$$

When $r_j = 0$,

$$l_j(\theta, r_j) = \frac{1}{2} \frac{(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}})^2}{\sigma_{Y_j}^2} - \frac{1}{2} \frac{\theta^2 \cdot \sigma_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2}.$$

When $r_j < 0$, we have $l_j(\theta, r_j)$ reach its local minimum when $r_j = \widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}} + \gamma \cdot \sigma_{Y_j}^2$.

$$l_j(\theta, r_j) = \frac{1}{2} \frac{\gamma^2 \cdot \sigma_{Y_j}^4}{\sigma_{Y_j}^2} - \gamma \cdot (\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}} + \gamma \cdot \sigma_{Y_j}^2).$$

and

$$\begin{aligned} l_j(\theta, 0) - l_j(\theta, r_j) &= \frac{1}{2} \frac{(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}})^2}{\sigma_{Y_j}^2} - \frac{1}{2} \frac{\theta^2 \cdot \sigma_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2} - \frac{1}{2} \frac{\gamma^2 \cdot \sigma_{Y_j}^4}{\sigma_{Y_j}^2} + \gamma \cdot (\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}} + \gamma \cdot \sigma_{Y_j}^2) \\ &= \frac{1}{2} \frac{(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}})^2}{\sigma_{Y_j}^2} + \frac{1}{2} \frac{\gamma^2 \cdot \sigma_{Y_j}^4}{\sigma_{Y_j}^2} + \gamma \cdot (\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}}) - \frac{1}{2} \frac{\theta^2 \cdot \sigma_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2} \\ &= \frac{1}{2} \frac{(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}} + \gamma \cdot \sigma_{Y_j}^2)^2}{\sigma_{Y_j}^2} - \frac{1}{2} \frac{\theta^2 \cdot \sigma_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2}. \end{aligned}$$

Thus when $\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}} + \gamma \cdot \sigma_{Y_j}^2 < -|\theta| \cdot \sigma_{X_j, \text{RB}}$, $l_j(\theta, r_j)$ achieves the minimum when

$$r_j = \widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}} + \gamma \cdot \sigma_{Y_j}^2.$$

Otherwise, $l_j(\theta, r_j)$ achieves the minimum when $r_j = 0$.

- In the case that $-\gamma \cdot \sigma_{Y_j}^2 \leq \widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j, \text{RB}} \leq \gamma \cdot \sigma_{Y_j}^2$, $l_j(\theta, r_j)$ achieves the minimum when $r_j = 0$.

S.3.2 Method 2

We consider the objective function

$$l(\theta, \{r_j\}_{j \in \mathcal{S}}) \triangleq \sum_{j \in \mathcal{S}} l_j(\theta, r_j),$$

where each component loss is given by

$$l_j(\theta, r_j) = \frac{1}{2} \frac{(\hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_j} - r_j)^2}{\sigma_{Y_j}^2} - \frac{1}{2} \frac{\theta^2 \cdot \sigma_{X_j}^2}{\sigma_{Y_j}^2} + (\gamma + \frac{1}{2} \frac{\theta^2 \cdot \sigma_{X_j}^2}{\sigma_{Y_j}^2}) |r_j|.$$

When θ is fixed, the optimization over r_j reduces to minimizing:

$$\tilde{l}_j(\theta, r_j) = \frac{1}{2} \frac{(\hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_j} - r_j)^2}{\sigma_{Y_j}^2} + \tilde{\lambda}_j |r_j|,$$

$$\text{with } \tilde{\lambda}_j = \gamma + \frac{1}{2} \frac{\theta^2 \cdot \sigma_{X_j}^2}{\sigma_{Y_j}^2}.$$

This takes the canonical form of the Lasso problem

$$\min_{r \in \mathbb{R}} \left\{ \frac{1}{2} (z - r)^2 + \lambda |r| \right\}.$$

Letting

$$a_j = \hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_j}, \quad \lambda_j = \tilde{\lambda}_j \cdot \sigma_{Y_j}^2,$$

we have:

$$\tilde{l}_j(r_j) = \frac{1}{2\sigma_{Y_j}^2} (a_j - r_j)^2 + \tilde{\lambda}_j |r_j| = \frac{1}{\sigma_{Y_j}^2} \left(\frac{1}{2} (a_j - r_j)^2 + \lambda_j |r_j| \right).$$

The minimizer of this expression is given by the soft-thresholding operator [12],

$$r_j^* = S_{\lambda_j}(a_j) = \begin{cases} \text{sign}(a_j) \cdot (|a_j| - \lambda_j), & \text{if } |a_j| > \lambda_j, \\ 0, & \text{otherwise.} \end{cases}$$

Substituting back, we obtain:

$$r_j = \begin{cases} \text{sign}(\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j}) \cdot \left(|\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j}| - \tilde{\lambda}_j \cdot \sigma_{Y_j}^2 \right), & \text{if } |\widehat{\beta}_{Y_j} - \theta \cdot \widehat{\beta}_{X_j}| > \tilde{\lambda}_j \cdot \sigma_{Y_j}^2, \\ 0, & \text{otherwise.} \end{cases}$$

where $\tilde{\lambda}_j = \gamma + \frac{1}{2} \frac{\theta^2 \cdot \sigma_{X_j}^2}{\sigma_{Y_j}^2}$ for all $j \in \mathcal{S}$.

S.4 Proof of Theorem 1

S.4.1 Notions and Assumptions

We first review some notions and assumptions that will be used in our proofs:

- The selected set of relevant IVs after randomization:

$$\mathcal{S}_\lambda = \left\{ j : \left| \frac{\widehat{\beta}_{X_j}}{\sigma_{X_j}} + Z_j \right| > \lambda, j = 1, \dots, p \right\}.$$

- Cardinality of the set of selected relevant IVs: $s_\lambda = |\mathcal{S}_\lambda|$.
- The average measure of instrument strength after selection:

$$\kappa_\lambda = \frac{1}{s_\lambda} \sum_{j \in \mathcal{S}_\lambda} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}.$$

- In our bagging strategy, we denote the b -th bootstrap sample as $\mathcal{S}_{\lambda,b}^*$ and the number of occurrences in $\mathcal{S}_{\lambda,b}^*$ for j -th IVs of \mathcal{S}_λ as w_{jb}^* . We also denote the selected set of valid IVs as

$$\widehat{\mathcal{V}}_b = \{j : \widehat{r}_{jb} = 0 \text{ and } j \in \mathcal{S}_{\lambda,b}^*\}$$

and the causal estimator as

$$\widehat{\theta}_b = A_b^{-1} \sum_{j \in \widehat{\mathcal{V}}_b} \widehat{\beta}_{Y_j} \widehat{\beta}_{X_j, \text{RB}} / \sigma_{Y_j}^2,$$

where

$$A_b = \sum_{j \in \widehat{\mathcal{V}}_b} (\widehat{\beta}_{X_j, \text{RB}}^2 - \widehat{\sigma}_{X_j, \text{RB}}^2) / \sigma_{Y_j}^2.$$

- For convenience, we also denote the conditional expectation taken with respect to bootstrap resampling as

$$\mathbb{E}^*[\cdot] = \mathbb{E}\left[\cdot \mid S_\lambda, \{(\widehat{\beta}_{Y_j}, \widehat{\beta}_{X_j, \text{RB}})\}_{j \in S_\lambda}\right].$$

Our final estimator is obtained by taking bootstrap aggregation

$$\tilde{\theta} = \frac{1}{B} \sum_{b=1}^B \hat{\theta}_b.$$

Assumption S1 (Measurement error model) (i) For any $j \neq j'$, $(\hat{\beta}_{Y_j}, \hat{\beta}_{X_j})$ and $(\hat{\beta}_{Y_{j'}}, \hat{\beta}_{X_{j'}})$ are mutually independent.

(ii) For each j , the association pair $(\hat{\beta}_{Y_j}, \hat{\beta}_{X_j})$ follows

$$\begin{bmatrix} \hat{\beta}_{X_j} \\ \hat{\beta}_{Y_j} \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \beta_{X_j} \\ \theta \beta_{X_j} + r_j \end{bmatrix}, \begin{bmatrix} \sigma_{X_j}^2 & 0 \\ 0 & \sigma_{Y_j}^2 \end{bmatrix} \right).$$

Furthermore, there exists positive constants l and u such that $\frac{m}{n} \leq \sigma_{X_j}^2 \leq \frac{M}{n}$, $\frac{m}{n} \leq \sigma_{Y_j}^2 \leq \frac{M}{n}$ for $j = 1, \dots, p$.

Assumption S2 (Variance stabilization) There exists a variance stabilizing quantity a_λ and a vector $\tau \in \mathbb{R}^{s_\lambda}$ in which each component is independent of $\{(u_j, \nu_j)\}_{j \in S_\lambda}$ and uniformly bounded away from infinity in probability in the sense that

$$\sup_{j \in S_\lambda} \left| a_\lambda \cdot \mathbb{E}^* \left[A_b^{-1} \cdot \hat{w}_{jb} \right] - \tau_j \right| = o_p(1),$$

where $A_b = \sum_{k \in S_\lambda} \hat{w}_{kb} \cdot (\hat{\beta}_{X_k, \text{RB}}^2 - \hat{\sigma}_{X_k, \text{RB}}^2) / \sigma_{Y_k}^2$, and

$$\hat{w}_{jb} = \begin{cases} w_{jb}^* \cdot \text{I}(\hat{r}_{jb} = 0) & \text{if } w_{jb}^* \geq 1, \\ 0 & \text{if } w_{jb}^* = 0. \end{cases}$$

In addition, there is no dominating instrument in the sense that

$$\frac{\max_{j \in S_\lambda} \beta_{X_j}^2}{\sum_{j \in S_\lambda} \beta_{X_j}^2} \xrightarrow{\text{P}} 0.$$

Assumption S3 (Negligible invalid IV induced bias) There is negligible bias induced by

potential imperfect screening of invalid IVs after bootstrap aggregation in the sense that

$$\frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \mathbb{E}^* \left[A_b^{-1} \sum_{j \in S_\lambda} \widehat{\beta}_{X_{j,\text{RB}}} \cdot r_j \cdot \widehat{w}_{jb} / \sigma_{Y_j}^2 \right] = o_p(1).$$

Assumption S4 (Instrument Selection) Define $\underline{\eta} = \min_{1 \leq j \leq p} \eta_j$ and $\bar{\eta} = \max_{1 \leq j \leq p} \eta_j$, then both $\underline{\eta}$ and $\bar{\eta}$ are bounded and bounded away from zero.

S.4.2 Proof

We begin by decomposing $\frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} (\tilde{\theta} - \theta_0)$ and want to show that there is a leading term in the decomposition converging to a Gaussian distribution. While the remained terms converges to zero in probability. We notice that

$$\frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} (\tilde{\theta} - \theta_0) = \frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \cdot \mathbb{E}^* \left[\widehat{\theta}_b - \theta_0 \right].$$

Here

$$\begin{aligned} \widehat{\theta}_b - \theta_0 &= A_b^{-1} \left\{ \sum_{j \in \widehat{V}_b} \tilde{u}_j / \sigma_{Y_j}^2 + \sum_{j \in \widehat{V}_b} \widehat{\beta}_{X_{j,\text{RB}}} \cdot r_j / \sigma_{Y_j}^2 \right\} \\ &= A_b^{-1} \left\{ \sum_{j \in S_\lambda} \widehat{w}_{jb} \cdot \tilde{u}_j / \sigma_{Y_j}^2 + \sum_{j \in S_\lambda} \widehat{w}_{jb} \cdot \widehat{\beta}_{X_{j,\text{RB}}} \cdot r_j / \sigma_{Y_j}^2 \right\}. \end{aligned}$$

where $\tilde{u}_j = \beta_{X_j} \cdot (\nu_j - \theta_0 \cdot u_j) + (\nu_j \cdot u_j - \theta_0 \cdot (u_j^2 - \widehat{\sigma}_{X_{j,\text{RB}}}^2))$ and

$$\widehat{w}_{jb} = \begin{cases} w_{jb}^* \cdot \mathbb{I}(\widehat{r}_{jb} = 0) & \text{if } w_{jb}^* \geq 1, \\ 0 & \text{if } w_{jb}^* = 0. \end{cases}$$

We then can decompose $\frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} (\tilde{\theta} - \theta_0)$ into two terms:

$$\frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} (\tilde{\theta} - \theta_0) = \underbrace{\frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \sum_{j \in S_\lambda} \mathbb{E}^* \left[A_b^{-1} \cdot \widehat{w}_{jb} \right] \cdot \tilde{u}_j / \sigma_{Y_j}^2}_{(I)} + \underbrace{\frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \mathbb{E}^* \left[A_b^{-1} \sum_{j \in S_\lambda} \widehat{w}_{jb} \cdot \widehat{\beta}_{X_{j,\text{RB}}} \cdot r_j / \sigma_{Y_j}^2 \right]}_{(II)}.$$

Assumption S3 shows that the second term in the above formula satisfies $(II) = o_p(1)$.

For the term (I), we further decompose it as

$$\begin{aligned}
& \frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \sum_{j \in \mathcal{S}_\lambda} \mathbb{E}^* \left[A_b^{-1} \cdot \hat{w}_{jb} \right] \cdot \tilde{u}_j / \sigma_{Y_j}^2 \\
&= \underbrace{\frac{1}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \sum_{j \in \mathcal{S}_\lambda} \tau_j \cdot \tilde{u}_j / \sigma_{Y_j}^2}_{(I.1)} \\
&+ \underbrace{\frac{1}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \sum_{j \in \mathcal{S}_\lambda} \left\{ a_\lambda \cdot \mathbb{E}^* \left[A_b^{-1} \cdot \hat{w}_{jb} \right] - \tau_j \right\} \cdot \tilde{u}_j / \sigma_{Y_j}^2}_{(I.2)}.
\end{aligned}$$

Here (I.2) has the following upper bounds,

$$(I.2) \leq \sup_{j \in \mathcal{S}_\lambda} \left| a_\lambda \cdot \mathbb{E}^* \left[A_b^{-1} \cdot \hat{w}_{jb} \right] - \tau_j \right| \cdot \frac{1}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \sum_{j \in \mathcal{S}_\lambda} \cdot \tilde{u}_j / \sigma_{Y_j}^2.$$

Under Assumption S2, we can prove $(I.2) = o_p(1)$.

Combining all the above results, we have

$$\frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} (\tilde{\theta} - \theta_0) = \frac{1}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \sum_{j \in \mathcal{S}_\lambda} \tau_j \cdot \tilde{u}_j + o_p(1).$$

Using the proof of Theorem 1 in [33], we can show that when Assumption S1 and Assumption S4 hold and $\frac{\max_{j \in \mathcal{S}_\lambda} \beta_{X_j}^2}{\sum_{j \in \mathcal{S}_\lambda} \beta_{X_j}^2} \xrightarrow{P} 0$, conditional on the selection event \mathcal{S}_λ , $\frac{1}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \sum_{j \in \mathcal{S}_\lambda} \tau_j \cdot \tilde{u}_j$ converges to a Gaussian distribution as $s_\lambda \xrightarrow{P} \infty$ and $\frac{\kappa_\lambda}{\lambda^2} \xrightarrow{P} \infty$.

Therefore, we can conclude that $\frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} (\tilde{\theta} - \theta_0)$ converges to a Gaussian distribution.

S.4.3 Verifying the Assumption S2 in the case with perfect screening property

To cast more insights into Assumption S2, we next consider a special case where perfect IV screening is achieved. In the case of perfect IV screening, we have

$$A_b = \sum_{k \in \mathcal{V}_\lambda} w_{kb}^* \cdot (\hat{\beta}_{X_k, \text{RB}}^2 - \hat{\sigma}_{X_k, \text{RB}}^2) / \sigma_{Y_j}^2.$$

In what follows, we argue that Assumption S2 holds for both valid and invalid IVs in \mathcal{S}_λ :

- For valid IVs in \mathcal{V}_λ (\mathcal{V}_λ is the collection of all valid IVs in \mathcal{S}_λ), we define

$$\tau_j = a_\lambda \cdot \mathbb{E}^* \left[\frac{w_{jb}^*}{\sum_{k \in \mathcal{V}_\lambda} w_{kb}^* \cdot \beta_{X_k}^2 / \sigma_{Y_k}^2} \right], \quad a_\lambda = \sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2,$$

and we have τ_j independent of $\{(u_j, \nu_j)\}_{j \in \mathcal{S}_\lambda}$. In this context, we have $\widehat{w}_{jb} = w_{jb}^*$ and can show that

$$\left| a_\lambda \cdot \mathbb{E}^* \left[A_b^{-1} \cdot \widehat{w}_{jb} \right] - \tau_j \right| = \left| a_\lambda \cdot \mathbb{E}^* \left[A_b^{-1} \cdot w_{jb}^* \right] - \tau_j \right| = o_p(1).$$

For this bound to hold uniformly for $j \in \mathcal{V}_\lambda$ as stated in the assumption, given that w_{jb}^* follows a multinomial distribution with an equal mean, we conjecture that this condition is likely to hold as long as A_b converges to a center that is independent of j . In fact, under appropriate conditions (See Section S.4.4 in the Supplement Material for full theoretical justifications), we can show that,

$$A_b = \sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2 \cdot (1 + o_p(1)),$$

which is indeed independent of j .

- For invalid IV $j \in \mathcal{S}_\lambda / \mathcal{V}_\lambda$, under perfect screening property, we have $\widehat{r}_{jb} = 0$ and therefore $\widehat{w}_{jb} = 0$. Set $\tau_j = 0$ for $j \in \mathcal{S}_\lambda / \mathcal{V}_\lambda$, we have

$$\sup_{j \in \mathcal{S}_\lambda / \mathcal{V}_\lambda} \left| a_\lambda \cdot \mathbb{E}^* \left[A_b^{-1} \cdot \widehat{w}_{jb} \right] - \tau_j \right| = o_p(1).$$

Combining these two parts of results, we can verify that the Assumption S2 is satisfied.

S.4.4 The asymptotic analysis of A_b under perfect screening property

Notice that

$$A_b = \sum_{k \in \mathcal{V}_\lambda} w_{kb}^* \cdot (\widehat{\beta}_{X_k, \text{RB}}^2 - \widehat{\sigma}_{X_k, \text{RB}}^2) / \sigma_{Y_k}^2.$$

We want to show $A_b = \sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2 \cdot (1 + o_p(1))$ under these two conditions

$$\frac{\max_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2}{\sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2} \rightarrow 0 \text{ and } \frac{v_\lambda \cdot \max_{k \in \mathcal{V}_\lambda} |(\widehat{\beta}_{X_k}^2 - \widehat{\sigma}_{X_k, \text{RB}}^2) / \sigma_{Y_k}^2 - \beta_{X_k}^2 / \sigma_{Y_k}^2|}{\sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2} = o_p(1).$$

To prove this result, we begin with the following decomposition,

$$\begin{aligned} A_b - \sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2 &= \sum_{k \in \mathcal{V}_\lambda} w_{kb}^* \cdot (\widehat{\beta}_{X_k}^2 - \widehat{\sigma}_{X_k, \text{RB}}^2) / \sigma_{Y_k}^2 - \sum_{k \in \mathcal{V}_\lambda} w_{kb}^* \cdot \beta_{X_k}^2 / \sigma_{Y_k}^2 + \sum_{k \in \mathcal{V}_\lambda} w_{kb}^* \cdot \beta_{X_k}^2 / \sigma_{Y_k}^2 - \sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2 \\ &= \sum_{k \in \mathcal{V}_\lambda} w_{kb}^* \cdot (\widehat{\beta}_{X_k}^2 - \widehat{\sigma}_{X_k, \text{RB}}^2) / \sigma_{Y_k}^2 - \sum_{k \in \mathcal{V}_\lambda} w_{kb}^* \cdot \beta_{X_k}^2 / \sigma_{Y_k}^2 + \sum_{k \in \mathcal{V}_\lambda} (w_{kb}^* - 1) \cdot \beta_{X_k}^2 / \sigma_{Y_k}^2 \\ &= \sum_{k \in \mathcal{V}_\lambda} w_{kb}^* \cdot \left((\widehat{\beta}_{X_k}^2 - \widehat{\sigma}_{X_k, \text{RB}}^2) / \sigma_{Y_k}^2 - \beta_{X_k}^2 / \sigma_{Y_k}^2 \right) + \sum_{k \in \mathcal{V}_\lambda} (w_{kb}^* - 1) \cdot \beta_{X_k}^2 / \sigma_{Y_k}^2. \end{aligned}$$

It suffices to prove the two terms on the right-hand side are of the asymptotic order $o_p(\sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2)$.

Notice that $[w_{1,b}^*, \dots, w_{s_\lambda, b}^*]$ follows a multinomial distribution with $\mathbb{E}[w_{k,b}^*] = 1$ for all $k \in \mathcal{S}_\lambda$ and

$$\begin{aligned} \text{Var}[w_{k,b}^*] &= \frac{1}{s_\lambda} \cdot \left(1 - \frac{1}{s_\lambda}\right), \text{ for all } k \in \mathcal{S}_\lambda, \\ \text{Cov}(w_{i,b}^*, w_{j,b}^*) &= -\frac{1}{s_\lambda} \text{ for all } i, j \in \mathcal{S}_\lambda \text{ such that } i \neq j. \end{aligned}$$

- To show $\frac{\sum_{k \in \mathcal{V}_\lambda} (w_{kb}^* - 1) \cdot \beta_{X_k}^2 / \sigma_{Y_k}^2}{\sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2} = o_p(1)$,
we have $\sum_{k \in \mathcal{V}_\lambda} (w_{kb}^* - 1) \cdot \beta_{X_k}^2 / \sigma_{Y_k}^2 = O_p(\sqrt{\text{Var}[\sum_{k \in \mathcal{V}_\lambda} (w_{kb}^* - 1) \cdot \beta_{X_k}^2 / \sigma_{Y_k}^2]})$ and

$$\begin{aligned} \text{Var}[\sum_{k \in \mathcal{V}_\lambda} (w_{kb}^* - 1) \cdot \beta_{X_k}^2 / \sigma_{Y_k}^2] &= \sum_{k \in \mathcal{V}_\lambda} \left(1 - \frac{1}{s_\lambda}\right) \cdot (\beta_{X_k}^2 / \sigma_{Y_k}^2)^2 - \frac{1}{s_\lambda} \sum_{i,j \in \mathcal{V}_\lambda, i \neq j} (\beta_{X_i}^2 / \sigma_{Y_i}^2) \cdot (\beta_{X_j}^2 / \sigma_{Y_j}^2) \\ &= \sum_{k \in \mathcal{V}_\lambda} (\beta_{X_j}^2 / \sigma_{Y_k}^2)^2 - \frac{1}{s_\lambda} \left(\sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2\right)^2. \end{aligned}$$

Notice that

$$\begin{aligned}
\frac{Var[\sum_{k \in \mathcal{V}_\lambda} (w_{kb}^* - 1) \cdot \beta_{X_k}^2 / \sigma_{Y_k}^2]}{(\sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2)^2} &= \frac{\sum_{k \in \mathcal{V}_\lambda} (\beta_{X_k}^2 / \sigma_{Y_k}^2)^2 - \frac{1}{s_\lambda} (\sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2)^2}{(\sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2)^2} \\
&= \frac{\sum_{k \in \mathcal{V}_\lambda} (\beta_{X_k}^2 / \sigma_{Y_k}^2)^2}{(\sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2)^2} - \frac{1}{s_\lambda} \\
&\leq \frac{\max_{j \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2 \cdot \sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2}{(\sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2)^2} - \frac{1}{s_\lambda} \\
&= \frac{\max_{j \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2}{\sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2} - \frac{1}{s_\lambda},
\end{aligned}$$

if we have

$$\frac{\max_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2}{\sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2} \rightarrow 0,$$

$$\frac{\sum_{k \in \mathcal{V}_\lambda} (w_{kb}^* - 1) \cdot \beta_{X_k}^2 / \sigma_{Y_k}^2}{\sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2} = o_p(1) \text{ directly follows.}$$

- To show $\frac{\sum_{k \in \mathcal{V}_\lambda} w_{kb}^* \cdot ((\hat{\beta}_{X_k}^2 - \hat{\sigma}_{X_k, \text{RB}}^2) / \sigma_{Y_k}^2 - \beta_{X_k}^2 / \sigma_{Y_k}^2)}{\sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2} = o_p(1)$, we further decompose it into two terms

$$\frac{\sum_{k \in \mathcal{V}_\lambda} (w_{kb}^* - 1) \cdot ((\hat{\beta}_{X_k}^2 - \hat{\sigma}_{X_k, \text{RB}}^2) / \sigma_{Y_k}^2 - \beta_{X_k}^2 / \sigma_{Y_k}^2)}{\sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2} + \frac{\sum_{k \in \mathcal{V}_\lambda} ((\hat{\beta}_{X_k}^2 - \hat{\sigma}_{X_k, \text{RB}}^2) / \sigma_{Y_k}^2 - \beta_{X_k}^2 / \sigma_{Y_k}^2)}{\sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2}.$$

$\frac{\sum_{k \in \mathcal{V}_\lambda} ((\hat{\beta}_{X_k}^2 - \hat{\sigma}_{X_k, \text{RB}}^2) / \sigma_{Y_k}^2 - \beta_{X_k}^2 / \sigma_{Y_k}^2)}{\sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2} = o_p(1)$ can directly follow from Lemma S.13 of the Supplemental Material of [33].

To prove $\frac{\sum_{k \in \mathcal{V}_\lambda} (w_{kb}^* - 1) \cdot ((\hat{\beta}_{X_k}^2 - \hat{\sigma}_{X_k, \text{RB}}^2) / \sigma_{Y_k}^2 - \beta_{X_k}^2 / \sigma_{Y_k}^2)}{\sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2} = o_p(1)$, we use

$$\begin{aligned}
&\frac{\sum_{k \in \mathcal{V}_\lambda} (w_{kb}^* - 1) \cdot ((\hat{\beta}_{X_k}^2 - \hat{\sigma}_{X_k, \text{RB}}^2) / \sigma_{Y_k}^2 - \beta_{X_k}^2 / \sigma_{Y_k}^2)}{\sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2} \\
&= O_p\left(\frac{\mathbb{E}|\sum_{k \in \mathcal{V}_\lambda} (w_{kb}^* - 1) \cdot ((\hat{\beta}_{X_k}^2 - \hat{\sigma}_{X_k, \text{RB}}^2) / \sigma_{Y_k}^2 - \beta_{X_k}^2 / \sigma_{Y_k}^2)|}{\sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2}\right).
\end{aligned}$$

Notice that

$$\begin{aligned} & \mathbb{E} \left| \sum_{k \in \mathcal{V}_\lambda} (w_{kb}^* - 1) \cdot \left((\widehat{\beta}_{X_k}^2 - \widehat{\sigma}_{X_k, \text{RB}}^2) / \sigma_{Y_k}^2 - \beta_{X_k}^2 / \sigma_{Y_k}^2 \right) \right| \\ & \leq \max_{j \in \mathcal{V}_\lambda} \left| (\widehat{\beta}_{X_k}^2 - \widehat{\sigma}_{X_k, \text{RB}}^2) / \sigma_{Y_k}^2 - \beta_{X_k}^2 / \sigma_{Y_k}^2 \right| \cdot \sum_{k \in \mathcal{V}_\lambda} \mathbb{E} |w_{kb}^* - 1|, \\ & \text{and } \mathbb{E} |w_{kb}^* - 1| = \mathbb{E} (w_{kb}^* - 1) + 2 \cdot \mathbb{P}(w_{kb}^* = 0) = 2 \cdot \left(1 - \frac{1}{s_\lambda}\right). \end{aligned}$$

If we have

$$\frac{v_\lambda \cdot \max_{k \in \mathcal{V}_\lambda} \left| (\widehat{\beta}_{X_k}^2 - \widehat{\sigma}_{X_k, \text{RB}}^2) / \sigma_{Y_k}^2 - \beta_{X_k}^2 / \sigma_{Y_k}^2 \right|}{\sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2} = o_p(1).$$

Then can show $\frac{\sum_{k \in \mathcal{V}_\lambda} (w_{kb}^* - 1) \cdot \left((\widehat{\beta}_{X_k}^2 - \widehat{\sigma}_{X_k, \text{RB}}^2) / \sigma_{Y_k}^2 - \beta_{X_k}^2 / \sigma_{Y_k}^2 \right)}{\sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2} = o_p(1)$ and therefore

$$\frac{\sum_{k \in \mathcal{V}_\lambda} w_{kb}^* \cdot \left((\widehat{\beta}_{X_k}^2 - \widehat{\sigma}_{X_k, \text{RB}}^2) / \sigma_{Y_k}^2 - \beta_{X_k}^2 / \sigma_{Y_k}^2 \right)}{\sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2} = o_p(1).$$

Combining all these results, we have $A_b = \sum_{k \in \mathcal{V}_\lambda} \beta_{X_k}^2 / \sigma_{Y_k}^2 \cdot (1 + o_p(1))$.

S.5 Invalid IV screening consistency

In this section, we show that under Conditions 1-7, the proposed invalid IV screening procedure is “nearly perfect” as s_λ goes to infinity.

S.5.1 Notations

We first introduce notations to be used in the sufficient conditions and our proofs below:

- The correct set of valid IVs in S_λ :

$$\mathcal{V}_\lambda = \{j \in \mathcal{S}_\lambda : \beta_{X_j} \neq 0 \text{ and } r_j = 0\}.$$

- Cardinality of the set of valid IVs: $v_\lambda = |\mathcal{V}_\lambda|$.

- The selected set of valid IVs:

$$\widehat{\mathcal{V}}_\lambda = \{j : \widehat{r}_j = 0 \text{ and } j \in \mathcal{S}_\lambda\}.$$

- Cardinality of the set of selected valid IVs: $\widehat{v}_\lambda = |\widehat{\mathcal{V}}_\lambda|$.
- For any $\mathcal{V} \subseteq \mathcal{S}_\lambda$, we use the following notation:
 - Cardinality of the set: $v = |\mathcal{V}|$.
 - The measure of average instrument strength of \mathcal{V} : $\kappa_\lambda(\mathcal{V}) = \frac{1}{v} \sum_{j \in \mathcal{V}} \frac{\beta_{Xj}^2}{\sigma_{Yj}^2}$.
 - The measure of average pleiotropic effects of \mathcal{V} : $r_\lambda(\mathcal{V}) = \frac{1}{v} \sum_{j \in \mathcal{V}} \frac{r_j^2}{\sigma_{Yj}^2}$.
 - Correlation between instrument strength and pleiotropic effects of \mathcal{V} when \mathcal{V} has at least one non-zero r_j :

$$\rho(\mathcal{V}) = \text{Corr}^2\left(\{\beta_{Xj}\}_{j \in \mathcal{V}}, \{r_j\}_{j \in \mathcal{V}}\right) = \frac{\left(\sum_{j \in \mathcal{V}} \frac{r_j \cdot \beta_{Xj}}{\sigma_{Yj}^2}\right)^2}{\sum_{j \in \mathcal{V}} \frac{r_j^2}{\sigma_{Yj}^2} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{Xj}^2}{\sigma_{Yj}^2}}.$$

To identify invalid IVs, we use the following measurement error models.

$$\widehat{\beta}_{Yj} = \theta \cdot \beta_{Xj} + r_j + \nu_j, \quad \widehat{\beta}_{X_{j,\text{RB}}} = \beta_{Xj} + u_j, \quad j \in \mathcal{S}_\lambda.$$

and let $n_1 = n_2 = n$ be the sample sizes of the two GWAS summary datasets for X and Y , respectively.

The invalid IV screening is obtained by solving

$$\min_{\theta \in \mathbb{R}, r_j} \widehat{l}(\theta, \{r_j\}_{j \in \mathcal{S}_\lambda}, \{\widehat{\beta}_{Yj}, \sigma_{Yj}, \widehat{\beta}_{X_{j,\text{RB}}}, \widehat{\sigma}_{X_{j,\text{RB}}}\}_{j \in \mathcal{S}_\lambda}), \text{ s.t. } \sum_{j \in \mathcal{S}_\lambda} \mathbf{1}_{r_j=0} = v.$$

where

$$\widehat{l}(\theta, \{r_j\}_{j \in \mathcal{S}_\lambda}, \{\widehat{\beta}_{Yj}, \sigma_{Yj}, \widehat{\beta}_{X_{j,\text{RB}}}, \widehat{\sigma}_{X_{j,\text{RB}}}\}_{j \in \mathcal{S}_\lambda}) = \sum_{j \in \mathcal{S}_\lambda} \frac{(\widehat{\beta}_{Yj} - \theta \cdot \widehat{\beta}_{X_{j,\text{RB}}} - r_j)^2}{\sigma_{Yj}^2} - \sum_{j \in \mathcal{S}_\lambda} \frac{\theta^2 \cdot \widehat{\sigma}_{X_{j,\text{RB}}^2}}{\sigma_{Yj}^2} \cdot \mathbf{1}_{r_j=0}.$$

Here $\mathbb{1}(\cdot)$ is the indicator function and v is a tuning parameter representing the unknown number of valid IVs. We propose a generalized Bayesian Information Criterion(GBIC) to select the best v :

$$GBIC(v) = \hat{l}(\hat{\theta}, \{\hat{r}_j\}_{j \in \mathcal{S}_\lambda}, \left\{ \hat{\beta}_{Y_j}, \sigma_{Y_j}, \hat{\beta}_{X_{j,\text{RB}}}, \hat{\sigma}_{X_{j,\text{RB}}} \right\}_{j \in \mathcal{S}_\lambda}) + \kappa_n \cdot (s_\lambda - v).$$

Then we select $\hat{v} = \arg \min_v GBIC(v)$ and estimate $\hat{\mathcal{V}}_\lambda = \{j : \hat{r}_{j,\hat{v}} = 0 \text{ and } j \in \mathcal{S}_\lambda\}$, which is the set of the estimated invalid IVs..

S.5.2 Sufficient conditions

Condition 1 (Bound of Orlicz norm) *Fix λ , $\frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}$ is a sub-exponential random variable for all $j \in \mathcal{S}_\lambda$ and we have $\|\frac{\nu_j}{\sigma_{Y_j}}\|_{\psi_2}^2$, $\|\frac{u_j}{\sigma_{Y_j}}\|_{\psi_2}^2$, $\|\sqrt{\frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2}}\|_{\psi_2}$, $\|\sqrt{\frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}}\|_{\psi_2}$, $\|\sqrt{\frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}}\|_{\psi_2}$ bounded away from ∞ uniformly for all $j \in \mathcal{S}_\lambda$.*

This condition is a technical condition. It places some restrictions on the tail distributions of the noise terms, aiming to ensure that they have good concentration behaviors.

Condition 2 (Orders of the variances and sample sizes) *There exist positive constants m and M such that we have $\frac{m}{n} \leq \sigma_{X_j}^2$, $\sigma_{Y_j}^2 \leq \frac{M}{n}$ for $j = 1, \dots, p$.*

In this condition, we require the variances of both $\hat{\beta}_{X_j}$ and $\hat{\beta}_{Y_j}$ have the orders $\frac{1}{n}$ uniformly for all $j \in \mathcal{S}_\lambda$, which is a normal assumption in two-sample summary Mendelian Randomization literature.

Condition 3 (Plurality and no perfect correlation) *For all $\mathcal{V} \subseteq \mathcal{S}_\lambda$ and \mathcal{V} contains at least one $r_j \neq 0$, whenever $\rho(\mathcal{V}) = 1$, we have $|\mathcal{V}_\lambda| > |\mathcal{V}|$; whenever $\rho(\mathcal{V}) < 1$, we have the correlation coefficient $\rho(\mathcal{V}) < 1$ is upper bounded by a constant c_0 smaller than one.*

Here $\rho(\mathcal{V})$ measures the correlation between instrument strength and pleiotropic effects of \mathcal{V} , which is defined as

$$\rho(\mathcal{V}) = \text{Corr}^2\left(\{\beta_{X_j}\}_{j \in \mathcal{V}}, \{r_j\}_{j \in \mathcal{V}}\right) = \frac{\left(\sum_{j \in \mathcal{V}} \frac{r_j \cdot \beta_{X_j}}{\sigma_{Y_j}^2}\right)^2}{\sum_{j \in \mathcal{V}} \frac{r_j^2}{\sigma_{Y_j}^2} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}}.$$

This condition is closely related to the plural validity assumption commonly made in two-sample summary data Mendelian Randomization literature [28, 20], which ensures the uniqueness and identifiability of the causal effect θ . By Cauchy-Schwartz inequality, we can see that $\rho(\mathcal{V}) = 1$ indicates that there exists a $c \in \mathbb{R}$ such that $\frac{r_j}{\beta_{X_j}} = c$ holds for all $j \in \mathcal{V}$. If the first part of this condition does not hold, there will be a \mathcal{V}^* such that $|\mathcal{V}^*| \geq |\mathcal{V}_\lambda|$ and $\frac{r_j}{\beta_{X_j}} = c > 0$ holds for all $j \in \mathcal{V}^*$. We expect that the invalid IV screening procedure will tend to screen out $\mathcal{S}_\lambda / \mathcal{V}^*$ and leave \mathcal{V}^* . Therefore, the sub-sequential causal estimation using \mathcal{V}^* will be centered around $\theta_0 + c$ instead of θ_0 , where θ_0 is the true causal effect. In this case, we fail to identify the true causal effect. Furthermore, the second part of this condition is to ensure that different clusters of IV set $\mathcal{V}_c = \{j \in \mathcal{S}_\lambda \mid \frac{r_j}{\beta_{X_j}} = c\}$ are sufficiently separable, so that there will not be a IV set $\mathcal{V}^* \neq \mathcal{V}_\lambda$ with $\rho(\mathcal{V}^*) \rightarrow 1$ selected by the invalid screening procedure. Without this, we might not be able to distinguish the IV set \mathcal{V}^* and \mathcal{V}_λ .

Condition 4 (Boundedness) *For any $\mathcal{V} \in \mathcal{S}_\lambda$, $|\widehat{\theta}(\mathcal{V})|$ is uniformly bounded away from ∞ with probability goes to 1.*

This condition requires that for any subset $\mathcal{V} \subseteq \mathcal{S}_\lambda$, the causal estimate

$$\widehat{\theta}(\mathcal{V}) = \frac{\sum_{j \in \mathcal{V}} \frac{\widehat{\beta}_{Y_j} \widehat{\beta}_{X_j, \text{RB}}}{\sigma_{Y_j}^2}}{\sum_{j \in \mathcal{V}} \frac{\widehat{\beta}_{X_j, \text{RB}}^2 - \widehat{\sigma}_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2}}$$

based on \mathcal{V} should not be too large. In fact, when the Condition 1 holds, this condition can be satisfied in the case that $\frac{r_j}{\beta_{X_j}}$ is bounded away from infinity for all $j \in \mathcal{S}_\lambda$ and β_{X_j} is sufficiently separated from 0 for all $j \in \mathcal{S}_\lambda$. To see this, we can decompose $\widehat{\theta}(\mathcal{V})$ as follows:

$$\begin{aligned} \widehat{\theta}(\mathcal{V}) &= \frac{\sum_{j \in \mathcal{V}} \frac{(\theta_0 \beta_{X_j} + r_j) \beta_{X_j}}{\sigma_{Y_j}^2} + \sum_{j \in \mathcal{V}} \frac{\beta_{X_j} \nu_j}{\sigma_{Y_j}^2} + \sum_{j \in \mathcal{V}} \frac{(\theta_0 \beta_{X_j} + r_j) u_j}{\sigma_{Y_j}^2} + \sum_{j \in \mathcal{V}} \frac{u_j \nu_j}{\sigma_{Y_j}^2}}{\sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2} + 2 \sum_{j \in \mathcal{V}} \frac{\beta_{X_j} u_j}{\sigma_{Y_j}^2} + \sum_{j \in \mathcal{V}} \frac{u_j^2 - \sigma_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2} - \sum_{j \in \mathcal{V}} \frac{\widehat{\sigma}_{X_j, \text{RB}}^2 - \sigma_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2}} \\ &= \frac{\sum_{j \in \mathcal{V}} (\theta_0 + \frac{r_j}{\beta_{X_j}}) \cdot \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2} + \sum_{j \in \mathcal{V}} \frac{\beta_{X_j} \nu_j}{\sigma_{Y_j}^2} + \sum_{j \in \mathcal{V}} (\theta_0 + \frac{r_j}{\beta_{X_j}}) \cdot \frac{\beta_{X_j} u_j}{\sigma_{Y_j}^2} + \sum_{j \in \mathcal{V}} \frac{u_j \nu_j}{\sigma_{Y_j}^2}}{\sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2} + 2 \sum_{j \in \mathcal{V}} \frac{\beta_{X_j} u_j}{\sigma_{Y_j}^2} + \sum_{j \in \mathcal{V}} \frac{u_j^2 - \sigma_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2} - \sum_{j \in \mathcal{V}} \frac{\widehat{\sigma}_{X_j, \text{RB}}^2 - \sigma_{X_j, \text{RB}}^2}{\sigma_{Y_j}^2}} \end{aligned}$$

If $\min_{j \in \mathcal{S}_\lambda} |\beta_{X_j}|$ is sufficiently separated from 0, with Condition 1, we can verify that there

exists a $c > 0$ such that the denominator is uniformly larger than $c \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}$ with probability going to one for any possible \mathcal{V} . Similarly, we can also show that when $\max_{j \in \mathcal{S}_\lambda} |\frac{r_j}{\beta_{X_j}}|$ is bounded away from infinity, there exists a $C > 0$ such that the numerator is bounded away from $C \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}$ with probability going to one for any possible \mathcal{V} . Therefore, we can verify that $|\hat{\theta}(\mathcal{V})|$ is uniformly bounded away from ∞ with probability going to one.

Condition 5 (Separation of $r_j \neq 0$ and 0)

$$\min_{j \in \mathcal{S}_\lambda, r_j \neq 0} r_j \gg \max \left\{ \frac{(s_\lambda \cdot \ln(s_\lambda))^{\frac{1}{2}}}{n^{\frac{1}{2}}}, \frac{\kappa_n^{\frac{1}{2}}}{n^{\frac{1}{2}}} \right\}.$$

In this condition, we require the pleiotropy effects $r_j \neq 0$ of invalid IVs to be bounded away from 0. In this case, our invalid IV screening procedure will be able to distinguish invalid IVs and valid IVs. This condition is similar to the "beta-min" condition in the high-dimension linear regression setting. The only difference is that it is made in r_j instead of the parameter of interest θ_0 . Without this condition, we will not be able to screen out some invalid IVs with r_j close to 0 in the invalid IV screening procedure, and the perfect screening property will not hold.

Condition 6 (The order of v_λ) *The number of valid IVs has the same order of s_λ . In other words, $\frac{v_\lambda}{s_\lambda}$ is bounded away from zero. There exists a constant c_1 such that $0 < c_1 < 1$. For all $\mathcal{V} \subseteq \mathcal{S}_\lambda$ containing at least one nonzero element $r_j \neq 0$, whenever $\rho(\mathcal{V}) = 1$, it holds that $v < c_1 \cdot v_\lambda$.*

The first part of this condition requires that the number of valid IVs in \mathcal{S}_λ should be sufficiently large. To be specific, it should be of the same order as the total number of IVs in \mathcal{S}_λ . This condition can be further weakened by adjusting the penalized coefficient of GBIC. The second part of this condition imposes constraints on the cardinality of the cluster $\{j \in \mathcal{S}_\lambda | \frac{r_j}{\beta_{X_j}} = c\}$. It requires for any $c \neq 0$, the cardinality of the IV clusters $\{j \in \mathcal{S}_\lambda | \frac{r_j}{\beta_{X_j}} = c\}$ should be sufficiently separated from the total number of valid IVs v_λ so that the algorithm will not fail to identify the true valid IV set \mathcal{V}_λ . If there exists a $c_0 \neq 0$ such that the cardinality of the IV clusters $\{j \in \mathcal{S}_\lambda | \frac{r_j}{\beta_{X_j}} = c_0\}$ is very close to v_λ , then the algorithm might fail to distinguish $\{j \in \mathcal{S}_\lambda | \frac{r_j}{\beta_{X_j}} = c_0\}$ and the valid IV set \mathcal{V}_λ .

Condition 7 (high dimension BIC)

$$\frac{\kappa_n}{n \cdot \min_{j \in \mathcal{S}_\lambda, r_j \neq 0} r_j^2} \rightarrow 0 \quad \text{and} \quad \kappa_n \gg \ln(s_\lambda).$$

Condition 7* (high dimension BIC)

$$\frac{\kappa_n}{n \cdot \min_{j \in \mathcal{S}_\lambda, r_j \neq 0} r_j^2} \rightarrow 0 \quad \text{and} \quad \kappa_n \gg s_\lambda \cdot \ln(s_\lambda).$$

S.5.3 Theoretical Results

Define a collection of set

$$\mathcal{V}_{\text{valid}} = \{\mathcal{V} | \mathcal{V} \subseteq \mathcal{S}_\lambda, r_j = 0 \text{ for all } j \in \mathcal{V}, \text{ and } |\mathcal{V}| \geq c_1 \cdot |\mathcal{V}_\lambda|\}.$$

Here c_1 is the constant that we introduce in Condition 6.

Theorem S1 *Under Condition 1-7, our IV screening procedure can consistently select the sets inside $\mathcal{V}_{\text{valid}}$. Mathematically, this property is expressed as:*

$$\mathbb{P}(\hat{\mathcal{V}}_\lambda \in \mathcal{V}_{\text{valid}}) \rightarrow 1, \text{ as } s_\lambda \rightarrow \infty.$$

where $\hat{\mathcal{V}}_\lambda$ represents the set of IVs selected by the screening procedure.

Theorem S2 *Under Condition 1-6 and Condition 7*, our IV screening procedure can consistently select the complete set of valid IVs \mathcal{V}_λ . Mathematically, this property is expressed as:*

$$\mathbb{P}(\hat{\mathcal{V}}_\lambda = \mathcal{V}_\lambda) \rightarrow 1, \text{ as } s_\lambda \rightarrow \infty.$$

where $\hat{\mathcal{V}}_\lambda$ represents the set of IVs selected by the screening procedure.

S.5.4 Proof of Theorem S1

For any $\mathcal{V} \subseteq \mathcal{S}_\lambda$, we denote a collection of sparse vectors

$$\mathcal{R}_V = \left\{ \mathbf{a} \in \mathbb{R}^{|\mathcal{S}_\lambda| \times 1} : a_j = 0, \text{ for } j \in \mathcal{V}, a_k \neq 0, \text{ for } k \in \mathcal{V}^c \right\}$$

and a function

$$\begin{aligned} h(\mathcal{V}, \theta) &= \min_{\mathbf{r} \in \mathcal{R}_V} \sum_{j \in \mathcal{S}_\lambda} \hat{l} \left(\theta, \mathbf{r}; \hat{\beta}_{Y_j}, \sigma_{Y_j}, \hat{\beta}_{X_{j,\text{RB}}}, \hat{\sigma}_{X_{j,\text{RB}}} \right) \\ &= \sum_{j \in \mathcal{V}} \frac{(\hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_{j,\text{RB}}})^2 - \theta^2 \cdot \hat{\sigma}_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}. \end{aligned}$$

Now we want to show $\mathbb{P}(\hat{\mathcal{V}}_\lambda \notin \mathcal{V}_{\text{valid}}) \rightarrow 0$ as $s_\lambda \rightarrow \infty$ by utilizing the following inequality:

$$\begin{aligned} \mathbb{P}(\hat{\mathcal{V}}_\lambda \notin \mathcal{V}_{\text{valid}}) &= \mathbb{P} \left(\min_{v \in \mathbb{N}_+, v \leq s_\lambda} \left[\min_{|\mathcal{V}|=v, \mathcal{V} \notin \mathcal{V}_{\text{valid}}} \min_{\theta \in \mathbb{R}} h(\mathcal{V}, \theta) - \kappa_n \cdot v \right] \leq \min_{\theta \in \mathbb{R}} h(\mathcal{V}_\lambda, \theta) - \kappa_n \cdot v_\lambda \right) \\ &\leq \bigcup_{\mathcal{V} \subseteq \mathcal{S}_\lambda, \mathcal{V} \notin \mathcal{V}_{\text{valid}}} \mathbb{P} \left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}, \theta) - \kappa_n \cdot |\mathcal{V}| \leq \min_{\theta \in \mathbb{R}} h(\mathcal{V}_\lambda, \theta) - \kappa_n \cdot v_\lambda \right) \\ &\leq \sum_{v=1}^{s_\lambda} \binom{s_\lambda}{v} \max_{|\mathcal{V}|=v, \mathcal{V} \notin \mathcal{V}_{\text{valid}}} \mathbb{P} \left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}, \theta) - \kappa_n \cdot v \leq h(\mathcal{V}_\lambda, \theta_0) - \kappa_n \cdot v_\lambda \right) \\ &\leq \sum_{v=1}^{s_\lambda} s_\lambda^v \max_{|\mathcal{V}|=v, \mathcal{V} \notin \mathcal{V}_{\text{valid}}} \mathbb{P} \left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}, \theta) - \kappa_n \cdot v \leq h(\mathcal{V}_\lambda, \theta_0) - \kappa_n \cdot v_\lambda \right) \\ &\leq \max_{\mathcal{V} \notin \mathcal{V}_{\text{valid}}} e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot \mathbb{P} \left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}, \theta) - \kappa_n \cdot |\mathcal{V}| \leq h(\mathcal{V}_\lambda, \theta_0) - \kappa_n \cdot v_\lambda \right) \\ &= e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot \mathbb{P} \left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_\lambda, \theta_0) - \kappa_n \cdot v_\lambda \right). \end{aligned} \tag{S1}$$

where $\mathcal{V}^* = \underset{\mathcal{V} \notin \mathcal{V}_{\text{valid}}}{\text{argmax}} e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot \mathbb{P} \left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}, \theta) - \kappa_n \cdot |\mathcal{V}| \leq h(\mathcal{V}_\lambda, \theta_0) - \kappa_n \cdot v_\lambda \right)$ and $v^* = |\mathcal{V}^*|$.

This is because the above inequality implies that as long as we show that

$$e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot \mathbb{P} \left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_\lambda, \theta_0) - \kappa_n \cdot v_\lambda \right) \rightarrow 0. \tag{S2}$$

then $\mathbb{P}(\hat{\mathcal{V}}_\lambda \notin \mathcal{V}_{\text{valid}}) \rightarrow 0$ holds. Here, we also note that the first equation in Eq (S1) follows from the definition of the optimization problem defined in Equation 2 in the manuscript, the second to the fifth inequalities in Eq (S1) hold following $\min_{\theta \in \mathbb{R}} h(\mathcal{V}_\lambda, \theta) \leq h(\mathcal{V}_\lambda, \theta_0)$, $\binom{s_\lambda}{v} \leq s_\lambda^v$ and some basic

calculations.

To prove Equation (2) goes to 0, we discuss $\mathcal{V}^* \notin \mathcal{V}_{\text{valid}}$ in three different cases.

- $|\mathcal{V}^*| = v^* \geq v_\lambda$ and $\mathcal{V}^* \neq \mathcal{V}_\lambda$,
- $c_1 \cdot v_\lambda \leq |\mathcal{V}^*| < v_\lambda$ but $\rho(\mathcal{V}^*) < 1$,
- $|\mathcal{V}^*| = v^* < c_1 \cdot v_\lambda$.

In each case, we show $e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot \mathbb{P} \left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_\lambda, \theta_0) - \kappa_n \cdot v_\lambda \right) \rightarrow 0$ and therefore (2) holds uniformly for all $\mathcal{V}^* \subseteq \mathcal{S}_\lambda$ and $\mathcal{V}^* \neq \mathcal{V}_\lambda$.

S.5.4.1 Case 1: $|\mathcal{V}^*| = v^* \geq v_\lambda$ and $\mathcal{V}^* \neq \mathcal{V}_\lambda$

To show the above results, we analyze the asymptotic properties of $h(\mathcal{V}_\lambda, \theta_0)$, $\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta)$ and κ_n .

We start with $h(\mathcal{V}_\lambda, \theta_0)$ and decompose it below following our notation defined in Section S.5.1.

$$\begin{aligned}
h(\mathcal{V}_\lambda, \theta_0) &= \sum_{j \in \mathcal{V}_\lambda} \frac{(\hat{\beta}_{Y_j} - \theta_0 \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} + \theta_0^2 \cdot \sum_{j \in \mathcal{V}_\lambda} \frac{(\hat{\beta}_{X_{j,\text{RB}}} - \beta_{X_j})^2 - \hat{\sigma}_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \\
&\quad - 2\theta_0 \cdot \sum_{j \in \mathcal{V}_\lambda} \frac{(\hat{\beta}_{Y_j} - \theta_0 \cdot \beta_{X_j})(\hat{\beta}_{X_{j,\text{RB}}} - \beta_{X_j})}{\sigma_{Y_j}^2} \\
&= \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j^2}{\sigma_{Y_j}^2} + \theta_0^2 \cdot \sum_{j \in \mathcal{V}_\lambda} \frac{u_j^2 - \hat{\sigma}_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} - 2\theta_0 \cdot \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j u_j}{\sigma_{Y_j}^2}. \tag{S3}
\end{aligned}$$

Next, we study the asymptotic property of $\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta)$. We denote $\hat{\theta}(\mathcal{V}^*) = \arg \min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta)$ and decompose $h(\mathcal{V}^*, \hat{\theta}(\mathcal{V}^*))$ in a similar way as $h(\mathcal{V}_\lambda, \theta_0)$, following our notation defined in Section

S.5.1.

$$\begin{aligned}
h(\mathcal{V}^*, \widehat{\theta}(\mathcal{V}^*)) &= \sum_{j \in \mathcal{V}^*} \frac{(\widehat{\beta}_{Y_j} - \widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} + \widehat{\theta}(\mathcal{V}^*)^2 \sum_{j \in \mathcal{V}^*} \frac{(\widehat{\beta}_{X_{j,\text{RB}}} - \beta_{X_j})^2 - \widehat{\sigma}_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \\
&\quad + 2\widehat{\theta}(\mathcal{V}^*) \sum_{j \in \mathcal{V}^*} \frac{(\widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot (\widehat{\beta}_{X_{j,\text{RB}}} - \beta_{X_j})}{\sigma_{Y_j}^2} \\
&\quad - 2\widehat{\theta}(\mathcal{V}^*) \sum_{j \in \mathcal{V}^*} \frac{(\widehat{\beta}_{Y_j} - \theta_0 \cdot \beta_{X_j} - r_j)(\widehat{\beta}_{X_{j,\text{RB}}} - \beta_{X_j})}{\sigma_{Y_j}^2} \\
&= \sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j + \nu_j - \widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} + \widehat{\theta}(\mathcal{V}^*)^2 \sum_{j \in \mathcal{V}^*} \frac{u_j^2 - \widehat{\sigma}_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \\
&\quad + 2\widehat{\theta}(\mathcal{V}^*) \sum_{j \in \mathcal{V}^*} \frac{(\widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot u_j}{\sigma_{Y_j}^2} - 2\widehat{\theta}(\mathcal{V}^*) \sum_{j \in \mathcal{V}^*} \frac{\nu_j u_j}{\sigma_{Y_j}^2} \\
&= \sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} + \sum_{j \in \mathcal{V}^*} \frac{\nu_j^2}{\sigma_{Y_j}^2} + 2 \sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j}) \cdot \nu_j}{\sigma_{Y_j}^2} \\
&\quad + \widehat{\theta}(\mathcal{V}^*)^2 \sum_{j \in \mathcal{V}^*} \frac{u_j^2 - \widehat{\sigma}_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} + 2\widehat{\theta}(\mathcal{V}^*) \sum_{j \in \mathcal{V}^*} \frac{(\widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot u_j}{\sigma_{Y_j}^2} - 2\widehat{\theta}(\mathcal{V}^*) \sum_{j \in \mathcal{V}^*} \frac{\nu_j u_j}{\sigma_{Y_j}^2}.
\end{aligned}$$

With these decomposition, the probability in Equation (2) can be rewritten as

$$\begin{aligned}
&\mathbb{P}\left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_\lambda, \theta_0) - \kappa_n \cdot v_\lambda\right) \\
&= \mathbb{P}\left(\sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} + \kappa_n \cdot (v_\lambda - v^*) \leq - \sum_{j \in \mathcal{V}^*} \frac{\nu_j^2}{\sigma_{Y_j}^2} - 2 \sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j}) \cdot \nu_j}{\sigma_{Y_j}^2}\right. \\
&\quad - \widehat{\theta}(\mathcal{V}^*)^2 \sum_{j \in \mathcal{V}^*} \frac{u_j^2 - \widehat{\sigma}_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} + \widehat{\theta}(\mathcal{V}^*)^2 \sum_{j \in \mathcal{V}^*} \frac{\widehat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} - 2\widehat{\theta}(\mathcal{V}^*) \sum_{j \in \mathcal{V}^*} \frac{(\widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot u_j}{\sigma_{Y_j}^2} \\
&\quad \left. + 2\widehat{\theta}(\mathcal{V}^*) \sum_{j \in \mathcal{V}^*} \frac{\nu_j u_j}{\sigma_{Y_j}^2} + \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j^2}{\sigma_{Y_j}^2} + \theta_0^2 \sum_{j \in \mathcal{V}_\lambda} \frac{u_j^2 - \widehat{\sigma}_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} - \theta_0^2 \sum_{j \in \mathcal{V}_\lambda} \frac{\widehat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} - 2\theta_0 \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j u_j}{\sigma_{Y_j}^2}\right).
\end{aligned}$$

When $|\mathcal{V}^*| = v^* \geq v_\lambda$ and $\mathcal{V}^* \neq \mathcal{V}_\lambda$, we know that there is at least one $r_j \in \mathcal{V}^*$ such that $r_j \neq 0$.

So $\rho(\mathcal{V}^*)$ is well-defined and by Condition 3 we have $\rho(\mathcal{V}^*) \leq c_0$. By some calculations, we can see that

$$\sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} \geq \min_{\theta \in \mathbb{R}} \sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \theta \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} = \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2} \cdot (1 - \rho(\mathcal{V}^*)).$$

with probability 1. Therefore,

$$\sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} \geq \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2} \cdot (1 - c_0).$$

with probability 1.

Then under Condition 4, there exists a $C_0 > 0$ such that

$$\begin{aligned} & \mathbb{P}\left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_\lambda, \theta_0) - \kappa_n \cdot v_\lambda\right) \\ & \leq \mathbb{P}\left(\sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} - \kappa_n \cdot (v^* - v_\lambda) \leq - \sum_{j \in \mathcal{V}^*} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} - \hat{\theta}(\mathcal{V}^*)^2 \sum_{j \in \mathcal{V}^*} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right. \\ & \quad \left. + \hat{\theta}(\mathcal{V}^*)^2 \sum_{j \in \mathcal{V}^*} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} - 2 \sum_{j \in \mathcal{V}^*} \frac{(\hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot \nu_j}{\sigma_{Y_j}^2} - 2\hat{\theta}(\mathcal{V}^*) \sum_{j \in \mathcal{V}^*} \frac{(\hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot u_j}{\sigma_{Y_j}^2}\right. \\ & \quad \left. + 2\hat{\theta}(\mathcal{V}^*) \sum_{j \in \mathcal{V}^*} \frac{\nu_j u_j}{\sigma_{Y_j}^2} + \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} + \theta_0^2 \sum_{j \in \mathcal{V}_\lambda} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} - \theta_0^2 \sum_{j \in \mathcal{V}_\lambda} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} - 2\theta_0 \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j u_j}{\sigma_{Y_j}^2}\right) \\ & \leq \mathbb{P}\left(\sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} - \kappa_n \cdot (v^* - v_\lambda) \leq \left| \sum_{j \in \mathcal{V}^*} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} \right| + \hat{\theta}(\mathcal{V}^*)^2 \left| \sum_{j \in \mathcal{V}^*} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right|\right. \\ & \quad \left. + 2\left| \hat{\theta}(\mathcal{V}^*) \sum_{j \in \mathcal{V}^*} \frac{(\hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot u_j}{\sigma_{Y_j}^2} \right| + 2 \left| \sum_{j \in \mathcal{V}^*} \frac{(\hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot \nu_j}{\sigma_{Y_j}^2} \right| + 2\left| \hat{\theta}(\mathcal{V}^*) \sum_{j \in \mathcal{V}^*} \frac{\nu_j u_j}{\sigma_{Y_j}^2} \right|\right. \\ & \quad \left. + \hat{\theta}(\mathcal{V}^*)^2 \left| \sum_{j \in \mathcal{V}^*} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} \right| + \theta_0^2 \left| \sum_{j \in \mathcal{V}_\lambda} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + \theta_0^2 \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + 2\theta_0 \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j u_j}{\sigma_{Y_j}^2} \right|\right) \\ & \leq \mathbb{P}\left(\sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} - \kappa_n \cdot (v^* - v_\lambda) \leq \left| \sum_{j \in \mathcal{V}^*} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} \right| + C_0^2 \left| \sum_{j \in \mathcal{V}^*} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right|\right. \\ & \quad \left. + C_0^2 \left| \sum_{j \in \mathcal{V}^*} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + 2C_0 \left| \sum_{j \in \mathcal{V}^*} \frac{(\hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot u_j}{\sigma_{Y_j}^2} \right| + 2 \left| \sum_{j \in \mathcal{V}^*} \frac{(\hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot \nu_j}{\sigma_{Y_j}^2} \right|\right. \\ & \quad \left. + 2C_0 \left| \sum_{j \in \mathcal{V}^*} \frac{\nu_j u_j}{\sigma_{Y_j}^2} \right| + \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} \right| + \theta_0^2 \left| \sum_{j \in \mathcal{V}_\lambda} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + \theta_0^2 \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + 2\theta_0 \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j u_j}{\sigma_{Y_j}^2} \right|\right). \end{aligned}$$

Denote the event

$$\begin{aligned} & \sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} - \kappa_n \cdot (v^* - v_\lambda) \leq \left| \sum_{j \in \mathcal{V}^*} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} \right| + C_0^2 \left| \sum_{j \in \mathcal{V}^*} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| \\ & \quad + C_0^2 \left| \sum_{j \in \mathcal{V}^*} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + 2C_0 \left| \sum_{j \in \mathcal{V}^*} \frac{(\hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot u_j}{\sigma_{Y_j}^2} \right| + 2 \left| \sum_{j \in \mathcal{V}^*} \frac{(\hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot \nu_j}{\sigma_{Y_j}^2} \right| \\ & \quad + 2C_0 \left| \sum_{j \in \mathcal{V}^*} \frac{\nu_j u_j}{\sigma_{Y_j}^2} \right| + \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} \right| + \theta_0^2 \left| \sum_{j \in \mathcal{V}_\lambda} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + \theta_0^2 \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + 2\theta_0 \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j u_j}{\sigma_{Y_j}^2} \right| \end{aligned}$$

as $\mathcal{C}(\mathcal{V}^*)$ and

$$\delta(\mathcal{V}^*) = \sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} - \kappa_n \cdot (v^* - v_\lambda),$$

we have

$$\begin{aligned} \mathcal{C}(\mathcal{V}^*) \subseteq & \left\{ C_0^2 \cdot \left| \sum_{j \in \mathcal{V}^*} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}^2}}{\sigma_{Y_j}^2} \right| \geq \frac{\delta(\mathcal{V}^*)}{10} \right\} \cup \left\{ \theta_0^2 \cdot \left| \sum_{j \in \mathcal{V}_\lambda} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}^2}}{\sigma_{Y_j}^2} \right| \geq \frac{\delta(\mathcal{V}^*)}{10} \right\} \\ & \cup \left\{ \left| \sum_{j \in \mathcal{V}^*} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} \right| \geq \frac{\delta(\mathcal{V}^*)}{10} \right\} \cup \left\{ \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} \right| \geq \frac{\delta(\mathcal{V}^*)}{10} \right\} \cup \left\{ \theta_0^2 \cdot \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}^2}}{\sigma_{Y_j}^2} \right| \geq \frac{\delta(\mathcal{V}^*)}{10} \right\} \\ & \cup \left\{ C_0^2 \cdot \left| \sum_{j \in \mathcal{V}^*} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}^2}}{\sigma_{Y_j}^2} \right| \geq \frac{\delta(\mathcal{V}^*)}{10} \right\} \cup \left\{ 2C_0 \cdot \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j u_j}{\sigma_{Y_j}^2} \right| \geq \frac{\delta(\mathcal{V}^*)}{10} \right\} \cup \left\{ 2\theta_0 \cdot \left| \sum_{j \in \mathcal{V}^*} \frac{\nu_j u_j}{\sigma_{Y_j}^2} \right| \geq \frac{\delta(\mathcal{V}^*)}{10} \right\} \\ & \cup \left\{ 2 \left| \sum_{j \in \mathcal{V}^*} \frac{(\hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot \nu_j}{\sigma_{Y_j}^2} \right| \geq \frac{\delta(\mathcal{V}^*)}{10} \right\} \\ & \cup \left\{ 2C_0 \cdot \left| \sum_{j \in \mathcal{V}^*} \frac{(\hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot u_j}{\sigma_{Y_j}^2} \right| \geq \frac{\delta(\mathcal{V}^*)}{10} \right\}. \end{aligned}$$

When $|\mathcal{V}^*| = v^* \geq v_\lambda$, we know that the number of r_j that is not equal to zero for $j \in \mathcal{V}^*$ is at least $v^* - v_\lambda$. Then if $\min_{j \in \mathcal{S}_\lambda, r_j \neq 0} \frac{\kappa_n}{\sigma_{Y_j}^2} \rightarrow 0$ (Condition 7), we have $\frac{\kappa_n \cdot (v^* - v_\lambda)}{\sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2}} \rightarrow 0$. So there exists a $c > 0$ such that

$$\delta(\mathcal{V}^*) = \sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} - \kappa_n \cdot (v^* - v_\lambda) \geq \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2} \cdot c.$$

uniformly holds for all $|\mathcal{V}^*| = v^* \geq v_\lambda$ and $\mathcal{V}^* \neq \mathcal{V}_\lambda$. Then $\mathbb{P}(\mathcal{C}(\mathcal{V}^*))$ is bounded by

$$\begin{aligned}
& \mathbb{P}\left(C_0^2 \cdot \left|\sum_{j \in \mathcal{V}^*} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}^2}}{\sigma_{Y_j}^2}\right| \geq \frac{c}{10} \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2}\right) + \mathbb{P}\left(\theta_0^2 \cdot \left|\sum_{j \in \mathcal{V}_\lambda} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}^2}}{\sigma_{Y_j}^2}\right| \geq \frac{c}{10} \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2}\right) \\
& + \mathbb{P}\left(C_0^2 \cdot \left|\sum_{j \in \mathcal{V}^*} \frac{\widehat{\sigma}_{X_{j,\text{RB}}^2} - \sigma_{X_{j,\text{RB}}^2}}{\sigma_{Y_j}^2}\right| \geq \frac{c}{10} \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2}\right) + \mathbb{P}\left(\theta_0^2 \cdot \left|\sum_{j \in \mathcal{V}_\lambda} \frac{\widehat{\sigma}_{X_{j,\text{RB}}^2} - \sigma_{X_{j,\text{RB}}^2}}{\sigma_{Y_j}^2}\right| \geq \frac{c}{10} \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2}\right) \\
& + \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}^*} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2}\right| \geq \frac{c}{10} \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2}\right) + \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2}\right| \geq \frac{c}{10} \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2}\right) \\
& + \mathbb{P}\left(2\theta_0 \cdot \left|\sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j u_j}{\sigma_{Y_j}^2}\right| \geq \frac{c}{10} \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2}\right) + \mathbb{P}\left(2C_0 \cdot \left|\sum_{j \in \mathcal{V}^*} \frac{\nu_j u_j}{\sigma_{Y_j}^2}\right| \geq \frac{c}{10} \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2}\right) \\
& + \mathbb{P}\left(2\left|\sum_{j \in \mathcal{V}^*} \frac{(\widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot \nu_j}{\sigma_{Y_j}^2}\right| \geq \frac{1}{10} \left(\sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} - \kappa_n \cdot (v^* - v_\lambda)\right)\right) \\
& + \mathbb{P}\left(2C_0 \cdot \left|\sum_{j \in \mathcal{V}^*} \frac{(\widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot u_j}{\sigma_{Y_j}^2}\right| \geq \frac{1}{10} \left(\sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} - \kappa_n \cdot (v^* - v_\lambda)\right)\right).
\end{aligned}$$

Using Lemma 1, we know that there exists a $c' > 0$ such that the first eight terms are bounded by $2 \cdot e^{-c' \cdot \min\left\{\frac{v^{*2} \cdot r_\lambda^2(\mathcal{V}^*)}{v_\lambda}, v^* \cdot r_\lambda^2(\mathcal{V}^*), v^* \cdot r_\lambda(\mathcal{V}^*)\right\}}$. We also have

$$\begin{aligned}
\left|\sum_{j \in \mathcal{V}^*} \frac{(\widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot u_j}{\sigma_{Y_j}^2}\right| & \leq \sqrt{\sum_{j \in \mathcal{V}^*} \frac{(\widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j)^2}{\sigma_{Y_j}^2}} \sqrt{\sum_{j \in \mathcal{V}^*} \frac{u_j^2}{\sigma_{Y_j}^2}} \text{ and} \\
\left|\sum_{j \in \mathcal{V}^*} \frac{(\widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot \nu_j}{\sigma_{Y_j}^2}\right| & \leq \sqrt{\sum_{j \in \mathcal{V}^*} \frac{(\widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j)^2}{\sigma_{Y_j}^2}} \sqrt{\sum_{j \in \mathcal{V}^*} \frac{\nu_j^2}{\sigma_{Y_j}^2}}.
\end{aligned}$$

That means there exists a $c'' > 0$ such that $\mathbb{P}\left(2\left|\sum_{j \in \mathcal{V}^*} \frac{(\widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot \nu_j}{\sigma_{Y_j}^2}\right| \geq \frac{\delta(\mathcal{V}^*)}{10}\right)$ is further bounded by $\mathbb{P}\left(\sum_{j \in \mathcal{V}^*} \frac{\nu_j^2}{\sigma_{Y_j}^2} \geq c'' \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2}\right)$ through

$$\begin{aligned}
& \mathbb{P}\left(2\left|\sum_{j \in \mathcal{V}^*} \frac{(\widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot \nu_j}{\sigma_{Y_j}^2}\right| \geq \frac{1}{10} \left(\sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} - \kappa_n \cdot (v^* - v_\lambda)\right)\right) \\
& \leq \mathbb{P}\left(2\sqrt{\sum_{j \in \mathcal{V}^*} \frac{\nu_j^2}{\sigma_{Y_j}^2}} \geq \frac{\frac{1}{10} \left(\sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} - \kappa_n \cdot (v^* - v_\lambda)\right)}{\sqrt{\sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2}}}\right) \\
& \leq \mathbb{P}\left(\sum_{j \in \mathcal{V}^*} \frac{\nu_j^2}{\sigma_{Y_j}^2} \geq c'' \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2}\right).
\end{aligned}$$

Similarly,

$$\begin{aligned} & \mathbb{P}\left(2C_0 \cdot \left| \sum_{j \in \mathcal{V}^*} \frac{(\hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot u_j}{\sigma_{Y_j}^2} \right| \geq \frac{1}{10} \left(\sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} - \kappa_n \cdot (v^* - v_\lambda) \right) \right) \\ & \leq \mathbb{P}\left(\sum_{j \in \mathcal{V}^*} \frac{u_j^2}{\sigma_{Y_j}^2} \geq c'' \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2} \right). \end{aligned}$$

Using Lemma 1, we can also show that these two terms are bounded by $2 \cdot e^{-c' \cdot \min\left\{\frac{v^{*2} \cdot r_\lambda^2(\mathcal{V}^*)}{v_\lambda}, v^* \cdot r_\lambda^2(\mathcal{V}^*), v^* \cdot r_\lambda(\mathcal{V}^*)\right\}}$.

To prove $e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot \mathbb{P}\left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_\lambda, \theta_0) - \kappa_n \cdot v_\lambda\right) \rightarrow 0$ for any $\mathcal{V}^* \subseteq \mathcal{S}_\lambda$ such that $|\mathcal{V}^*| = v^* \geq v_\lambda$, we only need to show

$$2e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot e^{-c' \cdot \min\left\{v^* \cdot r_\lambda^2(\mathcal{V}^*), v^* \cdot r_\lambda(\mathcal{V}^*), \frac{v^{*2} \cdot r_\lambda^2(\mathcal{V}^*)}{v_\lambda}\right\}} \rightarrow 0.$$

Using $v^* \geq v_\lambda$, it suffices to show

$$2e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot e^{-c' \cdot \min\left\{v^* \cdot r_\lambda^2(\mathcal{V}^*), v^* \cdot r_\lambda(\mathcal{V}^*)\right\}} \rightarrow 0.$$

We prove this formula in Lemma 2 and conclude that

$$e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot \mathbb{P}\left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_\lambda, \theta_0) - \kappa_n \cdot v_\lambda\right) \rightarrow 0.$$

uniformly for all $\mathcal{V}^* \subseteq \mathcal{S}_\lambda$ such that $|\mathcal{V}^*| = v^* \geq v_\lambda$ and $\mathcal{V}^* \neq \mathcal{V}_\lambda$.

S.5.4.2 Case 2: When $c_1 \cdot v_\lambda \leq |\mathcal{V}^*| < v_\lambda$ but $\rho(\mathcal{V}^*) < 1$

When $c_1 \cdot v_\lambda \leq |\mathcal{V}^*| < v_\lambda$ but $\rho(\mathcal{V}^*) < 1$, we can know from Condition 3 that $\rho(\mathcal{V}^*) \leq c_0$. Therefore,

$$\sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} \geq \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2} \cdot (1 - c_0).$$

with probability 1. Similarly, under Condition 4, as $n \rightarrow \infty$, there exists a $C_0 > 0$ such that

$$\begin{aligned}
& \mathbb{P}\left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_\lambda, \theta_0) - \kappa_n \cdot v_\lambda\right) \\
& \leq \mathbb{P}\left(\sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} - \kappa_n \cdot (v^* - v_\lambda) \leq \left|\sum_{j \in \mathcal{V}^*} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2}\right| + C_0^2 \left|\sum_{j \in \mathcal{V}^*} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right|\right. \\
& \quad + C_0^2 \left|\sum_{j \in \mathcal{V}^*} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right| + 2C_0 \left|\sum_{j \in \mathcal{V}^*} \frac{(\hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot u_j}{\sigma_{Y_j}^2}\right| + 2 \left|\sum_{j \in \mathcal{V}^*} \frac{(\hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot \nu_j}{\sigma_{Y_j}^2}\right| \\
& \quad \left. + 2C_0 \left|\sum_{j \in \mathcal{V}^*} \frac{\nu_j u_j}{\sigma_{Y_j}^2}\right| + \left|\sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2}\right| + \theta_0^2 \left|\sum_{j \in \mathcal{V}_\lambda} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right| + \theta_0^2 \left|\sum_{j \in \mathcal{V}_\lambda} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right| + 2\theta_0 \left|\sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j u_j}{\sigma_{Y_j}^2}\right|\right) \\
& = \mathbb{P}(\mathcal{C}(\mathcal{V}^*)).
\end{aligned}$$

We also have

$$\delta(\mathcal{V}^*) = \sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} - \kappa_n \cdot (v^* - v_\lambda) \geq \sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} \geq \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2} \cdot (1 - c_0).$$

This is because $v^* < v_\lambda$ and $\kappa_n \cdot (v^* - v_\lambda) < 0$.

Then the probability $\mathbb{P}(\mathcal{C}(\mathcal{V}^*))$ is bounded by

$$\begin{aligned}
& \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}^*} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq \frac{1 - c_0}{10} \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2}\right) + \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}_\lambda} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq \frac{1 - c_0}{10} \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2}\right) \\
& + \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}^*} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq \frac{1 - c_0}{10} \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2}\right) + \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}_\lambda} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq \frac{1 - c_0}{10} \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2}\right) \\
& + \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}^*} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2}\right| \geq \frac{1 - c_0}{10} \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2}\right) + \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2}\right| \geq \frac{1 - c_0}{10} \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2}\right) \\
& + \mathbb{P}\left(2\left|\sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j u_j}{\sigma_{Y_j}^2}\right| \geq \frac{1 - c_0}{10} \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2}\right) + \mathbb{P}\left(2\left|\sum_{j \in \mathcal{V}^*} \frac{\nu_j u_j}{\sigma_{Y_j}^2}\right| \geq \frac{1 - c_0}{10} \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2}\right) \\
& + \mathbb{P}\left(2\left|\sum_{j \in \mathcal{V}^*} \frac{(\hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot \nu_j}{\sigma_{Y_j}^2}\right| \geq \frac{1}{10} \sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2}\right) \\
& + \mathbb{P}\left(2\left|\sum_{j \in \mathcal{V}^*} \frac{(\hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot u_j}{\sigma_{Y_j}^2}\right| \geq \frac{1}{10} \sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2}\right).
\end{aligned}$$

Knowing that

$$\left| \sum_{j \in \mathcal{V}^*} \frac{(\widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot u_j}{\sigma_{Y_j}^2} \right| \leq \sqrt{\sum_{j \in \mathcal{V}^*} \frac{(\widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j)^2}{\sigma_{Y_j}^2}} \sqrt{\sum_{j \in \mathcal{V}^*} \frac{u_j^2}{\sigma_{Y_j}^2}}$$

$$\left| \sum_{j \in \mathcal{V}^*} \frac{(\widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot \nu_j}{\sigma_{Y_j}^2} \right| \leq \sqrt{\sum_{j \in \mathcal{V}^*} \frac{(\widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j)^2}{\sigma_{Y_j}^2}} \sqrt{\sum_{j \in \mathcal{V}^*} \frac{\nu_j^2}{\sigma_{Y_j}^2}}$$

We can similarly show that there exists a $c'' > 0$ such that the last two terms are further bounded by

$$\mathbb{P}\left(\sum_{j \in \mathcal{V}^*} \frac{\nu_j^2}{\sigma_{Y_j}^2} \geq c'' \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2}\right) + \mathbb{P}\left(\sum_{j \in \mathcal{V}^*} \frac{u_j^2}{\sigma_{Y_j}^2} \geq c'' \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2}\right).$$

Using Lemma 1, we know that there exists a $c' > 0$ such that all these terms are bounded by $2 \cdot e^{-c' \cdot \min\left\{\frac{v^{*2} \cdot r_\lambda^2(\mathcal{V}^*)}{v_\lambda}, v^* \cdot r_\lambda^2(\mathcal{V}^*), v^* \cdot r_\lambda(\mathcal{V}^*)\right\}}$.

To prove $e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot \mathbb{P}\left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_\lambda, \theta_0) - \kappa_n \cdot v_\lambda\right) \rightarrow 0$ for any \mathcal{V}^* such that $c_1 \cdot v_\lambda \leq |\mathcal{V}^*| < v_\lambda$ but $\rho(\mathcal{V}^*) < 1$, we only need to show

$$2e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot e^{-c' \cdot \min\left\{v^* \cdot r_\lambda^2(\mathcal{V}^*), v^* \cdot r_\lambda(\mathcal{V}^*), \frac{v^{*2} \cdot r_\lambda^2(\mathcal{V}^*)}{v_\lambda}\right\}} \rightarrow 0.$$

Knowing that $|\mathcal{V}^*| < v_\lambda$, it suffices to show that

$$2e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot e^{-c' \cdot \min\left\{v^* \cdot r_\lambda(\mathcal{V}^*), \frac{v^{*2} \cdot r_\lambda^2(\mathcal{V}^*)}{v_\lambda}\right\}} \rightarrow 0.$$

We prove this formula in Lemma 2 and conclude that

$$e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot \mathbb{P}\left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_\lambda, \theta_0) - \kappa_n \cdot v_\lambda\right) \rightarrow 0.$$

uniformly for all $\mathcal{V}^* \subseteq \mathcal{S}_\lambda$ such that $c_1 \cdot v_\lambda \leq |\mathcal{V}^*| < v_\lambda$ but $\rho(\mathcal{V}^*) < 1$.

S.5.4.3 Case 3: When $|\mathcal{V}^*| = v^* < c_1 \cdot v_\lambda$

When $v^* < c_1 \cdot v_\lambda$, we decompose $h(\mathcal{V}^*, \widehat{\theta}(\mathcal{V}^*))$ in a different way,

$$\begin{aligned} h(\mathcal{V}^*, \widehat{\theta}(\mathcal{V}^*)) &= \sum_{j \in \mathcal{V}^*} \frac{(\widehat{\beta}_{Y_j} - \widehat{\theta}(\mathcal{V}^*) \cdot \widehat{\beta}_{X_{j,\text{RB}}})^2}{\sigma_{Y_j}^2} - \widehat{\theta}(\mathcal{V}^*)^2 \sum_{j \in \mathcal{V}^*} \frac{\widehat{\sigma}_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \\ &= \sum_{j \in \mathcal{V}^*} \frac{(\widehat{\beta}_{Y_j} - \widehat{\theta}(\mathcal{V}^*) \cdot \widehat{\beta}_{X_{j,\text{RB}}})^2}{\sigma_{Y_j}^2} - \widehat{\theta}(\mathcal{V}^*)^2 \sum_{j \in \mathcal{V}^*} \frac{\widehat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} - \widehat{\theta}(\mathcal{V}^*)^2 \sum_{j \in \mathcal{V}^*} \frac{\sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}. \end{aligned}$$

Under Condition 4, there exists a $C_0 > 0$ such that

$$\begin{aligned} &\mathbb{P}\left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_\lambda, \theta_0) - \kappa_n \cdot v_\lambda\right) \\ &\leq \mathbb{P}\left(\sum_{j \in \mathcal{V}^*} \frac{(\widehat{\beta}_{Y_j} - \widehat{\theta}(\mathcal{V}^*) \cdot \widehat{\beta}_{X_{j,\text{RB}}})^2}{\sigma_{Y_j}^2} + \kappa_n \cdot (v_\lambda - v^*) \leq \widehat{\theta}(\mathcal{V}^*)^2 \sum_{j \in \mathcal{V}^*} \frac{\widehat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} + \widehat{\theta}(\mathcal{V}^*)^2 \sum_{j \in \mathcal{V}^*} \frac{\sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right. \\ &\quad \left. + \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} + \theta_0^2 \sum_{j \in \mathcal{V}_\lambda} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} - \theta_0^2 \sum_{j \in \mathcal{V}_\lambda} \frac{\widehat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} - 2\theta_0 \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j u_j}{\sigma_{Y_j}^2}\right) \\ &\leq \mathbb{P}\left(\sum_{j \in \mathcal{V}^*} \frac{(\widehat{\beta}_{Y_j} - \widehat{\theta}(\mathcal{V}^*) \cdot \widehat{\beta}_{X_{j,\text{RB}}})^2}{\sigma_{Y_j}^2} + \kappa_n \cdot (v_\lambda - v^*) \leq C_0^2 \left| \sum_{j \in \mathcal{V}^*} \frac{\widehat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + C_0^2 \cdot v^*\right. \\ &\quad \left. + \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} \right| + \theta_0^2 \left| \sum_{j \in \mathcal{V}_\lambda} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + \theta_0^2 \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\widehat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + 2\theta_0 \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j u_j}{\sigma_{Y_j}^2} \right|\right) \\ &\leq \mathbb{P}\left(\kappa_n \cdot (v_\lambda - v^*) - C_0^2 \cdot v^* \leq C_0^2 \left| \sum_{j \in \mathcal{V}^*} \frac{\widehat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} \right|\right. \\ &\quad \left. + \theta_0^2 \left| \sum_{j \in \mathcal{V}_\lambda} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + \theta_0^2 \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\widehat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + 2\theta_0 \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j u_j}{\sigma_{Y_j}^2} \right|\right). \end{aligned}$$

Denote the event

$$\begin{aligned} \kappa_n \cdot (v_\lambda - v^*) - C_0^2 \cdot v^* &\leq C_0^2 \cdot \left| \sum_{j \in \mathcal{V}^*} \frac{\widehat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} \right| \\ &\quad + \theta_0^2 \cdot \left| \sum_{j \in \mathcal{V}_\lambda} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + \theta_0^2 \cdot \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\widehat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + 2\theta_0 \cdot \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j u_j}{\sigma_{Y_j}^2} \right| \end{aligned}$$

as $\mathcal{D}(\mathcal{V}^*)$, we have

$$\begin{aligned}
\mathcal{D}(\mathcal{V}^*) \subseteq & \left\{ \theta_0^2 \cdot \left| \sum_{j \in \mathcal{V}_\lambda} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| \geq \frac{1}{5} \cdot (\kappa_n \cdot (v_\lambda - v^*) - C_0^2 \cdot v^*) \right\} \\
& \cup \left\{ 2\theta_0 \cdot \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j u_j}{\sigma_{Y_j}^2} \right| \geq \frac{1}{5} \cdot (\kappa_n \cdot (v_\lambda - v^*) - C_0^2 \cdot v^*) \right\} \\
& \cup \left\{ \theta_0^2 \cdot \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| \geq \frac{1}{5} \cdot (\kappa_n \cdot (v_\lambda - v^*) - C_0^2 \cdot v^*) \right\} \\
& \cup \left\{ \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} \right| \geq \frac{1}{5} \cdot (\kappa_n \cdot (v_\lambda - v^*) - C_0^2 \cdot v^*) \right\} \\
& \cup \left\{ C_0^2 \cdot \left| \sum_{j \in \mathcal{V}^*} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| \geq \frac{1}{5} \cdot (\kappa_n \cdot (v_\lambda - v^*) - C_0^2 \cdot v^*) \right\}.
\end{aligned}$$

Using $v^* < c_1 \cdot v_\lambda$ and Condition 7, we know that there must be a $c''' > 0$ such that

$$\kappa_n \cdot (v_\lambda - v^*) - C_0^2 \cdot v^* \geq \kappa_n \cdot (1 - c_1) \cdot v_\lambda - C_0^2 \cdot v^* \geq c''' \cdot \kappa_n \cdot v_\lambda.$$

then the probability $\mathbb{P}(\mathcal{D}(\mathcal{V}^*))$ is bounded by

$$\begin{aligned}
& \mathbb{P}\left(\theta_0^2 \cdot \left| \sum_{j \in \mathcal{V}_\lambda} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| \geq \frac{c'''}{5} \cdot \kappa_n \cdot v_\lambda\right) + \mathbb{P}\left(C_0^2 \cdot \left| \sum_{j \in \mathcal{V}^*} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| \geq \frac{c'''}{5} \cdot \kappa_n \cdot v_\lambda\right) \\
& + \mathbb{P}\left(\theta_0^2 \cdot \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| \geq \frac{c'''}{5} \cdot \kappa_n \cdot v_\lambda\right) + \mathbb{P}\left(2\theta_0 \cdot \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j u_j}{\sigma_{Y_j}^2} \right| \geq \frac{c'''}{5} \cdot \kappa_n \cdot v_\lambda\right) \\
& + \mathbb{P}\left(\left| \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} \right| \geq \frac{c'''}{5} \cdot \kappa_n \cdot v_\lambda\right).
\end{aligned}$$

Using Lemma 1, we know that there exists a $c' > 0$ such that these five terms are bounded by $2 \cdot e^{-c' \cdot \min\{\kappa_n^2 v_\lambda, \frac{\kappa_n^2 v_\lambda^2}{v^*}, \kappa_n \cdot v_\lambda\}}$. To prove $e^{(s_\lambda + 1) \cdot \ln(s_\lambda)} \cdot \mathbb{P}\left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_\lambda, \theta_0) - \kappa_n \cdot v_\lambda\right) \rightarrow 0$, we only need to show

$$2e^{(s_\lambda + 1) \cdot \ln(s_\lambda)} \cdot e^{-c' \cdot \kappa_n \cdot v_\lambda} \rightarrow 0.$$

We can prove this by Lemma 3 and conclude that $e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot \mathbb{P} \left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_\lambda, \theta_0) - \kappa_n \cdot v_\lambda \right) \rightarrow 0$ uniformly for all $\mathcal{V}^* \subseteq \mathcal{S}_\lambda$ such that $v^* < c_1 \cdot v_\lambda$.

Therefore, we conclude that $e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot \mathbb{P} \left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_\lambda, \theta_0) - \kappa_n \cdot v_\lambda \right) \rightarrow 0$ uniformly for all $\mathcal{V}^* \subseteq \mathcal{S}_\lambda$ such that $\mathcal{V}^* \neq \mathcal{V}_\lambda$.

S.5.5 Proof of Perfect Screening Property

Following the similar procedure in Section 3.4, for any $\mathcal{V} \subset \mathcal{S}_\lambda$, we denote a function

$$h(\mathcal{V}, \theta) = \sum_{j \in \mathcal{V}} \frac{(\hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_{j,\text{RB}}})^2 - \theta^2 \cdot \hat{\sigma}_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}.$$

and show

$$\mathbb{P}(\hat{\mathcal{V}}_\lambda \neq \mathcal{V}_\lambda) \leq e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot \mathbb{P} \left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_\lambda, \theta_0) - \kappa_n \cdot v_\lambda \right).$$

where $\mathcal{V}^* = \underset{\mathcal{V} \neq \mathcal{V}_\lambda}{\text{argmax}} e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot \mathbb{P} \left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}, \theta) - \kappa_n \cdot |\mathcal{V}| \leq h(\mathcal{V}_\lambda, \theta_0) - \kappa_n \cdot v_\lambda \right)$ and $v^* = |\mathcal{V}^*|$.

As long as we show that

$$e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot \mathbb{P} \left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_\lambda, \theta_0) - \kappa_n \cdot v_\lambda \right) \rightarrow 0.$$

then $\mathbb{P}(\hat{\mathcal{V}}_\lambda \neq \mathcal{V}_\lambda) \rightarrow 0$ holds.

To prove it, we discuss \mathcal{V}^* in two different cases.

- $|\mathcal{V}^*| = v^* \geq v_\lambda$ and $\mathcal{V}^* \neq \mathcal{V}_\lambda$,
- $|\mathcal{V}^*| = v^* < v_\lambda$.

In each case, we show $e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot \mathbb{P} \left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_\lambda, \theta_0) - \kappa_n \cdot v_\lambda \right) \rightarrow 0$ and therefore $\mathbb{P}(\hat{\mathcal{V}}_\lambda \neq \mathcal{V}_\lambda) \rightarrow 0$ holds.

S.5.5.1 Case 1: $|\mathcal{V}^*| = v^* \geq v_\lambda$ and $\mathcal{V}^* \neq \mathcal{V}_\lambda$

The proof of this section is the same as the one in Section 3.4.1.

S.5.5.2 Case 2: When $|\mathcal{V}^*| = v^* < v_\lambda$

When $v^* < v_\lambda$, following the proof in Section 3.4.3, we show

$$\begin{aligned}
& \mathbb{P}\left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_\lambda, \theta_0) - \kappa_n \cdot v_\lambda\right) \\
& \leq \mathbb{P}\left(\kappa_n \cdot (v_\lambda - v^*) - C_0^2 \cdot v^* \leq C_0^2 \left| \sum_{j \in \mathcal{V}^*} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} \right| \right. \\
& \quad \left. + \theta_0^2 \left| \sum_{j \in \mathcal{V}_\lambda} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + \theta_0^2 \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + 2\theta_0 \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j u_j}{\sigma_{Y_j}^2} \right| \right) \\
& = \mathbb{P}(\mathcal{D}(\mathcal{V}^*)).
\end{aligned}$$

Where $\mathcal{D}(\mathcal{V}^*)$ is the event

$$\begin{aligned}
& \kappa_n \cdot (v_\lambda - v^*) - C_0^2 \cdot v^* \leq C_0^2 \cdot \left| \sum_{j \in \mathcal{V}^*} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} \right| \\
& \quad + \theta_0^2 \cdot \left| \sum_{j \in \mathcal{V}_\lambda} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + \theta_0^2 \cdot \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + 2\theta_0 \cdot \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j u_j}{\sigma_{Y_j}^2} \right|.
\end{aligned}$$

We have

$$\begin{aligned}
\mathcal{D}(\mathcal{V}^*) & \subseteq \left\{ \theta_0^2 \cdot \left| \sum_{j \in \mathcal{V}_\lambda} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| \geq \frac{1}{5} \cdot (\kappa_n \cdot (v_\lambda - v^*) - C_0^2 \cdot v^*) \right\} \\
& \cup \left\{ 2\theta_0 \cdot \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j u_j}{\sigma_{Y_j}^2} \right| \geq \frac{1}{5} \cdot (\kappa_n \cdot (v_\lambda - v^*) - C_0^2 \cdot v^*) \right\} \\
& \cup \left\{ \theta_0^2 \cdot \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| \geq \frac{1}{5} \cdot (\kappa_n \cdot (v_\lambda - v^*) - C_0^2 \cdot v^*) \right\} \\
& \cup \left\{ \left| \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} \right| \geq \frac{1}{5} \cdot (\kappa_n \cdot (v_\lambda - v^*) - C_0^2 \cdot v^*) \right\} \\
& \cup \left\{ C_0^2 \cdot \left| \sum_{j \in \mathcal{V}^*} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| \geq \frac{1}{5} \cdot (\kappa_n \cdot (v_\lambda - v^*) - C_0^2 \cdot v^*) \right\}.
\end{aligned}$$

Using $v^* < v_\lambda$ and Condition 7*, we know that there must be a $c'' > 0$ such that

$$\kappa_n \cdot (v_\lambda - v^*) - C_0^2 \cdot v^* \geq \kappa_n - C_0^2 \cdot v^* \geq c''' \cdot \kappa_n.$$

then the probability $\mathbb{P}(\mathcal{D}(\mathcal{V}^*))$ is bounded by

$$\begin{aligned} & \mathbb{P}\left(\theta_0^2 \cdot \left|\sum_{j \in \mathcal{V}_\lambda} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq \frac{c'''}{5} \cdot \kappa_n\right) + \mathbb{P}\left(C_0^2 \cdot \left|\sum_{j \in \mathcal{V}^*} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq \frac{c'''}{5} \cdot \kappa_n\right) \\ & + \mathbb{P}\left(\theta_0^2 \cdot \left|\sum_{j \in \mathcal{V}_\lambda} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq \frac{c'''}{5} \cdot \kappa_n\right) + \mathbb{P}\left(2\theta_0 \cdot \left|\sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j u_j}{\sigma_{Y_j}^2}\right| \geq \frac{c'''}{5} \cdot \kappa_n\right) \\ & + \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2}\right| \geq \frac{c'''}{5} \cdot \kappa_n\right). \end{aligned}$$

Using Lemma 1, we know that there exists a $c' > 0$ such that these five terms are bounded by $2 \cdot e^{-c' \cdot \min\left\{\frac{\kappa_n^2}{v_\lambda}, \frac{\kappa_n^2}{v^*}, \kappa_n\right\}}$. To prove $e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot \mathbb{P}\left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_\lambda, \theta_0) - \kappa_n \cdot v_\lambda\right) \rightarrow 0$,

we only need to show

$$2e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot e^{-c' \cdot \min\left\{\frac{\kappa_n^2}{v_\lambda}, \frac{\kappa_n^2}{v^*}, \kappa_n\right\}} = 2e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot e^{-c' \cdot \min\left\{\frac{\kappa_n^2}{v_\lambda}, \kappa_n\right\}} \rightarrow 0.$$

We can prove this by Lemma 3 and conclude that $e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot \mathbb{P}\left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_\lambda, \theta_0) - \kappa_n \cdot v_\lambda\right) \rightarrow 0$ uniformly for all $\mathcal{V}^* \subseteq \mathcal{S}_\lambda$ such that $v^* < v_\lambda$.

Therefore, we conclude that $e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot \mathbb{P}\left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_\lambda, \theta_0) - \kappa_n \cdot v_\lambda\right) \rightarrow 0$ uniformly for all $\mathcal{V}^* \subseteq \mathcal{S}_\lambda$ such that $\mathcal{V}^* \neq \mathcal{V}_\lambda$.

S.5.6 Lemmas

Lemma 1 *Under Condition 1,*

$$\mathbb{P}\left(\left|\sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j u_j}{\sigma_{Y_j}^2}\right| \geq t\right) \leq 2 \cdot e^{-c \cdot \min\left\{\frac{t^2}{v_\lambda}, t\right\}}, \quad \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}} \frac{\nu_j u_j}{\sigma_{Y_j}^2}\right| \geq t\right) \leq 2 \cdot e^{-c \cdot \min\left\{\frac{t^2}{v^*}, t\right\}}$$

$$\mathbb{P}\left(\left|\sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2}\right| \geq t\right) \leq 2 \cdot e^{-c \cdot \min\left\{\frac{t^2}{v_\lambda}, t\right\}}, \quad \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2}\right| \geq t\right) \leq 2 \cdot e^{-c \cdot \min\left\{\frac{t^2}{v^*}, t\right\}}$$

$$\mathbb{P}\left(\left|\sum_{j \in \mathcal{V}_\lambda} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq t\right) \leq 2 \cdot e^{-c \cdot \min\left\{\frac{t^2}{v_\lambda}, t\right\}}, \quad \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq t\right) \leq 2 \cdot e^{-c \cdot \min\left\{\frac{t^2}{v^*}, t\right\}}$$

$$\mathbb{P}\left(\left|\sum_{j \in \mathcal{V}_\lambda} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq t\right) \leq 2 \cdot e^{-c \cdot \min\left\{\frac{t^2}{v_\lambda}, t\right\}}, \quad \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq t\right) \leq 2 \cdot e^{-c \cdot \min\left\{\frac{t^2}{v^*}, t\right\}}$$

Lemma 2 Under Condition 2, 5 and 6, we have

$$\frac{\ln(s_\lambda)}{r_\lambda(\mathcal{V})} \rightarrow 0, \text{ and } \frac{\ln(s_\lambda)}{r_\lambda^2(\mathcal{V})} \rightarrow 0.$$

uniformly holds for $|\mathcal{V}| \geq c_1 \cdot v_\lambda$ and $\mathcal{V} \neq \mathcal{V}_\lambda$.

Therefore, when $|\mathcal{V}| \geq v_\lambda$ and $\mathcal{V} \neq \mathcal{V}_\lambda$, we have

$$2e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot e^{-c' \cdot \min\{v \cdot r_\lambda^2(\mathcal{V}), v \cdot r_\lambda(\mathcal{V})\}} \rightarrow 0.$$

When $c_1 \cdot v_\lambda \leq |\mathcal{V}| < v_\lambda$ and $\mathcal{V} \neq \mathcal{V}_\lambda$, we have

$$2e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot e^{-c \cdot \min\{v \cdot r_\lambda(\mathcal{V}), \frac{v^2 \cdot r_\lambda^2(\mathcal{V})}{v_\lambda}\}} \rightarrow 0.$$

Lemma 3 Under Condition 6 and 7,

$$\frac{\ln(s_\lambda)}{\kappa_n} \rightarrow 0,$$

Therefore we have

$$2e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot e^{-c' \cdot \kappa_n \cdot v_\lambda} \rightarrow 0.$$

Furthermore, under Condition 6 and Condition 7*, which is a stronger condition of κ_n , we have

$$2e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot e^{-c' \cdot \min\left\{\frac{\kappa_n^2}{v_\lambda}, \kappa_n\right\}} \rightarrow 0.$$

S.5.7 Proof of Lemmas

S.5.7.1 Proof of Lemma 1

We first prove u_j is a sub-Gaussian random variable. We know from the definition that the n -th moment of u_j conditional on selection is

$$\begin{aligned}\mathbb{E}[u_j^n | S_j > 0] &= \frac{\sigma_{X_j}^n}{\mathbb{P}[S_j > 0]} \int_{-\infty}^{\infty} \left(y - \frac{1}{\eta_j} \frac{\phi(B_{j,+}(y)) - \phi(B_{j,-}(y))}{1 - \Phi(B_{j,+}(y)) + \Phi(B_{j,-}(y))} \right)^n \\ &\quad \phi(y) [1 - \Phi(B_{j,+}(y)) + \Phi(B_{j,-}(y))] dy.\end{aligned}$$

where

$$\mathbb{P}[S_j > 0] = \Phi\left(\frac{-\lambda + \frac{\beta_{X_j}}{\sigma_{X_j}}}{\sqrt{1 + \eta_j^2}}\right) + \Phi\left(\frac{-\lambda - \frac{\beta_{X_j}}{\sigma_{X_j}}}{\sqrt{1 + \eta_j^2}}\right).$$

and

$$B_{j,\pm}(y) = -\left(\frac{\beta_{X_j}}{\sigma_{X_j} \eta_j} + \frac{y}{\eta_j}\right) \pm \frac{\lambda}{\eta_j}.$$

Given that the calculations can be quite involved, we let $\eta_j = 1$ and $\gamma_j = 0$ to streamline the presentation. That is, we consider

$$\frac{\mathbb{E}[u_{X_j}^n | S_j > 0]}{\sigma_{X_j}^n} = \frac{1}{\mathbb{P}[S_j > 0]} \int_{-\infty}^{\infty} \left(y - \frac{\phi(-\lambda + y) - \phi(-\lambda - y)}{\Phi(-\lambda + y) + \Phi(-\lambda - y)} \right)^n \phi(y) [\Phi(-\lambda + y) + \Phi(-\lambda - y)] dy.$$

Here

$$\mathbb{P}[S_j > 0] = 2 \cdot \Phi\left(\frac{-\lambda}{\sqrt{2}}\right).$$

Then when n is an odd number, $\frac{\mathbb{E}[u_{X_j}^n | S_j > 0]}{\sigma_{X_j}^n} = 0$. When n is an even number, we have

$$\frac{\mathbb{E}[u_{X_j}^n | S_j > 0]}{\sigma_{X_j}^n} = \frac{2}{\mathbb{P}[S_j > 0]} \int_0^{\infty} \left(y - \frac{\phi(-\lambda + y) - \phi(-\lambda - y)}{\Phi(-\lambda + y) + \Phi(-\lambda - y)} \right)^n \phi(y) [\Phi(-\lambda + y) + \Phi(-\lambda - y)] dy.$$

$$\frac{\phi(-\lambda+y) - \phi(-\lambda-y)}{\Phi(-\lambda+y) + \Phi(-\lambda-y)} \leq \frac{\phi(-\lambda+y) - \phi(-\lambda-y)}{\Phi(-\lambda) + \Phi(-\lambda)} \leq \frac{\phi(-\lambda+y)}{\Phi(-\lambda) + \Phi(-\lambda)} \leq \frac{1/\sqrt{2\pi}}{\Phi(-\lambda) + \Phi(-\lambda)}.$$

$$\begin{aligned}
\frac{\mathbb{E}[u_{X_j}^n | S_j > 0]}{\sigma_{X_j}^n} &= \frac{2}{\mathbb{P}[S_j > 0]} \int_0^{\frac{1/\sqrt{2\pi}}{\Phi(-\lambda) + \Phi(-\lambda)}} \left(y - \frac{\phi(-\lambda+y) - \phi(-\lambda-y)}{\Phi(-\lambda+y) + \Phi(-\lambda-y)} \right)^n \phi(y) [\Phi(-\lambda+y) + \Phi(-\lambda-y)] dy \\
&\quad + \frac{2}{\mathbb{P}[S_j > 0]} \int_{\frac{1/\sqrt{2\pi}}{\Phi(-\lambda) + \Phi(-\lambda)}}^{\infty} \left(y - \frac{\phi(-\lambda+y) - \phi(-\lambda-y)}{\Phi(-\lambda+y) + \Phi(-\lambda-y)} \right)^n \phi(y) [\Phi(-\lambda+y) + \Phi(-\lambda-y)] dy \\
&\leq \frac{2}{\mathbb{P}[S_j > 0]} \int_0^{\frac{1/\sqrt{2\pi}}{\Phi(-\lambda) + \Phi(-\lambda)}} \left(\frac{1/\sqrt{2\pi}}{\Phi(-\lambda) + \Phi(-\lambda)} \right)^n \phi(y) [\Phi(-\lambda+y) + \Phi(-\lambda-y)] dy \\
&\quad + \frac{2}{\mathbb{P}[S_j > 0]} \int_{\frac{1/\sqrt{2\pi}}{\Phi(-\lambda) + \Phi(-\lambda)}}^{\infty} y^n \phi(y) [\Phi(-\lambda+y) + \Phi(-\lambda-y)] dy \\
&\leq \frac{2}{\mathbb{P}[S_j > 0]} \int_0^{\infty} \left(\frac{1/\sqrt{2\pi}}{\Phi(-\lambda) + \Phi(-\lambda)} \right)^n \phi(y) [\Phi(-\lambda+y) + \Phi(-\lambda-y)] dy \\
&\quad + \frac{2}{\mathbb{P}[S_j > 0]} \int_0^{\infty} y^n \phi(y) [\Phi(-\lambda+y) + \Phi(-\lambda-y)] dy \\
&= 2 \left(\frac{1/\sqrt{2\pi}}{\Phi(-\lambda) + \Phi(-\lambda)} \right)^n + \frac{2}{\mathbb{P}[S_j > 0]} \int_0^{\infty} y^n \phi(y) [\Phi(-\lambda+y) + \Phi(-\lambda-y)] dy \\
&\leq \left(\frac{1/\sqrt{2\pi}}{\Phi(-\lambda) + \Phi(-\lambda)} \right)^n + \frac{1}{\mathbb{P}[S_j > 0]} \int_{-\infty}^{\infty} y^n \phi(y) dy.
\end{aligned}$$

Then by the property of sub-Gaussian random variable, we know that there exists a K such that

$$\left(\int_{-\infty}^{\infty} y^n \phi(y) dy \right)^{\frac{1}{n}} \leq K\sqrt{n}.$$

for any $n \geq 1$.

$$\frac{\mathbb{E}[u_{X_j}^n | S_j > 0]}{\sigma_{X_j}^n} \leq \left(\frac{1/\sqrt{2\pi}}{\Phi(-\lambda) + \Phi(-\lambda)} \right)^n + \frac{1}{2 \cdot \Phi\left(\frac{-\lambda}{\sqrt{2}}\right)} \cdot K^n \cdot n^{\frac{n}{2}}.$$

Then we can know that there exists a $K' > 0$ such that

$$\left(\frac{\mathbb{E}[u_{X_j}^n | S_j > 0]}{\sigma_{X_j}^n} \right)^{\frac{1}{n}} \leq K' \cdot \sqrt{n}.$$

for any $n \geq 1$.

This proves that conditional on re-randomized selection, u_j is a sub-Gaussian random variable.

Since ν_j is a Gaussian random variable, it's also a sub-Gaussian random variable. Then we know

$\nu_j u_j$, u_j^2 and ν_j^2 are all subexponential random variables.

Also we know from Condition 1 that $\frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}$ is a sub-exponential random variable.

Then by Bernstein's inequality, we have

$$\begin{aligned} \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j u_j}{\sigma_{Y_j}^2}\right| \geq t\right) &\leq 2 \cdot e^{-c \cdot \min\left(\frac{t^2}{\sum_{j \in \mathcal{V}_\lambda} \|\frac{\nu_j u_j}{\sigma_{Y_j}^2}\|_{\psi_1}^2}, \frac{t}{\max_i \|\frac{\nu_j u_j}{\sigma_{Y_j}^2}\|_{\psi_1}}\right)} \\ &\leq 2 \cdot e^{-c \cdot \min\left(\frac{t^2}{\sum_{j \in \mathcal{V}_\lambda} \|\frac{\nu_j}{\sigma_{Y_j}^2}\|_{\psi_2}^2 \cdot \|\frac{u_j}{\sigma_{Y_j}^2}\|_{\psi_2}^2}, \frac{t}{\max_i \sqrt{\|\frac{\nu_j}{\sigma_{Y_j}^2}\|_{\psi_2}^2 \cdot \|\frac{u_j}{\sigma_{Y_j}^2}\|_{\psi_2}^2}}\right)}. \end{aligned}$$

$$\begin{aligned} \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}} \frac{\nu_j u_j}{\sigma_{Y_j}^2}\right| \geq t\right) &\leq 2 \cdot e^{-c \cdot \min\left(\frac{t^2}{\sum_{j \in \mathcal{V}} \|\frac{\nu_j u_j}{\sigma_{Y_j}^2}\|_{\psi_1}^2}, \frac{t}{\max_i \|\frac{\nu_j u_j}{\sigma_{Y_j}^2}\|_{\psi_1}}\right)} \\ &\leq 2 \cdot e^{-c \cdot \min\left(\frac{t^2}{\sum_{j \in \mathcal{V}} \|\frac{\nu_j}{\sigma_{Y_j}^2}\|_{\psi_2}^2 \cdot \|\frac{u_j}{\sigma_{Y_j}^2}\|_{\psi_2}^2}, \frac{t}{\max_i \sqrt{\|\frac{\nu_j}{\sigma_{Y_j}^2}\|_{\psi_2}^2 \cdot \|\frac{u_j}{\sigma_{Y_j}^2}\|_{\psi_2}^2}}\right)}. \end{aligned}$$

$$\begin{aligned} \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2}\right| \geq t\right) &\leq 2 \cdot e^{-c \cdot \min\left(\frac{t^2}{\sum_{j \in \mathcal{V}_\lambda} \|\frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2}\|_{\psi_1}^2}, \frac{t}{\max_i \|\frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2}\|_{\psi_1}}\right)} \\ &\leq 2 \cdot e^{-c \cdot \min\left(\frac{t^2}{\sum_{j \in \mathcal{V}_\lambda} \sqrt{\|\frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2}\|_{\psi_2}^4}}, \frac{t}{\max_i \sqrt{\|\frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2}\|_{\psi_2}^2}}\right)}. \end{aligned}$$

$$\begin{aligned} \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2}\right| \geq t\right) &\leq 2 \cdot e^{-c \cdot \min\left(\frac{t^2}{\sum_{j \in \mathcal{V}} \|\frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2}\|_{\psi_1}^2}, \frac{t}{\max_i \|\frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2}\|_{\psi_1}}\right)} \\ &\leq 2 \cdot e^{-c \cdot \min\left(\frac{t^2}{\sum_{j \in \mathcal{V}} \sqrt{\|\frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2}\|_{\psi_2}^4}}, \frac{t}{\max_i \sqrt{\|\frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2}\|_{\psi_2}^2}}\right)}. \end{aligned}$$

$$\begin{aligned}
& \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}_\lambda} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq t\right) \leq 2 \cdot e^{-c \cdot \min\left(\frac{t^2}{\sum_{j \in \mathcal{V}_\lambda} \|\frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\|_{\psi_1}^2}, \frac{t}{\max_i \|\frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\|_{\psi_1}}\right)} \\
& \leq 2 \cdot e^{-c \cdot \min\left(\frac{t^2}{\sum_{j \in \mathcal{V}_\lambda} \|\frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\|_{\psi_2}^4}, \frac{t}{\sqrt{\sum_{j \in \mathcal{V}_\lambda} \|\frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\|_{\psi_2}^4}}\right)}.
\end{aligned}$$

$$\begin{aligned}
& \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq t\right) \leq 2 \cdot e^{-c \cdot \min\left(\frac{t^2}{\sum_{j \in \mathcal{V}} \|\frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\|_{\psi_1}^2}, \frac{t}{\max_i \|\frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\|_{\psi_1}}\right)} \\
& \leq 2 \cdot e^{-c \cdot \min\left(\frac{t^2}{\sum_{j \in \mathcal{V}} \|\frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\|_{\psi_2}^4}, \frac{t}{\sqrt{\sum_{j \in \mathcal{V}} \|\frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\|_{\psi_2}^4}}\right)}.
\end{aligned}$$

$$\begin{aligned}
& \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}_\lambda} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq t\right) \leq 2 \cdot e^{-c \cdot \min\left(\frac{t^2}{\sum_{j \in \mathcal{V}_\lambda} \|\frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\|_{\psi_1}^2}, \frac{t}{\max_i \|\frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\|_{\psi_1}}\right)} \\
& \leq 2 \cdot e^{-c \cdot \min\left(\frac{t^2}{\sum_{j \in \mathcal{V}_\lambda} \|\frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\|_{\psi_2}^4}, \frac{t}{\sqrt{\sum_{j \in \mathcal{V}_\lambda} \|\frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\|_{\psi_2}^4}}\right)}.
\end{aligned}$$

$$\begin{aligned}
& \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq t\right) \leq 2 \cdot e^{-c \cdot \min\left(\frac{t^2}{\sum_{j \in \mathcal{V}} \|\frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\|_{\psi_1}^2}, \frac{t}{\max_i \|\frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\|_{\psi_1}}\right)} \\
& \leq 2 \cdot e^{-c \cdot \min\left(\frac{t^2}{\sum_{j \in \mathcal{V}} \|\frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\|_{\psi_2}^4}, \frac{t}{\sqrt{\sum_{j \in \mathcal{V}} \|\frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\|_{\psi_2}^4}}\right)}.
\end{aligned}$$

Under Condition 1, we can have the conclusion in Lemma 1.

$$\mathbb{P}\left(\left|\sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j u_j}{\sigma_{Y_j}^2}\right| \geq t\right) \leq 2 \cdot e^{-c \cdot \min\left\{\frac{t^2}{v_\lambda}, t\right\}}, \quad \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}} \frac{\nu_j u_j}{\sigma_{Y_j}^2}\right| \geq t\right) \leq 2 \cdot e^{-c \cdot \min\left\{\frac{t^2}{v^*}, t\right\}}.$$

$$\mathbb{P}\left(\left|\sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2}\right| \geq t\right) \leq 2 \cdot e^{-c \cdot \min\left\{\frac{t^2}{v_\lambda}, t\right\}}, \quad \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2}\right| \geq t\right) \leq 2 \cdot e^{-c \cdot \min\left\{\frac{t^2}{v^*}, t\right\}}.$$

$$\mathbb{P}\left(\left|\sum_{j \in \mathcal{V}_\lambda} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq t\right) \leq 2 \cdot e^{-c \cdot \min\left\{\frac{t^2}{v_\lambda}, t\right\}}, \quad \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq t\right) \leq 2 \cdot e^{-c \cdot \min\left\{\frac{t^2}{v^*}, t\right\}}.$$

$$\mathbb{P}\left(\left|\sum_{j \in \mathcal{V}_\lambda} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq t\right) \leq 2 \cdot e^{-c \cdot \min\left\{\frac{t^2}{v_\lambda}, t\right\}}, \quad \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq t\right) \leq 2 \cdot e^{-c \cdot \min\left\{\frac{t^2}{v^*}, t\right\}}.$$

S.5.7.2 Proof of Lemma 2

Proof. Without loss of generality, we can assume that $\sigma_{Y_j}^2 = \frac{1}{n}$, then we have

$$r_\lambda(\mathcal{V}) = \frac{1}{v} \sum_{j \in \mathcal{V}} \frac{r_j^2}{\sigma_{Y_j}^2} \geq \frac{1}{v} \cdot n \min_{j \in \mathcal{S}_\lambda, r_j \neq 0} r_j^2.$$

Under Condition 5, we have

$$\begin{aligned} \frac{\ln(s_\lambda)}{r_\lambda(\mathcal{V})} &= \frac{v \cdot \ln(s_\lambda)}{n \min_{j \in \mathcal{S}_\lambda, r_j \neq 0} r_j^2} \leq \frac{s_\lambda \cdot \ln(s_\lambda)}{n \min_{j \in \mathcal{S}_\lambda, r_j \neq 0} r_j^2} \rightarrow 0. \\ \frac{\ln(s_\lambda)}{r_\lambda^2(\mathcal{V})} &\leq \frac{\ln(s_\lambda) \cdot v^2}{n^2 \min_{j \in \mathcal{S}_\lambda, r_j \neq 0} r_j^4} \leq \frac{s_\lambda^2 \cdot \ln(s_\lambda)}{n^2 \min_{j \in \mathcal{S}_\lambda, r_j \neq 0} r_j^4} \rightarrow 0. \end{aligned}$$

Notice that when $|\mathcal{V}| \geq v_\lambda$ and $\mathcal{V} \neq \mathcal{V}_\lambda$, under Condition 6, we have

$$\frac{(s_\lambda + 1) \cdot \ln(s_\lambda)}{v \cdot r_\lambda(\mathcal{V})} \leq \frac{(s_\lambda + 1) \cdot \ln(s_\lambda)}{v_\lambda \cdot r_\lambda(\mathcal{V})} \rightarrow 0, \quad \frac{(s_\lambda + 1) \cdot \ln(s_\lambda)}{v \cdot r_\lambda^2(\mathcal{V})} \leq \frac{(s_\lambda + 1) \cdot \ln(s_\lambda)}{v_\lambda \cdot r_\lambda^2(\mathcal{V})} \rightarrow 0.$$

Then we can show

$$2e^{(s_\lambda + 1) \cdot \ln(s_\lambda)} \cdot e^{-c' \cdot \min\{v \cdot r_\lambda^2(\mathcal{V}), v \cdot r_\lambda(\mathcal{V})\}} \rightarrow 0.$$

Notice that when $c_1 \cdot v_\lambda \leq |\mathcal{V}| < v_\lambda$ and $\mathcal{V} \neq \mathcal{V}_\lambda$, under Condition 6, we have

$$\frac{(s_\lambda + 1) \cdot \ln(s_\lambda)}{v \cdot r_\lambda(\mathcal{V})} \leq \frac{(s_\lambda + 1) \cdot \ln(s_\lambda)}{c_1 \cdot v_\lambda \cdot r_\lambda(\mathcal{V})} \rightarrow 0, \quad \frac{(s_\lambda + 1) \cdot \ln(s_\lambda) \cdot v_\lambda}{v^2 \cdot r_\lambda^2(\mathcal{V})} \leq \frac{(s_\lambda + 1) \cdot \ln(s_\lambda)}{c_1^2 \cdot v_\lambda \cdot r_\lambda^2(\mathcal{V})} \rightarrow 0.$$

Then we can show

$$2e^{(s_\lambda + 1) \cdot \ln(s_\lambda)} \cdot e^{-c \cdot \min\left\{v \cdot r_\lambda(\mathcal{V}), \frac{v^2 \cdot r_\lambda^2(\mathcal{V})}{v_\lambda}\right\}} \rightarrow 0.$$

■

S.5.7.3 Proof of Lemma 3

Proof. Under Condition 6 and 7, we know that

$$\kappa_n \gg \ln(s_\lambda), \quad \text{and} \quad \frac{v_\lambda}{s_\lambda} \text{ is bounded away from 0.}$$

Then

$$\frac{(s_\lambda + 1) \cdot \ln(s_\lambda)}{\kappa_n \cdot v_\lambda} \rightarrow 0.$$

Then we know

$$2e^{(s_\lambda + 1) \cdot \ln(s_\lambda)} \cdot e^{-c' \kappa_n \cdot v_\lambda} \rightarrow 0.$$

Under Condition 6 and Condition 7*, we have

$$\frac{(s_\lambda + 1) \cdot \ln(s_\lambda)}{\kappa_n} \rightarrow 0.$$

Then we know

$$2e^{(s_\lambda + 1) \cdot \ln(s_\lambda)} \cdot e^{-c' \cdot \min\left\{\frac{\kappa_n^2}{v_\lambda}, \kappa_n\right\}} \rightarrow 0.$$

■

S.6 An example that Assumption S3 is satisfied without perfect screening

S.6.1 Main results

With a slight abuse of notation, we consider a special case where the instruments can be divided into three clusters:

$$\mathcal{V}_1 = \{j : r_j = 0, \beta_{X_j} = \beta_0\} \text{ with } |\mathcal{V}_1| = v_1,$$

$$\mathcal{V}_2 = \{j : r_j = r_2, \beta_{X_j} = \beta_0\} \text{ with } |\mathcal{V}_2| = v_2,$$

$$\mathcal{V}_3 = \{j : r_j = r_3, \beta_{X_j} = \beta_0\} \text{ with } |\mathcal{V}_3| = v_3.$$

Here, \mathcal{V}_1 represents the set of valid IVs with $r_j = 0$, and \mathcal{V}_2 represents the set of invalid IVs with vanishing pleiotropic effects with r_2 tending to zero at an appropriate rate (see Lemma 5 and Theorem S3 for its precise characterization), and \mathcal{V}_3 represents the set of invalid IVs with non-vanishing pleiotropic effects. We note that it is not necessary to restrict all β_{X_j} 's to have the same magnitude, and our results presented in this section can be extended to cases where the standardized IV strength lies in a neighborhood of β_0/σ_{X_j} , in the sense that $\frac{\beta_{X_j}}{\sigma_{X_j}} \in \left[\frac{\beta_0}{\sigma_{X_j}} \pm \delta \times \frac{\beta_0}{\sigma_{X_j}} \right]$ with δ tending to zero.

To further simplify our theoretical derivation, we consider $\sigma_{Y_j}^2 = \frac{1}{n}$ for all $j \in \mathcal{S}_\lambda$. Next, for any subset $\mathcal{V} \subseteq \mathcal{S}_\lambda$, we further define the following quantities:

$$p_1(\mathcal{V}) = \frac{|\mathcal{V}_1 \cap \mathcal{V}|}{|\mathcal{V}|}, \quad p_2(\mathcal{V}) = \frac{|\mathcal{V}_2 \cap \mathcal{V}|}{|\mathcal{V}|}, \quad p_3(\mathcal{V}) = \frac{|\mathcal{V}_3 \cap \mathcal{V}|}{|\mathcal{V}|}.$$

Lastly, we denote $A_n = (\ln(s_\lambda) \vee \kappa_n) \cdot \sqrt{s_\lambda/n}$.

In what follows, we will argue that to satisfy Assumption 3, our invalid IV screening procedure does not need to have a perfect screening property. In other words, our estimator remains asymptotically unbiased even if our IV screening procedure does not select \mathcal{V}_1 with probability approaching one. As shall be made clear in Lemma 4 and Lemma 5, our method avoids the need

for perfect IV screening by showing that the selected IV set $\widehat{\mathcal{V}}$ can include both invalid IVs from \mathcal{V}_2 and a vanishing portion of invalid IVs from \mathcal{V}_3 in the selected set $\widehat{\mathcal{V}}$.

We impose the following conditions:

Condition 8 (The order of the number of valid IVs) *The number of valid IVs v_1 is of the same order as s_λ . For the number of invalid IVs, there exists a positive constant $c_1 \in (0, 1)$ such that $(v_2/v_1 \vee v_3/v_1) \leq c_1$.*

The above condition requires that the majority of the IVs included in MR are valid IVs. The next condition is needed so that our optimization problem does not suffer from potential over-fitting issues in high-dimensional settings:

Condition 9 (High-dimensional BIC) $\kappa_n \gg \ln(s_\lambda)$.

Next, for any given $\varepsilon > 0$, define a collection of sets

$$\begin{aligned} \mathcal{V}_{\text{bias}}(\varepsilon) &= \mathcal{V}(\varepsilon) \cup \mathcal{V}_{\text{BIC}} \\ &= \left\{ \mathcal{V} \mid \mathcal{V} \subseteq \mathcal{S}_\lambda, p_3(\mathcal{V}) \geq \frac{\varepsilon}{\sqrt{ns_\lambda} \cdot r_3} \right\} \cup \left\{ \mathcal{V} \mid \mathcal{V} \subseteq \mathcal{S}_\lambda, |\mathcal{V}| = v < \frac{1+c_1}{2} \cdot v_1 \right\}. \end{aligned}$$

$\mathcal{V}_{\text{bias}}(\varepsilon)$ is a union of two types of sets that will be screened out by our invalid IV screening procedure. The first set $\mathcal{V}(\varepsilon)$ consists of all possible sets with a non-vanishing proportion of IVs in \mathcal{V}_3 , defined by the condition $p_3(\mathcal{V}) \geq \frac{\varepsilon}{\sqrt{ns_\lambda} \cdot r_3}$. Consequently, if our selected set $\widehat{\mathcal{V}}$ belongs to $\mathcal{V}(\varepsilon)$, the resulting causal effect estimator is biased. The second set \mathcal{V}_{BIC} comprises all sets containing a total number of IVs smaller than v_1 . As our IV screening procedure adopts l_0 penalty with BIC to screen out invalid IVs, our selected set $\widehat{\mathcal{V}}$ tends to select an IV set with cardinality larger than v_1 . Therefore, $\widehat{\mathcal{V}}$ does not belong to \mathcal{V}_{BIC} as well. The following lemma provides rigorous statement about our selected IV set $\widehat{\mathcal{V}}$:

Lemma 4 *For any given $\varepsilon > 0$, if r_3 is sufficiently large in the sense that $A_n/(r_3\varepsilon) = o(1)$ and $|\widehat{\theta}(\mathcal{V})|$ is bounded by a constant for all $\mathcal{V} \in \mathcal{S}_\lambda$, then under Condition 1, 2, 8 and 9, the selected IV set $\widehat{\mathcal{V}}$ using our procedure satisfies $\mathbb{P}(\widehat{\mathcal{V}} \in \mathcal{V}_{\text{bias}}(\varepsilon)) \rightarrow 0$.*

Next, we demonstrate that when r_2 tends to zero at an appropriate rate and β_0 are sufficiently large, for the set \mathcal{V} that does not fall into $\mathcal{V}_{\text{bias}}(\varepsilon)$, the bias term described in Assumption 3 is asymptotically negligible:

Lemma 5 *We choose $a_\lambda \asymp s_\lambda \cdot \kappa_\lambda = ns_\lambda \cdot \beta_0^2$ to stabilize the variance (other choices for a_λ can also be adopted). For any given $\varepsilon > 0$, whenever $r_2 < \frac{\varepsilon}{\sqrt{ns_\lambda}}$ and $\frac{r_3}{\beta_0^2 \varepsilon \cdot \sqrt{ns_\lambda}} = o(1)$, under Condition 1 and 2, we have*

$$\left| \frac{a_\lambda}{\sqrt{ns_\lambda} \cdot \kappa_\lambda} \cdot \frac{\sum_{j \in \mathcal{V}} r_j \cdot \hat{\beta}_{X_{j,\text{RB}}}}{\sum_{j \in \mathcal{V}} \hat{\beta}_{X_{j,\text{RB}}}^2 - \hat{\sigma}_{X_{j,\text{RB}}}^2} \right| = O_p(\varepsilon).$$

for any $\mathcal{V} \notin \mathcal{V}_{bias}(\varepsilon)$.

In Lemma 5, ε can tend to zero at different rates, each affecting the conditions on r_2 , r_3 , and β_0 differently. To cast some insights into this result, we consider a simple example. For a positive constant δ , we let

$$\varepsilon = \frac{1}{\ln^\delta(s_\lambda)} = o(1), \quad \kappa_n = \ln^{1+\delta}(s_\lambda) \gg \ln(s_\lambda), \quad r_3 = \ln^{1+3\delta}(s_\lambda) \cdot \sqrt{\frac{s_\lambda}{n}}.$$

If r_2 and β_0 satisfy

$$r_2 < \frac{1}{\ln^\delta(s_\lambda) \cdot \sqrt{ns_\lambda}}, \quad \beta_0 \gg \sqrt{\frac{\ln^{1+4\delta}(s_\lambda)}{n}},$$

the conditions of Lemma 5 are met. Here, the above requirement on the magnitude of β_0 is rather mild, as the selected IV strength in \mathcal{S}_λ typically has an order greater than $\sqrt{\log p/n}$, since the cut-off value λ is often of the order $\sqrt{\log p}$. In practice, since relevant IVs often constitute only a small fraction of all candidate IVs, s_λ should be a term of smaller order compared to p . Therefore, the condition $\beta_0 \gg \sqrt{\ln^{1+4\delta}(s_\lambda)/n}$ that we impose here is rather mild.

With these two lemmas, we are ready to show that the set $\hat{\mathcal{V}}$ selected by our proposed invalid screening procedure induces negligible bias:

Theorem S3 *For a vanishing number $\varepsilon > 0$, we assume that*

- (i) *r_3 is sufficiently large in the sense that $A_n/(r_3\varepsilon) = o(1)$, which ensures our invalid screening procedure to effectively screen out IVs from \mathcal{V}_3 .*
- (ii) *r_2 is a vanishing number in the sense that $r_2 < \frac{\varepsilon}{\sqrt{ns_\lambda}}$, which ensures IVs from \mathcal{V}_2 to have vanishing pleiotrophic effects.*
- (iii) *β_0 is sufficiently large in the sense that $\frac{\sqrt{s_\lambda} \ln(s_\lambda)}{\sqrt{n}} \frac{r_3}{\varepsilon \beta_0^2} \rightarrow 0$.*

If $|\widehat{\theta}(\mathcal{V})|$ is bounded by a constant for all $\mathcal{V} \in \mathcal{S}_\lambda$, under Condition 1, 2, 8 and 9, choosing $a_\lambda \asymp s_\lambda \cdot \kappa_\lambda = ns_\lambda \cdot \beta_0^2$ to stabilize the variance, we can prove that

$$\frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \cdot \frac{\sum_{j \in \widehat{\mathcal{V}}} r_j \cdot \widehat{\beta}_{X_{j,\text{RB}}}}{\sum_{j \in \widehat{\mathcal{V}}} \widehat{\beta}_{X_{j,\text{RB}}}^2 - \widehat{\sigma}_{X_{j,\text{RB}}}^2} = o_p(1).$$

We note that the third condition in the above theorem is slightly stronger than what was assumed in Lemma 6, as we applied a union bound, needed to account for uniformity across all possible subsets of \mathcal{S}_λ . The conditions we impose here are sufficient but by no means necessary.

S.6.2 Proof of Theorem S3

For any given $\varepsilon > 0$, we have

$$\begin{aligned} \frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \cdot \frac{\sum_{j \in \widehat{\mathcal{V}}} r_j \cdot \widehat{\beta}_{X_{j,\text{RB}}}}{\sum_{j \in \widehat{\mathcal{V}}} \widehat{\beta}_{X_{j,\text{RB}}}^2 - \widehat{\sigma}_{X_{j,\text{RB}}}^2} &= \sum_{\mathcal{V} \in \mathcal{V}_{\text{bias}}(\varepsilon)} \frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \cdot \frac{\sum_{j \in \mathcal{V}} r_j \cdot \widehat{\beta}_{X_{j,\text{RB}}}}{\sum_{j \in \mathcal{V}} \widehat{\beta}_{X_{j,\text{RB}}}^2 - \widehat{\sigma}_{X_{j,\text{RB}}}^2} \cdot 1(\widehat{\mathcal{V}} = \mathcal{V}) \\ &\quad + \sum_{\mathcal{V} \notin \mathcal{V}_{\text{bias}}(\varepsilon)} \frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \cdot \frac{\sum_{j \in \mathcal{V}} r_j \cdot \widehat{\beta}_{X_{j,\text{RB}}}}{\sum_{j \in \mathcal{V}} \widehat{\beta}_{X_{j,\text{RB}}}^2 - \widehat{\sigma}_{X_{j,\text{RB}}}^2} \cdot 1(\widehat{\mathcal{V}} = \mathcal{V}) \end{aligned}$$

For any $\varepsilon_0 > 0$, the first term in the right hand side can be bounded using the following inequality:

$$\mathbb{P}\left(\left|\sum_{\mathcal{V} \in \mathcal{V}_{\text{bias}}(\varepsilon)} \frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \cdot \frac{\sum_{j \in \mathcal{V}} r_j \cdot \widehat{\beta}_{X_{j,\text{RB}}}}{\sum_{j \in \mathcal{V}} \widehat{\beta}_{X_{j,\text{RB}}}^2 - \widehat{\sigma}_{X_{j,\text{RB}}}^2} \cdot 1(\widehat{\mathcal{V}} = \mathcal{V})\right| > \varepsilon_0\right) \leq \mathbb{P}(\cup_{\mathcal{V} \in \mathcal{V}_{\text{bias}}(\varepsilon)} \{\widehat{\mathcal{V}} = \mathcal{V}\}) = \mathbb{P}(\widehat{\mathcal{V}} \in \mathcal{V}_{\text{bias}}(\varepsilon))$$

By using the result in Lemma 4 and letting $\varepsilon_0 \rightarrow 0$, we are able to show

$$\sum_{\mathcal{V} \in \mathcal{V}_{\text{bias}}(\varepsilon)} \frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \cdot \frac{\sum_{j \in \mathcal{V}} r_j \cdot \widehat{\beta}_{X_{j,\text{RB}}}}{\sum_{j \in \mathcal{V}} \widehat{\beta}_{X_{j,\text{RB}}}^2 - \widehat{\sigma}_{X_{j,\text{RB}}}^2} \cdot 1(\widehat{\mathcal{V}} = \mathcal{V}) = o_p(1).$$

Thus it suffices to show that the second term on the right-hand side satisfies

$$\sum_{\mathcal{V} \notin \mathcal{V}_{\text{bias}}(\varepsilon)} \frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \cdot \frac{\sum_{j \in \mathcal{V}} r_j \cdot \widehat{\beta}_{X_{j,\text{RB}}}}{\sum_{j \in \mathcal{V}} \widehat{\beta}_{X_{j,\text{RB}}}^2 - \widehat{\sigma}_{X_{j,\text{RB}}}^2} \cdot 1(\widehat{\mathcal{V}} = \mathcal{V}) = o_p(1).$$

In Lemma 5, we show that under the event $\mathcal{A}(\mathcal{V}, \varepsilon)$, we have

$$\left| \frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \cdot \frac{\sum_{j \in \mathcal{V}} r_j \cdot \hat{\beta}_{X_{j, \text{RB}}}}{\sum_{j \in \mathcal{V}} \hat{\beta}_{X_{j, \text{RB}}}^2 - \hat{\sigma}_{X_{j, \text{RB}}}^2} \right| < 9\varepsilon$$

for any given $\mathcal{V} \notin \mathcal{V}_{\text{bias}}(\varepsilon)$.

Under the event $\bigcap_{\mathcal{V} \notin \mathcal{V}_{\text{bias}}(\varepsilon)} \mathcal{A}(\mathcal{V}, \varepsilon)$, we can show that this holds uniformly for all $\mathcal{V} \notin \mathcal{V}_{\text{bias}}(\varepsilon)$.

Thus we have

$$\sum_{\mathcal{V} \notin \mathcal{V}_{\text{bias}}(\varepsilon)} \frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \cdot \frac{\sum_{j \in \mathcal{V}} r_j \cdot \hat{\beta}_{X_{j, \text{RB}}}}{\sum_{j \in \mathcal{V}} \hat{\beta}_{X_{j, \text{RB}}}^2 - \hat{\sigma}_{X_{j, \text{RB}}}^2} \cdot 1(\hat{\mathcal{V}} = \mathcal{V}) < 9\varepsilon.$$

We also notice that

$$\mathbb{P}\left(\bigcap_{\mathcal{V} \notin \mathcal{V}_{\text{bias}}(\varepsilon)} \mathcal{A}(\mathcal{V}, \varepsilon)\right) = 1 - \mathbb{P}\left(\bigcup_{\mathcal{V} \notin \mathcal{V}_{\text{bias}}(\varepsilon)} \mathcal{A}^c(\mathcal{V}, \varepsilon)\right) \geq 1 - \sum_{\mathcal{V} \notin \mathcal{V}_{\text{bias}}(\varepsilon)} \mathbb{P}(\mathcal{A}^c(\mathcal{V}, \varepsilon))$$

Here $\mathcal{A}^c(\mathcal{V}, \varepsilon)$ is the complement of the event $\mathcal{A}(\mathcal{V}, \varepsilon)$.

To prove $\mathbb{P}\left(\bigcap_{\mathcal{V} \notin \mathcal{V}_{\text{bias}}(\varepsilon)} \mathcal{A}(\mathcal{V}, \varepsilon)\right) \rightarrow 1$, we only need to show

$$\sum_{\mathcal{V} \notin \mathcal{V}_{\text{bias}}(\varepsilon)} \mathbb{P}(\mathcal{A}^c(\mathcal{V}, \varepsilon)) \leq e^{(s_\lambda + 1) \cdot \ln(s_\lambda)} \max_{\mathcal{V} \notin \mathcal{V}_{\text{bias}}(\varepsilon)} \mathbb{P}(\mathcal{A}^c(\mathcal{V}, \varepsilon)) \rightarrow 0.$$

In Lemma 5, we have $\mathbb{P}(\mathcal{A}(\mathcal{V}, \varepsilon)) \geq 1 - 2 \cdot e^{-\frac{c}{16} n \cdot v \cdot \beta_0^2} - 4 \cdot e^{-c \cdot \min\{\frac{1}{16} \cdot n^2 \cdot v \beta_0^4, \frac{1}{4} \cdot n \cdot v \beta_0^2\}} - 2 \cdot e^{\frac{-c \cdot \varepsilon^2 \cdot \beta_0^2}{p_2(\mathcal{V})r_2^2 + p_3(\mathcal{V})r_3^2} \frac{v}{16s_\lambda}}$, thus $\mathbb{P}(\mathcal{A}^c(\mathcal{V}, \varepsilon)) < 2 \cdot e^{-\frac{c}{16} n \cdot v \cdot \beta_0^2} + 4 \cdot e^{-c \cdot \min\{\frac{1}{16} \cdot n^2 \cdot v \beta_0^4, \frac{1}{4} \cdot n \cdot v \beta_0^2\}} + 2 \cdot e^{\frac{-c \cdot \varepsilon^2 \cdot \beta_0^2}{p_2(\mathcal{V})r_2^2 + p_3(\mathcal{V})r_3^2} \frac{v}{16s_\lambda}}$. In addition, we have $v > \frac{1+c_1}{2} \cdot v_1$ for any $\mathcal{V} \notin \mathcal{V}_{\text{bias}}(\varepsilon)$. Under Condition 8, we have $v \asymp s_\lambda$ uniformly hold for any $\mathcal{V} \notin \mathcal{V}_{\text{bias}}(\varepsilon)$.

With these results, we can prove $e^{(s_\lambda + 1) \cdot \ln(s_\lambda)} \max_{\mathcal{V} \notin \mathcal{V}_{\text{bias}}(\varepsilon)} \mathbb{P}(\mathcal{A}^c(\mathcal{V}, \varepsilon)) \rightarrow 0$ if we have $\frac{\sqrt{s_\lambda} \ln(s_\lambda)}{\sqrt{n}} \frac{r_3}{\varepsilon \beta_0^2} \rightarrow 0$ and $r_2 < \frac{\varepsilon}{\sqrt{ns_\lambda}}$.

When ε is a vanishing number, we can show

$$\sum_{\mathcal{V} \notin \mathcal{V}_{\text{bias}}(\varepsilon)} \frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \cdot \frac{\sum_{j \in \mathcal{V}} r_j \cdot \hat{\beta}_{X_{j, \text{RB}}}}{\sum_{j \in \mathcal{V}} \hat{\beta}_{X_{j, \text{RB}}}^2 - \hat{\sigma}_{X_{j, \text{RB}}}^2} \cdot 1(\hat{\mathcal{V}} = \mathcal{V}) = o_p(1).$$

Thus,

$$\frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \cdot \frac{\sum_{j \in \hat{\mathcal{V}}} r_j \cdot \hat{\beta}_{X_{j,\text{RB}}}}{\sum_{j \in \hat{\mathcal{V}}} \hat{\beta}_{X_{j,\text{RB}}}^2 - \hat{\sigma}_{X_{j,\text{RB}}}^2} = o_p(1).$$

Therefore, although our selection procedure does not have perfect screening properties, the set that we select still has negligible bias.

S.6.3 Proof of Lemma 4

For any $\mathcal{V} \subseteq \mathcal{S}_\lambda$, we denote a collection of sparse vectors

$$\mathcal{R}_V = \left\{ \mathbf{a} \in \mathbb{R}^{|\mathcal{S}_\lambda| \times 1} : a_j = 0, \text{ for } j \in \mathcal{V}, a_k \neq 0, \text{ for } k \in \mathcal{V}^c \right\}$$

and a function

$$\begin{aligned} h(\mathcal{V}, \theta) &= \min_{\mathbf{r} \in \mathcal{R}_V} \sum_{j \in \mathcal{S}_\lambda} \hat{l} \left(\theta, \mathbf{r}; \hat{\beta}_{Y_j}, \sigma_{Y_j}, \hat{\beta}_{X_{j,\text{RB}}}, \hat{\sigma}_{X_{j,\text{RB}}} \right) \\ &= \sum_{j \in \mathcal{V}} \frac{(\hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_{j,\text{RB}}})^2 - \theta^2 \cdot \hat{\sigma}_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}. \end{aligned}$$

For any given $\varepsilon > 0$, we define the set $\mathcal{V}_{\text{bias}}(\varepsilon) = \{\mathcal{V} | \mathcal{V} \subseteq \mathcal{S}_\lambda, p_3(\mathcal{V}) \geq \frac{\varepsilon}{\sqrt{n s_\lambda \cdot r_3}}\} \cup \{\mathcal{V} | \mathcal{V} \subseteq \mathcal{S}_\lambda, |\mathcal{V}| = v < \frac{1+c_1}{2} \cdot v_1\}$. Now we want to analyze $\mathbb{P}(\hat{\mathcal{V}} \in \mathcal{V}_{\text{bias}}(\varepsilon))$ by utilizing the following

inequality:

$$\begin{aligned}
\mathbb{P}(\widehat{\mathcal{V}} \in \mathcal{V}_{\text{bias}}(\varepsilon)) &\leq \mathbb{P}\left(\min_{v \in \mathbb{N}_+, v \leq s_\lambda} \left[\min_{|\mathcal{V}|=v, \mathcal{V} \in \mathcal{V}_{\text{bias}}(\varepsilon)} \min_{\theta \in \mathbb{R}} h(\mathcal{V}, \theta) - \kappa_n \cdot v \right] \leq \min_{\theta \in \mathbb{R}} h(\mathcal{V}_1, \theta) - \kappa_n \cdot v_1\right) \\
&\leq \bigcup_{\mathcal{V} \in \mathcal{V}_{\text{bias}}(\varepsilon)} \mathbb{P}\left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}, \theta) - \kappa_n \cdot |\mathcal{V}| \leq \min_{\theta \in \mathbb{R}} h(\mathcal{V}_1, \theta) - \kappa_n \cdot v_1\right) \\
&\leq \sum_{v=1}^{s_\lambda} \binom{s_\lambda}{v} \max_{|\mathcal{V}|=v, \mathcal{V} \in \mathcal{V}_{\text{bias}}(\varepsilon)} \mathbb{P}\left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}, \theta) - \kappa_n \cdot v \leq h(\mathcal{V}_1, \theta_0) - \kappa_n \cdot v_1\right) \\
&\leq \sum_{v=1}^{s_\lambda} s_\lambda^v \max_{|\mathcal{V}|=v, \mathcal{V} \in \mathcal{V}_{\text{bias}}(\varepsilon)} \mathbb{P}\left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}, \theta) - \kappa_n \cdot v \leq h(\mathcal{V}_1, \theta_0) - \kappa_n \cdot v_1\right) \\
&\leq \max_{\mathcal{V} \in \mathcal{V}_{\text{bias}}(\varepsilon)} e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot \mathbb{P}\left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}, \theta) - \kappa_n \cdot |\mathcal{V}| \leq h(\mathcal{V}_1, \theta_0) - \kappa_n \cdot v_1\right) \\
&= e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot \mathbb{P}\left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_1, \theta_0) - \kappa_n \cdot v_1\right).
\end{aligned} \tag{S4}$$

where $\mathcal{V}^* = \arg\max_{\mathcal{V} \in \mathcal{V}_{\text{bias}}(\varepsilon)} e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot \mathbb{P}\left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}, \theta) - \kappa_n \cdot |\mathcal{V}| \leq h(\mathcal{V}_1, \theta_0) - \kappa_n \cdot v_1\right)$ and $v^* = |\mathcal{V}^*|$.

As we show that

$$e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot \mathbb{P}\left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_1, \theta_0) - \kappa_n \cdot v_1\right) \rightarrow 0, \tag{S5}$$

then $\mathbb{P}(\widehat{\mathcal{V}} \in \mathcal{V}_{\text{bias}}(\varepsilon)) \rightarrow 0$ holds. Here, we also note that the first equation in Eq (S4) follows from the definition of the optimization problem defined in Equation 2 in the manuscript, the second to the fifth inequalities in Eq (S4) hold following $\min_{\theta \in \mathbb{R}} h(\mathcal{V}_1, \theta) \leq h(\mathcal{V}_1, \theta_0)$, $\binom{s_\lambda}{v} \leq s_\lambda^v$ and some basic calculations.

To prove formula (S5), we need to analyze the asymptotic properties of $h(\mathcal{V}_1, \theta_0)$, $\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta)$ and κ_n .

We start with $h(\mathcal{V}_1, \theta_0)$ and decompose it below following our notation defined in Section S.5.1

$$\begin{aligned}
h(\mathcal{V}_1, \theta_0) &= \sum_{j \in \mathcal{V}_1} \frac{(\widehat{\beta}_{Y_j} - \theta_0 \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} + \theta_0^2 \cdot \sum_{j \in \mathcal{V}_1} \frac{(\widehat{\beta}_{X_{j,\text{RB}}} - \beta_{X_j})^2 - \widehat{\sigma}_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \\
&\quad - 2\theta_0 \cdot \sum_{j \in \mathcal{V}_1} \frac{(\widehat{\beta}_{Y_j} - \theta_0 \cdot \beta_{X_j})(\widehat{\beta}_{X_{j,\text{RB}}} - \beta_{X_j})}{\sigma_{Y_j}^2} \\
&= \sum_{j \in \mathcal{V}_1} \frac{\nu_j^2}{\sigma_{Y_j}^2} + \theta_0^2 \cdot \sum_{j \in \mathcal{V}_1} \frac{u_j^2 - \widehat{\sigma}_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} - 2\theta_0 \cdot \sum_{j \in \mathcal{V}_1} \frac{\nu_j u_j}{\sigma_{Y_j}^2}.
\end{aligned} \tag{S6}$$

Next, we study the asymptotic property of $\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta)$. We denote $\widehat{\theta}(\mathcal{V}^*) = \arg \min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta)$ and decompose $h(\mathcal{V}^*, \widehat{\theta}(\mathcal{V}^*))$ in a similar way as $h(\mathcal{V}_1, \theta_0)$, following our notation in Section S.5.1.

$$\begin{aligned}
h(\mathcal{V}^*, \widehat{\theta}(\mathcal{V}^*)) &= \sum_{j \in \mathcal{V}^*} \frac{(\widehat{\beta}_{Y_j} - \widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} + \widehat{\theta}(\mathcal{V}^*)^2 \sum_{j \in \mathcal{V}^*} \frac{(\widehat{\beta}_{X_{j,\text{RB}}} - \beta_{X_j})^2 - \widehat{\sigma}_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \\
&\quad + 2\widehat{\theta}(\mathcal{V}^*) \sum_{j \in \mathcal{V}^*} \frac{(\widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot (\widehat{\beta}_{X_{j,\text{RB}}} - \beta_{X_j})}{\sigma_{Y_j}^2} \\
&\quad - 2\widehat{\theta}(\mathcal{V}^*) \sum_{j \in \mathcal{V}^*} \frac{(\widehat{\beta}_{Y_j} - \theta_0 \cdot \beta_{X_j} - r_j)(\widehat{\beta}_{X_{j,\text{RB}}} - \beta_{X_j})}{\sigma_{Y_j}^2} \\
&= \sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j + \nu_j - \widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} + \widehat{\theta}(\mathcal{V}^*)^2 \sum_{j \in \mathcal{V}^*} \frac{u_j^2 - \widehat{\sigma}_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \\
&\quad + 2\widehat{\theta}(\mathcal{V}^*) \sum_{j \in \mathcal{V}^*} \frac{(\widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot u_j}{\sigma_{Y_j}^2} - 2\widehat{\theta}(\mathcal{V}^*) \sum_{j \in \mathcal{V}^*} \frac{\nu_j u_j}{\sigma_{Y_j}^2} \\
&= \sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} + \sum_{j \in \mathcal{V}^*} \frac{\nu_j^2}{\sigma_{Y_j}^2} + 2 \sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j}) \cdot \nu_j}{\sigma_{Y_j}^2} \\
&\quad + \widehat{\theta}(\mathcal{V}^*)^2 \sum_{j \in \mathcal{V}^*} \frac{u_j^2 - \widehat{\sigma}_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} + 2\widehat{\theta}(\mathcal{V}^*) \sum_{j \in \mathcal{V}^*} \frac{(\widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot u_j}{\sigma_{Y_j}^2} - 2\widehat{\theta}(\mathcal{V}^*) \sum_{j \in \mathcal{V}^*} \frac{\nu_j u_j}{\sigma_{Y_j}^2}.
\end{aligned}$$

With these decompositions, the Equation (2) can be rewritten as

$$\begin{aligned}
&\mathbb{P}\left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_1, \theta_0) - \kappa_n \cdot v_1\right) \\
&= \mathbb{P}\left(\sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} + \kappa_n \cdot (v_1 - v^*) \leq - \sum_{j \in \mathcal{V}^*} \frac{\nu_j^2}{\sigma_{Y_j}^2} - 2 \sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j}) \cdot \nu_j}{\sigma_{Y_j}^2}\right. \\
&\quad - \widehat{\theta}(\mathcal{V}^*)^2 \sum_{j \in \mathcal{V}^*} \frac{u_j^2 - \widehat{\sigma}_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} + \widehat{\theta}(\mathcal{V}^*)^2 \sum_{j \in \mathcal{V}^*} \frac{\widehat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} - 2\widehat{\theta}(\mathcal{V}^*) \sum_{j \in \mathcal{V}^*} \frac{(\widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot u_j}{\sigma_{Y_j}^2} \\
&\quad \left. + 2\widehat{\theta}(\mathcal{V}^*) \sum_{j \in \mathcal{V}^*} \frac{\nu_j u_j}{\sigma_{Y_j}^2} + \sum_{j \in \mathcal{V}_1} \frac{\nu_j^2}{\sigma_{Y_j}^2} + \theta_0^2 \sum_{j \in \mathcal{V}_1} \frac{u_j^2 - \widehat{\sigma}_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} - \theta_0^2 \sum_{j \in \mathcal{V}_1} \frac{\widehat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} - 2\theta_0 \sum_{j \in \mathcal{V}_1} \frac{\nu_j u_j}{\sigma_{Y_j}^2}\right).
\end{aligned}$$

By some calculations, we can see that

$$\sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} \geq \min_{\theta \in \mathbb{R}} \sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \theta \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} = \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2} - \frac{\left(\sum_{j \in \mathcal{V}^*} \frac{r_j \cdot \beta_{X_j}}{\sigma_{Y_j}^2}\right)^2}{\sum_{j \in \mathcal{V}^*} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}}.$$

with probability 1.

Let

$$\Delta(\mathcal{V}^*) = \frac{1}{v^*} \left(\sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2} - \frac{\left(\sum_{j \in \mathcal{V}^*} \frac{r_j \cdot \beta_{X_j}}{\sigma_{Y_j}^2} \right)^2}{\sum_{j \in \mathcal{V}^*} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}} \right), \quad \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2} - \frac{\left(\sum_{j \in \mathcal{V}^*} \frac{r_j \cdot \beta_{X_j}}{\sigma_{Y_j}^2} \right)^2}{\sum_{j \in \mathcal{V}^*} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}} = v^* \cdot \Delta(\mathcal{V}^*).$$

When $\mathcal{V}^* \in \mathcal{V}_{\text{bias}}(\varepsilon)$ such that $|\mathcal{V}^*| = v^* \geq \frac{1+c_1}{2} \cdot v_1$ and $\mathcal{V}^* \neq \mathcal{V}_1$, under Condition 4, there exists a $C_0 > 0$ such that

$$\begin{aligned} & \mathbb{P} \left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_1, \theta_0) - \kappa_n \cdot v_1 \right) \\ & \leq \mathbb{P} \left(\sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} - \kappa_n \cdot (v^* - v_\lambda) \leq - \sum_{j \in \mathcal{V}^*} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} - \hat{\theta}(\mathcal{V}^*)^2 \sum_{j \in \mathcal{V}^*} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right. \\ & \quad \left. + \hat{\theta}(\mathcal{V}^*)^2 \sum_{j \in \mathcal{V}^*} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} - 2 \sum_{j \in \mathcal{V}^*} \frac{(\hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot \nu_j}{\sigma_{Y_j}^2} - 2\hat{\theta}(\mathcal{V}^*) \sum_{j \in \mathcal{V}^*} \frac{(\hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot u_j}{\sigma_{Y_j}^2} \right. \\ & \quad \left. + 2\hat{\theta}(\mathcal{V}^*) \sum_{j \in \mathcal{V}^*} \frac{\nu_j u_j}{\sigma_{Y_j}^2} + \sum_{j \in \mathcal{V}_1} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} + \theta_0^2 \sum_{j \in \mathcal{V}_1} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} - \theta_0^2 \sum_{j \in \mathcal{V}_1} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} - 2\theta_0 \sum_{j \in \mathcal{V}_1} \frac{\nu_j u_j}{\sigma_{Y_j}^2} \right) \\ & \leq \mathbb{P} \left(\sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} - \kappa_n \cdot (v^* - v_\lambda) \leq \left| \sum_{j \in \mathcal{V}^*} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} \right| + \hat{\theta}(\mathcal{V}^*)^2 \left| \sum_{j \in \mathcal{V}^*} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| \right. \\ & \quad \left. + 2\left| \hat{\theta}(\mathcal{V}^*) \sum_{j \in \mathcal{V}^*} \frac{(\hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot u_j}{\sigma_{Y_j}^2} \right| + 2 \left| \sum_{j \in \mathcal{V}^*} \frac{(\hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot \nu_j}{\sigma_{Y_j}^2} \right| + 2\left| \hat{\theta}(\mathcal{V}^*) \sum_{j \in \mathcal{V}^*} \frac{\nu_j u_j}{\sigma_{Y_j}^2} \right| \right. \\ & \quad \left. + \hat{\theta}(\mathcal{V}^*)^2 \left| \sum_{j \in \mathcal{V}^*} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + \left| \sum_{j \in \mathcal{V}_1} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} \right| + \theta_0^2 \left| \sum_{j \in \mathcal{V}_1} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + \theta_0^2 \left| \sum_{j \in \mathcal{V}_1} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + 2\theta_0 \left| \sum_{j \in \mathcal{V}_1} \frac{\nu_j u_j}{\sigma_{Y_j}^2} \right| \right) \\ & \leq \mathbb{P} \left(\sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} - \kappa_n \cdot (v^* - v_\lambda) \leq \left| \sum_{j \in \mathcal{V}^*} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} \right| + C_0^2 \left| \sum_{j \in \mathcal{V}^*} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| \right. \\ & \quad \left. + C_0^2 \left| \sum_{j \in \mathcal{V}^*} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + 2C_0 \left| \sum_{j \in \mathcal{V}^*} \frac{(\hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot u_j}{\sigma_{Y_j}^2} \right| + 2 \left| \sum_{j \in \mathcal{V}^*} \frac{(\hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot \nu_j}{\sigma_{Y_j}^2} \right| \right. \\ & \quad \left. + 2C_0 \left| \sum_{j \in \mathcal{V}^*} \frac{\nu_j u_j}{\sigma_{Y_j}^2} \right| + \left| \sum_{j \in \mathcal{V}_1} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} \right| + \theta_0^2 \left| \sum_{j \in \mathcal{V}_1} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + \theta_0^2 \left| \sum_{j \in \mathcal{V}_1} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + 2\theta_0 \left| \sum_{j \in \mathcal{V}_1} \frac{\nu_j u_j}{\sigma_{Y_j}^2} \right| \right). \end{aligned}$$

For simplicity, we can assume $C_0 = 1$ and $\theta_0 = 1$. We also know that

$$\sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} - \kappa_n \cdot (v^* - v_1) \geq \sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2} - \frac{\left(\sum_{j \in \mathcal{V}^*} \frac{r_j \cdot \beta_{X_j}}{\sigma_{Y_j}^2} \right)^2}{\sum_{j \in \mathcal{V}^*} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}} - \kappa_n \cdot (v^* - v_1).$$

If we have

$$\frac{\kappa_n \cdot (v^* - v_1)}{\sum_{j \in \mathcal{V}^*} \frac{r_j^2}{\sigma_{Y_j}^2} - \frac{(\sum_{j \in \mathcal{V}^*} \frac{r_j \cdot \beta_{X_j}}{\sigma_{Y_j}^2})^2}{\sum_{j \in \mathcal{V}^*} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}}} = \frac{\kappa_n \cdot (v^* - v_1)}{v^* \Delta(\mathcal{V}^*)} \rightarrow 0. \quad (\text{S7})$$

uniformly holds, there should be a c such that

$$\sum_{j \in \mathcal{V}} \frac{r_j^2}{\sigma_{Y_j}^2} - \frac{(\sum_{j \in \mathcal{V}} \frac{r_j \cdot \beta_{X_j}}{\sigma_{Y_j}^2})^2}{\sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}} - \kappa_n \cdot (v^* - v_1) \geq c \cdot (\sum_{j \in \mathcal{V}} \frac{r_j^2}{\sigma_{Y_j}^2} - \frac{(\sum_{j \in \mathcal{V}} \frac{r_j \cdot \beta_{X_j}}{\sigma_{Y_j}^2})^2}{\sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}}).$$

We prove the Equation (S7) in Lemma 6.

With this result, the above probability is bounded by

$$\begin{aligned} & \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}^*} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq \frac{c}{10} \left(\sum_{j \in \mathcal{V}} \frac{r_j^2}{\sigma_{Y_j}^2} - \frac{(\sum_{j \in \mathcal{V}} \frac{r_j \cdot \beta_{X_j}}{\sigma_{Y_j}^2})^2}{\sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}}\right)\right) + \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}_1} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq \frac{c}{10} \left(\sum_{j \in \mathcal{V}} \frac{r_j^2}{\sigma_{Y_j}^2} - \frac{(\sum_{j \in \mathcal{V}} \frac{r_j \cdot \beta_{X_j}}{\sigma_{Y_j}^2})^2}{\sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}}\right)\right) \\ & + \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}^*} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq \frac{c}{10} \left(\sum_{j \in \mathcal{V}} \frac{r_j^2}{\sigma_{Y_j}^2} - \frac{(\sum_{j \in \mathcal{V}} \frac{r_j \cdot \beta_{X_j}}{\sigma_{Y_j}^2})^2}{\sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}}\right)\right) + \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}_1} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq \frac{c}{10} \left(\sum_{j \in \mathcal{V}} \frac{r_j^2}{\sigma_{Y_j}^2} - \frac{(\sum_{j \in \mathcal{V}} \frac{r_j \cdot \beta_{X_j}}{\sigma_{Y_j}^2})^2}{\sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}}\right)\right) \\ & + \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}^*} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2}\right| \geq \frac{c}{10} \left(\sum_{j \in \mathcal{V}} \frac{r_j^2}{\sigma_{Y_j}^2} - \frac{(\sum_{j \in \mathcal{V}} \frac{r_j \cdot \beta_{X_j}}{\sigma_{Y_j}^2})^2}{\sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}}\right)\right) + \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2}\right| \geq \frac{c}{10} \left(\sum_{j \in \mathcal{V}} \frac{r_j^2}{\sigma_{Y_j}^2} - \frac{(\sum_{j \in \mathcal{V}} \frac{r_j \cdot \beta_{X_j}}{\sigma_{Y_j}^2})^2}{\sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}}\right)\right) \\ & + \mathbb{P}\left(2\left|\sum_{j \in \mathcal{V}_1} \frac{\nu_j u_j}{\sigma_{Y_j}^2}\right| \geq \frac{c}{10} \left(\sum_{j \in \mathcal{V}} \frac{r_j^2}{\sigma_{Y_j}^2} - \frac{(\sum_{j \in \mathcal{V}} \frac{r_j \cdot \beta_{X_j}}{\sigma_{Y_j}^2})^2}{\sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}}\right)\right) + \mathbb{P}\left(2\left|\sum_{j \in \mathcal{V}^*} \frac{\nu_j u_j}{\sigma_{Y_j}^2}\right| \geq \frac{c}{10} \left(\sum_{j \in \mathcal{V}} \frac{r_j^2}{\sigma_{Y_j}^2} - \frac{(\sum_{j \in \mathcal{V}} \frac{r_j \cdot \beta_{X_j}}{\sigma_{Y_j}^2})^2}{\sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}}\right)\right) \\ & + \mathbb{P}\left(2\left|\sum_{j \in \mathcal{V}^*} \frac{(\hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot \nu_j}{\sigma_{Y_j}^2}\right| \geq \frac{1}{10} \left(\sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} - \kappa_n \cdot (v^* - v_1)\right)\right) \\ & + \mathbb{P}\left(2\left|\sum_{j \in \mathcal{V}^*} \frac{(\hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot u_j}{\sigma_{Y_j}^2}\right| \geq \frac{1}{10} \left(\sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} - \kappa_n \cdot (v^* - v_1)\right)\right). \end{aligned}$$

Also we have

$$\left|\sum_{j \in \mathcal{V}^*} \frac{(\hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot u_j}{\sigma_{Y_j}^2}\right| \leq \sqrt{\sum_{j \in \mathcal{V}^*} \frac{(\hat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j)^2}{\sigma_{Y_j}^2}} \sqrt{\sum_{j \in \mathcal{V}^*} \frac{u_j^2}{\sigma_{Y_j}^2}}$$

$$\left| \sum_{j \in \mathcal{V}^*} \frac{(\widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot \nu_j}{\sigma_{Y_j}^2} \right| \leq \sqrt{\sum_{j \in \mathcal{V}^*} \frac{(\widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j)^2}{\sigma_{Y_j}^2}} \sqrt{\sum_{j \in \mathcal{V}^*} \frac{\nu_j^2}{\sigma_{Y_j}^2}}$$

That means there exists a $c'' > 0$ such that the last two terms are further bounded by

$$\begin{aligned} & \mathbb{P}\left(\left| \sum_{j \in \mathcal{V}^*} \frac{(\widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot \nu_j}{\sigma_{Y_j}^2} \right| \geq \frac{1}{10} \left(\sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} - \kappa_n \cdot (v^* - v_1) \right)\right) \\ & \leq \mathbb{P}\left(\sqrt{\sum_{j \in \mathcal{V}^*} \frac{\nu_j^2}{\sigma_{Y_j}^2}} \geq \frac{\frac{1}{10} \left(\sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} - \kappa_n \cdot (v^* - v_1) \right)}{\sqrt{\sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2}}}\right) \\ & \leq \mathbb{P}\left(\sum_{j \in \mathcal{V}^*} \frac{\nu_j^2}{\sigma_{Y_j}^2} \geq c'' \left(\sum_{j \in \mathcal{V}} \frac{r_j^2}{\sigma_{Y_j}^2} - \frac{\left(\sum_{j \in \mathcal{V}} \frac{r_j \cdot \beta_{X_j}}{\sigma_{Y_j}^2} \right)^2}{\sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}} \right)\right). \end{aligned}$$

Similarly,

$$\begin{aligned} & \mathbb{P}\left(\left| \sum_{j \in \mathcal{V}^*} \frac{(\widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j} - \theta_0 \cdot \beta_{X_j} - r_j) \cdot u_j}{\sigma_{Y_j}^2} \right| \geq \frac{1}{10} \left(\sum_{j \in \mathcal{V}^*} \frac{(\theta_0 \cdot \beta_{X_j} + r_j - \widehat{\theta}(\mathcal{V}^*) \cdot \beta_{X_j})^2}{\sigma_{Y_j}^2} - \kappa_n \cdot (v^* - v_1) \right)\right) \\ & \leq \mathbb{P}\left(\sum_{j \in \mathcal{V}^*} \frac{u_j^2}{\sigma_{Y_j}^2} \geq c'' \left(\sum_{j \in \mathcal{V}} \frac{r_j^2}{\sigma_{Y_j}^2} - \frac{\left(\sum_{j \in \mathcal{V}} \frac{r_j \cdot \beta_{X_j}}{\sigma_{Y_j}^2} \right)^2}{\sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}} \right)\right). \end{aligned}$$

Using Lemma 1, we can show that these ten probabilities are bounded by $2 \cdot e^{-c' \cdot \min \left\{ \frac{v^{*2} \cdot \Delta^2(\mathcal{V}^*)}{v_1}, v^* \cdot \Delta^2(\mathcal{V}^*), v^* \cdot \Delta(\mathcal{V}^*) \right\}}$.

To prove $e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot \mathbb{P}\left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_1, \theta_0) - \kappa_n \cdot v_1\right) \rightarrow 0$ uniformly for any $\mathcal{V}^* \in \mathcal{V}_{\text{bias}}(\varepsilon)$ such that $|\mathcal{V}^*| = v^* \geq \frac{1+c_1}{2} \cdot v_1$ and $\mathcal{V}^* \neq \mathcal{V}_1$, we only need to show

$$2e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot e^{-c' \cdot \min \left\{ v^* \cdot \Delta(\mathcal{V}^*), v^* \cdot \Delta^2(\mathcal{V}^*), \frac{v^{*2} \cdot \Delta(\mathcal{V}^*)}{v_1} \right\}} \rightarrow 0.$$

The above formula can be converted into

$$2e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot e^{-c' \cdot \min \{v^* \cdot \Delta(\mathcal{V}^*)\}} \rightarrow 0. \quad (\text{S8})$$

We prove the Equation (S8) in Lemma 6.

When $v^* < \frac{1+c_1}{2} \cdot v_1$, we decompose $h(\mathcal{V}^*, \widehat{\theta}(\mathcal{V}^*))$ in a different way,

$$\begin{aligned} h(\mathcal{V}^*, \widehat{\theta}(\mathcal{V}^*)) &= \sum_{j \in \mathcal{V}^*} \frac{(\widehat{\beta}_{Y_j} - \widehat{\theta}(\mathcal{V}^*) \cdot \widehat{\beta}_{X_{j,\text{RB}}})^2}{\sigma_{Y_j}^2} - \widehat{\theta}(\mathcal{V}^*)^2 \sum_{j \in \mathcal{V}^*} \frac{\widehat{\sigma}_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \\ &= \sum_{j \in \mathcal{V}^*} \frac{(\widehat{\beta}_{Y_j} - \widehat{\theta}(\mathcal{V}^*) \cdot \widehat{\beta}_{X_{j,\text{RB}}})^2}{\sigma_{Y_j}^2} - \widehat{\theta}(\mathcal{V}^*)^2 \sum_{j \in \mathcal{V}^*} \frac{\widehat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} - \widehat{\theta}(\mathcal{V}^*)^2 \sum_{j \in \mathcal{V}^*} \frac{\sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \end{aligned}$$

Under Condition 4, there exists a $C_0 > 0$ such that

$$\begin{aligned} &\mathbb{P}\left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_1, \theta_0) - \kappa_n \cdot v_1\right) \\ &\leq \mathbb{P}\left(\sum_{j \in \mathcal{V}^*} \frac{(\widehat{\beta}_{Y_j} - \widehat{\theta}(\mathcal{V}^*) \cdot \widehat{\beta}_{X_{j,\text{RB}}})^2}{\sigma_{Y_j}^2} + \kappa_n \cdot (v_1 - v^*) \leq \widehat{\theta}(\mathcal{V}^*)^2 \sum_{j \in \mathcal{V}^*} \frac{\widehat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} + \widehat{\theta}(\mathcal{V}^*)^2 \sum_{j \in \mathcal{V}^*} \frac{\sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right. \\ &\quad \left. + \sum_{j \in \mathcal{V}_1} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} + \theta_0^2 \sum_{j \in \mathcal{V}_1} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} - \theta_0^2 \sum_{j \in \mathcal{V}_1} \frac{\widehat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} - 2\theta_0 \sum_{j \in \mathcal{V}_1} \frac{\nu_j u_j}{\sigma_{Y_j}^2}\right) \\ &\leq \mathbb{P}\left(\sum_{j \in \mathcal{V}^*} \frac{(\widehat{\beta}_{Y_j} - \widehat{\theta}(\mathcal{V}^*) \cdot \widehat{\beta}_{X_{j,\text{RB}}})^2}{\sigma_{Y_j}^2} + \kappa_n \cdot (v_1 - v^*) \leq C_0^2 \left| \sum_{j \in \mathcal{V}^*} \frac{\widehat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + C_0^2 \cdot v^*\right. \\ &\quad \left. + \left| \sum_{j \in \mathcal{V}_1} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} \right| + \theta_0^2 \left| \sum_{j \in \mathcal{V}_1} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + \theta_0^2 \left| \sum_{j \in \mathcal{V}_1} \frac{\widehat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + 2\theta_0 \left| \sum_{j \in \mathcal{V}_1} \frac{\nu_j u_j}{\sigma_{Y_j}^2} \right| \right) \\ &\leq \mathbb{P}\left(\kappa_n \cdot (v_1 - v^*) - C_0^2 \cdot v^* \leq C_0^2 \left| \sum_{j \in \mathcal{V}^*} \frac{\widehat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + \left| \sum_{j \in \mathcal{V}_1} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} \right|\right. \\ &\quad \left. + \theta_0^2 \left| \sum_{j \in \mathcal{V}_1} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + \theta_0^2 \left| \sum_{j \in \mathcal{V}_1} \frac{\widehat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| + 2\theta_0 \left| \sum_{j \in \mathcal{V}_1} \frac{\nu_j u_j}{\sigma_{Y_j}^2} \right| \right). \end{aligned}$$

For simplicity, we can assume $C_0 = 1$ and $\theta_0 = 1$.

Using $v^* < \frac{1+c_1}{2} \cdot v_1$, we know that there must be a $c > 0$ such that

$$\kappa_n \cdot (v_1 - v^*) - C_0^2 \cdot v^* \geq c \cdot \kappa_n \cdot v_1.$$

then the above probability is bounded by

$$\begin{aligned} &\mathbb{P}\left(\left| \sum_{j \in \mathcal{V}_1} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| \geq \frac{c}{5} \cdot \kappa_n \cdot v_1\right) + \mathbb{P}\left(\left| \sum_{j \in \mathcal{V}^*} \frac{\widehat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| \geq \frac{c}{5} \cdot \kappa_n \cdot v_1\right) \\ &+ \mathbb{P}\left(\left| \sum_{j \in \mathcal{V}_1} \frac{\widehat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| \geq \frac{c}{5} \cdot \kappa_n \cdot v_1\right) + \mathbb{P}\left(2 \left| \sum_{j \in \mathcal{V}_1} \frac{\nu_j u_j}{\sigma_{Y_j}^2} \right| \geq \frac{c}{5} \cdot \kappa_n \cdot v_1\right) + \mathbb{P}\left(\left| \sum_{j \in \mathcal{V}_\lambda} \frac{\nu_j^2 - \sigma_{Y_j}^2}{\sigma_{Y_j}^2} \right| \geq \frac{c}{5} \cdot \kappa_n \cdot v_1\right). \end{aligned}$$

Using Lemma 1, we know that there exists a $c' > 0$ such that the these five terms are bounded by $2 \cdot e^{-c' \cdot \min \left\{ \kappa_n^2 v_1, \frac{\kappa_n^2 v_1^2}{v^*}, \kappa_n \cdot v_1 \right\}}$.

To prove $e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot \mathbb{P} \left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_1, \theta_0) - \kappa_n \cdot v_1 \right) \rightarrow 0$, we only need to show

$$2e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot e^{-c' \cdot \kappa_n \cdot v_1} \rightarrow 0.$$

This can be easily verified by Condition 8 and 9. So we can conclude that $e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot \mathbb{P} \left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_1, \theta_0) - \kappa_n \cdot v_1 \right) \rightarrow 0$ uniformly for $\mathcal{V}^* \in \mathcal{V}_{\text{bias}}(\varepsilon)$ such that $v^* < \frac{1+c_1}{2} \cdot v_1$.

Therefore, we conclude that $e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot \mathbb{P} \left(\min_{\theta \in \mathbb{R}} h(\mathcal{V}^*, \theta) - \kappa_n \cdot v^* \leq h(\mathcal{V}_1, \theta_0) - \kappa_n \cdot v_1 \right) \rightarrow 0$ uniformly for all $\mathcal{V}^* \in \mathcal{V}_{\text{bias}}(\varepsilon)$.

S.6.4 Proof of Lemma 5

To prove this lemma, we first define the event:

$$\begin{aligned} \mathcal{A}(\mathcal{V}, \varepsilon) = & \left\{ \left| \sum_{j \in \mathcal{V}} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| < \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2} \right\} \cap \left\{ \left| \sum_{j \in \mathcal{V}} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} \right| \geq \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2} \right\} \\ & \cap \left\{ \left| \sum_{j \in \mathcal{V}} \frac{\beta_{X_j} u_j}{\sigma_{Y_j}^2} \right| < \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2} \right\} \cap \left\{ \frac{4a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \left| \sum_{j \in \mathcal{V}} \frac{r_j u_j}{\sigma_{Y_j}^2} \right| < \varepsilon \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2} \right\}. \end{aligned}$$

We want to show for any $\varepsilon > 0$ and any given $\mathcal{V} \notin \mathcal{V}_{\text{bias}}(\varepsilon)$, under event $\mathcal{A}(\mathcal{V}, \varepsilon)$

$$\frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \cdot \frac{\sum_{j \in \mathcal{V}} r_j \cdot \hat{\beta}_{X_{j,\text{RB}}}}{\sum_{j \in \mathcal{V}} \hat{\beta}_{X_{j,\text{RB}}}^2 - \hat{\sigma}_{X_{j,\text{RB}}}^2} < 9 \cdot \varepsilon.$$

and $\mathbb{P}(\mathcal{A}(\mathcal{V}, \varepsilon)) \rightarrow 1$.

To do this, we make the following decomposition,

$$\frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \frac{\sum_{j \in \mathcal{V}} r_j \cdot \hat{\beta}_{X_{j,\text{RB}}}}{\sum_{j \in \mathcal{V}} \hat{\beta}_{X_{j,\text{RB}}}^2 - \hat{\sigma}_{X_{j,\text{RB}}}^2} = \frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \frac{\sum_{j \in \mathcal{V}} r_j \cdot \beta_{X_j} + \sum_{j \in \mathcal{V}} r_j \cdot u_j}{\sum_{j \in \mathcal{V}} \beta_{X_j}^2 + \sum_{j \in \mathcal{V}} 2\beta_{X_j} u_j + \sum_{j \in \mathcal{V}} (\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2) + \sum_{j \in \mathcal{V}} (u_j^2 - \sigma_{X_{j,\text{RB}}}^2)}$$

and notice that under $\mathcal{A}(\mathcal{V}, \varepsilon)$ we have

$$\begin{aligned}
\frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \frac{\sum_{j \in \mathcal{V}} r_j \cdot \hat{\beta}_{X_{j,\text{RB}}}}{\sum_{j \in \mathcal{V}} \hat{\beta}_{X_{j,\text{RB}}}^2 - \hat{\sigma}_{X_{j,\text{RB}}}^2} &\leq \frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \frac{\sum_{j \in \mathcal{V}} r_j \cdot \beta_{X_j} + \sum_{j \in \mathcal{V}} r_j \cdot u_j}{\frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \beta_{X_j}^2} \\
&= \frac{4 \cdot a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \frac{\sum_{j \in \mathcal{V}} r_j \cdot \beta_{X_j}}{\sum_{j \in \mathcal{V}} \beta_{X_j}^2} + \frac{4 \cdot a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \frac{\sum_{j \in \mathcal{V}} r_j \cdot u_j}{\sum_{j \in \mathcal{V}} \beta_{X_j}^2} \\
&< \frac{4 \cdot a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \frac{\sum_{j \in \mathcal{V}} r_j \cdot \beta_{X_j}}{\sum_{j \in \mathcal{V}} \beta_{X_j}^2} + \varepsilon.
\end{aligned}$$

For the first term on the right-hand side

$$\frac{4 \cdot a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \cdot \frac{\sum_{j \in \mathcal{V}} r_j \cdot \beta_{X_j}}{\sum_{j \in \mathcal{V}} \beta_{X_j}^2}$$

, we can rewrite it as

$$\begin{aligned}
4 \cdot \sqrt{ns_\lambda \cdot \beta_0^2} \cdot \frac{\sum_{j \in \mathcal{V}} r_j \cdot \beta_{X_j}}{\sum_{j \in \mathcal{V}} \beta_{X_j}^2} &= 4 \cdot \sqrt{ns_\lambda \cdot \beta_0^2} \cdot \frac{(v'_2 k + v'_3) \cdot r_3 \beta_0}{v'_1 \beta_0^2 + v'_2 \beta_0^2 + v'_3 \beta_0^2} \\
&= 4 \cdot \sqrt{ns_\lambda \cdot \beta_0^2} \cdot \frac{v'_2 k + v'_3}{(v'_1 + v'_2 + v'_3)} \cdot \frac{r_3}{\beta_0} \\
&= 4 \cdot \sqrt{ns_\lambda \cdot \beta_0^2} \cdot (p_2(\mathcal{V}) \cdot k + p_3(\mathcal{V})) \frac{r_3}{\beta_0} \\
&= 4 \cdot \sqrt{ns_\lambda} \cdot (p_2(\mathcal{V}) \cdot k + p_3(\mathcal{V})) \cdot r_3.
\end{aligned}$$

where $k = \frac{r_2}{r_3}$. For any given $\varepsilon > 0$, we have $p_3(\mathcal{V}) < \frac{\varepsilon}{\sqrt{ns_\lambda} \cdot r_3}$ for all $\mathcal{V} \notin \mathcal{V}_{\text{bias}}(\varepsilon)$. If we have $r_2 < \frac{\varepsilon}{\sqrt{ns_\lambda}}$, then we know

$$4 \cdot \sqrt{ns_\lambda \cdot \beta_0^2} \cdot \frac{\sum_{j \in \mathcal{V}} r_j \cdot \beta_{X_j}}{\sum_{j \in \mathcal{V}} \beta_{X_j}^2} = 4 \cdot \sqrt{ns_\lambda} \cdot (p_2(\mathcal{V}) \cdot k + p_3(\mathcal{V})) \cdot r_3 < 8\varepsilon.$$

holds for any given $\mathcal{V} \notin \mathcal{V}_{\text{bias}}(\varepsilon)$.

Now we show that under $\mathcal{A}(\mathcal{V}, \varepsilon)$,

$$\frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \cdot \frac{\sum_{j \in \mathcal{V}} r_j \cdot \hat{\beta}_{X_{j,\text{RB}}}}{\sum_{j \in \mathcal{V}} \hat{\beta}_{X_{j,\text{RB}}}^2 - \hat{\sigma}_{X_{j,\text{RB}}}^2} < 9 \cdot \varepsilon.$$

It suffices to show that $\mathbb{P}(\mathcal{A}(\mathcal{V}, \varepsilon)) \rightarrow 1$. We have

$$\begin{aligned} \mathbb{P}(\mathcal{A}(\mathcal{V}, \varepsilon)) &\geq 1 - \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}} \frac{\hat{\sigma}_{X_{j, \text{RB}}}^2 - \sigma_{X_{j, \text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}\right) - \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}} \frac{u_j^2 - \sigma_{X_{j, \text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}\right) \\ &\quad - \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}} \frac{\beta_{X_j} u_j}{\sigma_{Y_j}^2}\right| \geq \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}\right) - \mathbb{P}\left(\frac{4a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \left|\sum_{j \in \mathcal{V}} \frac{r_j u_j}{\sigma_{Y_j}^2}\right| \geq \varepsilon \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}\right) \end{aligned}$$

Under Condition 1, we have

$$\begin{aligned} \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}} \frac{\hat{\sigma}_{X_{j, \text{RB}}}^2 - \sigma_{X_{j, \text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}\right) &\leq 2 \cdot e^{-c \cdot \min\left\{\frac{(\sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2})^2}{16 \cdot v}, \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}\right\}} = 2 \cdot e^{-c \cdot \min\left\{\frac{1}{16} \cdot n^2 \cdot v \beta_0^4, \frac{1}{4} \cdot n \cdot v \beta_0^2\right\}} \\ \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}} \frac{u_j^2 - \sigma_{X_{j, \text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}\right) &\leq 2 \cdot e^{-c \cdot \min\left\{\frac{(\sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2})^2}{16 \cdot v}, \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}\right\}} = 2 \cdot e^{-c \cdot \min\left\{\frac{1}{16} \cdot n^2 \cdot v \beta_0^4, \frac{1}{4} \cdot n \cdot v \beta_0^2\right\}} \\ \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}} \frac{\beta_{X_j} u_j}{\sigma_{Y_j}^2}\right| \geq \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}\right) &\leq 2 \cdot e^{-c \cdot \frac{(\sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2})^2}{16 \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}}} = 2 \cdot e^{-\frac{c}{16} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}} = 2 \cdot e^{-\frac{c}{16} n \cdot \sum_{j \in \mathcal{V}} \beta_{X_j}^2} = 2 \cdot e^{-\frac{c}{16} n \cdot v \cdot \beta_0^2}. \\ \mathbb{P}\left(\frac{4a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \left|\sum_{j \in \mathcal{V}} \frac{r_j u_j}{\sigma_{Y_j}^2}\right| \geq \varepsilon \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}\right) &\leq 2 \cdot e^{-c \cdot \frac{\varepsilon^2 \cdot (\sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2})^2}{16 \cdot \sum_{j \in \mathcal{V}} \frac{r_j^2}{\sigma_{Y_j}^2}} \cdot \frac{s_\lambda \cdot \kappa_\lambda}{a_\lambda^2}} = 2 \cdot e^{\frac{-c \cdot \varepsilon^2 \cdot n^2 \cdot v^2 \beta_0^4}{16n \sum_{j \in \mathcal{V}} r_j^2} \cdot \frac{1}{ns_\lambda \beta_0^2}} = 2 \cdot e^{\frac{-c \cdot \varepsilon^2 \cdot \beta_0^2}{p_2(\mathcal{V})r_2^2 + p_3(\mathcal{V})r_3^2} \cdot \frac{v}{16s_\lambda}}. \end{aligned}$$

Thus we have $\mathbb{P}(\mathcal{A}(\mathcal{V}, \varepsilon)) \geq 1 - 2 \cdot e^{-\frac{c}{16} n \cdot v \cdot \beta_0^2} - 4 \cdot e^{-c \cdot \min\left\{\frac{1}{16} \cdot n^2 \cdot v \beta_0^4, \frac{1}{4} \cdot n \cdot v \beta_0^2\right\}} - 2 \cdot e^{\frac{-c \cdot \varepsilon^2 \cdot \beta_0^2}{p_2(\mathcal{V})r_2^2 + p_3(\mathcal{V})r_3^2} \cdot \frac{v}{16s_\lambda}}$.

Since $|\mathcal{V}| = v \geq \frac{1+c_1}{2} s_\lambda$ and $p_3(\mathcal{V}) < \frac{\varepsilon}{\sqrt{ns_\lambda} \cdot r_3}$ for $\mathcal{V} \notin \mathcal{V}_{\text{bias}}(\varepsilon)$, if we have $k < \frac{\varepsilon}{\sqrt{ns_\lambda} \cdot r_3}$ then

$$\begin{aligned} \frac{p_2(\mathcal{V})r_2^2 + p_3(\mathcal{V})r_3^2}{\varepsilon^2 \cdot \beta_0^2} &\leq \frac{2\varepsilon}{\sqrt{ns_\lambda} r_3} \frac{r_3^2}{\varepsilon^2 \beta_0^2} \\ &= \frac{2\varepsilon}{\sqrt{ns_\lambda}} \frac{r_3}{\varepsilon^2 \beta_0^2} \\ &= \frac{2}{\sqrt{ns_\lambda}} \frac{r_3}{\varepsilon \beta_0^2}. \end{aligned}$$

If $\frac{1}{\sqrt{ns_\lambda}} \frac{r_3}{\varepsilon \beta_0^2} \rightarrow 0$ and $\frac{1}{n^2 \cdot s_\lambda \beta_0^4} \rightarrow 0$, we then can show $\mathbb{P}(\mathcal{A}(\mathcal{V}, \varepsilon)) \rightarrow 1$.

Thus we have

$$\frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \cdot \frac{\sum_{j \in \mathcal{V}} r_j \cdot \widehat{\beta}_{X_{j,\text{RB}}}}{\sum_{j \in \mathcal{V}} \widehat{\beta}_{X_{j,\text{RB}}}^2 - \widehat{\sigma}_{X_{j,\text{RB}}}^2} = O_p(\varepsilon),$$

for any given $\mathcal{V} \notin \mathcal{V}_{\text{bias}}(\varepsilon)$.

S.6.5 Additional Lemmas

Lemma 6 *Under Condition 2 and 8, if $A_n/(r_3\varepsilon) = o(1)$, we have*

$$2e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \cdot e^{-c' \cdot \min\{v \cdot \Delta(\mathcal{V})\}} \rightarrow 0 \text{ and } \frac{\kappa_n \cdot (v - v_1)}{v \cdot \Delta(\mathcal{V})} \rightarrow 0.$$

uniformly hold for all $\mathcal{V} \in \mathcal{V}_{\text{bias}}(\varepsilon)$ and $v = |\mathcal{V}| \geq \frac{1+c_1}{2}v_1$.

S.6.5.1 Proof of Lemma 6

For a given set $\mathcal{V} \in \mathcal{V}_{\text{bias}}(\varepsilon)$, we let $v'_1 = |\mathcal{V}_1 \cap \mathcal{V}|$, $v'_2 = |\mathcal{V}_2 \cap \mathcal{V}|$ and $v'_3 = |\mathcal{V}_3 \cap \mathcal{V}|$. Then

$$\begin{aligned} \Delta(\mathcal{V}) &= \frac{1}{v'_1 + v'_2 + v'_3} \left(\sum_{j \in \mathcal{V}} \frac{r_j^2}{\sigma_{Y_j}^2} - \frac{\left(\sum_{j \in \mathcal{V}} \frac{r_j \cdot \beta_{X_j}}{\sigma_{Y_j}^2} \right)^2}{\sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}} \right) = \frac{v'_2 k^2 + v'_3}{v'_1 + v'_2 + v'_3} \cdot n r_3^2 - n \cdot \frac{(v'_2 k \cdot r_3 \beta_0 + v'_3 \cdot r_3 \beta_0)^2}{v'_1 \beta_0^2 + v'_2 \beta_0^2 + v'_3 \beta_0^2} \cdot \frac{1}{v'_1 + v'_2 + v'_3} \\ &= \frac{v'_2 k^2 + v'_3}{v'_1 + v'_2 + v'_3} \cdot n r_3^2 - \frac{(v'_2 k + v'_3)^2 \cdot n r_3^2}{(v'_1 + v'_2 + v'_3)^2}. \end{aligned}$$

Using the definition of $p_1(\mathcal{V})$, $p_2(\mathcal{V})$ and $p_3(\mathcal{V})$, we have

$$\Delta(\mathcal{V}) = \frac{1}{v'_1 + v'_2 + v'_3} \left(\sum_{j \in \mathcal{V}} \frac{r_j^2}{\sigma_{Y_j}^2} - \frac{\left(\sum_{j \in \mathcal{V}} \frac{r_j \cdot \beta_{X_j}}{\sigma_{Y_j}^2} \right)^2}{\sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}} \right) = (p_2(\mathcal{V})k^2 + p_3(\mathcal{V})) \cdot n r_3^2 - (p_2(\mathcal{V})k + p_3(\mathcal{V}))^2 \cdot n r_3^2.$$

where $k = \frac{r_2}{r_3}$.

We have

$$\begin{aligned}
& p_2(\mathcal{V})k^2 + p_3(\mathcal{V}) - (p_2(\mathcal{V})k + p_3(\mathcal{V}))^2 \\
&= (p_2(\mathcal{V}) - p_2(\mathcal{V})^2)k^2 - 2p_2(\mathcal{V})p_3(\mathcal{V}) \cdot k + p_3(\mathcal{V}) - p_3(\mathcal{V})^2 \\
&= p_2(\mathcal{V})(p_1(\mathcal{V}) + p_3(\mathcal{V}))k^2 - 2p_2(\mathcal{V})p_3(\mathcal{V}) \cdot k + p_3(\mathcal{V})(p_1(\mathcal{V}) + p_2(\mathcal{V})) \\
&= p_2(\mathcal{V})p_3(\mathcal{V}) \cdot k^2 - 2p_2(\mathcal{V})p_3(\mathcal{V}) \cdot k + p_2(\mathcal{V})p_3(\mathcal{V}) + p_1(\mathcal{V})p_3(\mathcal{V}) + p_2(\mathcal{V})p_1(\mathcal{V}) \cdot k^2 \\
&= p_2(\mathcal{V})p_3(\mathcal{V}) \cdot (k - 1)^2 + p_1(\mathcal{V})p_3(\mathcal{V}) + p_2(\mathcal{V})p_1(\mathcal{V}) \cdot k^2
\end{aligned}$$

Note that we have

$$p_3(\mathcal{V}) < \frac{v'_3}{v'_1 + v'_2 + v'_3} < \frac{c_1 \cdot v_1}{\frac{1+c_1}{2} \cdot v_1} = \frac{2c_1}{1+c_1}.$$

Let $c'_1 = \frac{2c_1}{1+c_1}$. If $p_3(\mathcal{V}) \geq \frac{\varepsilon}{\sqrt{ns_\lambda} \cdot r_3}$, we have $1 - c'_1 < p_1(\mathcal{V}) + p_2(\mathcal{V}) \leq 1 - \frac{\varepsilon}{\sqrt{ns_\lambda} \cdot r_3}$. Then we consider two different situations:

- $p_2(\mathcal{V}) \geq \frac{1-c'_1}{2}$: by choosing $k < \frac{1}{2}$, we have

$$(p_2(\mathcal{V})k^2 + p_3(\mathcal{V})) - (p_2(\mathcal{V})k + p_3(\mathcal{V}))^2 \geq p_2(\mathcal{V})p_3(\mathcal{V}) \cdot (k - 1)^2 > \frac{(1 - c'_1)}{8} \frac{\varepsilon}{\sqrt{ns_\lambda} \cdot r_3}.$$

- $p_2(\mathcal{V}) < \frac{1-c'_1}{2}$: we have $p_1(\mathcal{V}) > \frac{1-c'_1}{2}$,

$$(p_2(\mathcal{V})k^2 + p_3(\mathcal{V})) - (p_2(\mathcal{V})k + p_3(\mathcal{V}))^2 \geq p_1(\mathcal{V})p_3(\mathcal{V}) > \frac{1 - c'_1}{2} \cdot \frac{\varepsilon}{\sqrt{ns_\lambda} \cdot r_3}.$$

So we have when $p_3(\mathcal{V}) \geq \frac{\varepsilon}{\sqrt{ns_\lambda} \cdot r_3}$,

$$(p_2(\mathcal{V})k^2 + p_3(\mathcal{V})) - (p_2(\mathcal{V})k + p_3(\mathcal{V}))^2 > \frac{(1 - c'_1)}{8} \cdot \frac{\varepsilon}{\sqrt{ns_\lambda} \cdot r_3}.$$

Now we consider

$$\frac{(s_\lambda + 1) \cdot \ln(s_\lambda)}{v \cdot \Delta(\mathcal{V})} = \frac{s_\lambda + 1}{v} \frac{\ln(s_\lambda)}{((p_2(\mathcal{V})k^2 + p_3(\mathcal{V})) - (p_2(\mathcal{V})k + p_3(\mathcal{V}))^2) \cdot nr_3^2} \leq \frac{s_\lambda + 1}{\frac{c_1+1}{2} \cdot v_1} \frac{8\sqrt{s_\lambda} \max\{\ln(s_\lambda), \kappa_n\}}{(1 - c'_1) \cdot \varepsilon \cdot \sqrt{nr_3}}.$$

$$\frac{\kappa_n \cdot (v - v_1)}{v \cdot \Delta(\mathcal{V})} \leq \frac{\kappa_n}{((p_2(\mathcal{V})k^2 + p_3(\mathcal{V})) - (p_2(\mathcal{V})k + p_3(\mathcal{V}))^2) \cdot nr_3^2} \leq \frac{8\sqrt{s_\lambda} \max\{ln(s_\lambda), \kappa_n\}}{(1 - c'_1) \cdot \varepsilon \cdot \sqrt{nr_3}}.$$

Using Condition 8 we have $\frac{s_\lambda + 1}{\frac{c_1 + 1}{2} \cdot v_1} = O(1)$. Then if we further have

$$\frac{A_n}{\varepsilon r_3} = o(1),$$

we then can show

$$2e^{(s_\lambda + 1) \cdot ln(s_\lambda)} \cdot e^{-c' \cdot \min\{v \cdot \Delta(\mathcal{V})\}} \rightarrow 0 \text{ and } \frac{\kappa_n \cdot (v - v_1)}{v \cdot \Delta(\mathcal{V})} \rightarrow 0.$$

S.6.6 Sufficient conditions for the Boundness condition

Condition 10 (Boundedness) *For any $\mathcal{V} \in S_\lambda$, $|\hat{\theta}(\mathcal{V})|$ is uniformly bounded away from ∞ with probability goes to 1.*

To see this, we can decompose $\hat{\theta}(\mathcal{V})$ as follows:

$$\begin{aligned} \hat{\theta}(\mathcal{V}) &= \frac{\sum_{j \in \mathcal{V}} \frac{(\theta_0 \beta_{X_j} + r_j) \beta_{X_j}}{\sigma_{Y_j}^2} + \sum_{j \in \mathcal{V}} \frac{\beta_{X_j} \nu_j}{\sigma_{Y_j}^2} + \sum_{j \in \mathcal{V}} \frac{(\theta_0 \beta_{X_j} + r_j) u_j}{\sigma_{Y_j}^2} + \sum_{j \in \mathcal{V}} \frac{u_j \nu_j}{\sigma_{Y_j}^2}}{\sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2} + 2 \sum_{j \in \mathcal{V}} \frac{\beta_{X_j} u_j}{\sigma_{Y_j}^2} + \sum_{j \in \mathcal{V}} \frac{u_j^2 - \sigma_{X_{j,RB}}^2}{\sigma_{Y_j}^2} - \sum_{j \in \mathcal{V}} \frac{\hat{\sigma}_{X_{j,RB}}^2 - \sigma_{X_{j,RB}}^2}{\sigma_{Y_j}^2}} \\ &= \frac{(\theta_0 + \max_{j \in \mathcal{S}_\lambda} |\frac{r_j}{\beta_{X_j}}|) \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}}{\sigma_{Y_j}^2} + \sum_{j \in \mathcal{V}} \frac{\beta_{X_j} \nu_j}{\sigma_{Y_j}^2} + (\theta_0 + \max_{j \in \mathcal{S}_\lambda} |\frac{r_j}{\beta_{X_j}}|) \sum_{j \in \mathcal{V}} \frac{\beta_{X_j} u_j}{\sigma_{Y_j}^2} + \sum_{j \in \mathcal{V}} \frac{u_j \nu_j}{\sigma_{Y_j}^2}}{\sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2} + 2 \sum_{j \in \mathcal{V}} \frac{\beta_{X_j} u_j}{\sigma_{Y_j}^2} + \sum_{j \in \mathcal{V}} \frac{u_j^2 - \sigma_{X_{j,RB}}^2}{\sigma_{Y_j}^2} - \sum_{j \in \mathcal{V}} \frac{\hat{\sigma}_{X_{j,RB}}^2 - \sigma_{X_{j,RB}}^2}{\sigma_{Y_j}^2}} \end{aligned}$$

Define the event $\mathcal{B}(\mathcal{V})$

$$\begin{aligned} \mathcal{B}(\mathcal{V}) &= \left\{ \left| \sum_{j \in \mathcal{V}} \frac{\hat{\sigma}_{X_{j,RB}}^2 - \sigma_{X_{j,RB}}^2}{\sigma_{Y_j}^2} \right| < \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2} \right\} \cap \left\{ \left| \sum_{j \in \mathcal{V}} \frac{u_j^2 - \sigma_{X_{j,RB}}^2}{\sigma_{Y_j}^2} \right| < \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2} \right\} \\ &\cap \left\{ 2 \left| \sum_{j \in \mathcal{V}} \frac{\beta_{X_j} u_j}{\sigma_{Y_j}^2} \right| < \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2} \right\} \cap \left\{ \left| \sum_{j \in \mathcal{V}} \frac{u_j \nu_j}{\sigma_{Y_j}^2} \right| < \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2} \right\} \cap \left\{ \left| \sum_{j \in \mathcal{V}} \frac{\beta_{X_j} \nu_j}{\sigma_{Y_j}^2} \right| < \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2} \right\}. \end{aligned}$$

Under $\mathcal{B}(\mathcal{V})$, we have

$$\begin{aligned}
\hat{\theta}(\mathcal{V}) &= \frac{(\theta_0 + \max_{j \in \mathcal{S}_\lambda} |\frac{r_j}{\beta_{X_j}}|) \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2} + \sum_{j \in \mathcal{V}} \frac{\beta_{X_j} \nu_j}{\sigma_{Y_j}^2} + (\theta_0 + \max_{j \in \mathcal{S}_\lambda} |\frac{r_j}{\beta_{X_j}}|) \sum_{j \in \mathcal{V}} \frac{\beta_{X_j} u_j}{\sigma_{Y_j}^2} + \sum_{j \in \mathcal{V}} \frac{u_j \nu_j}{\sigma_{Y_j}^2}}{\sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2} + 2 \sum_{j \in \mathcal{V}} \frac{\beta_{X_j} u_j}{\sigma_{Y_j}^2} + \sum_{j \in \mathcal{V}} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2} - \sum_{j \in \mathcal{V}} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}} \\
&\leq \frac{(\theta_0 + \max_{j \in \mathcal{S}_\lambda} |\frac{r_j}{\beta_{X_j}}|) \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2} + \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2} + (\theta_0 + \max_{j \in \mathcal{S}_\lambda} |\frac{r_j}{\beta_{X_j}}|) \cdot \frac{1}{8} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2} + \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}}{\frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}} \\
&= \frac{9}{2} \cdot (\theta_0 + |\frac{r_3}{\beta_0}|) + 2.
\end{aligned}$$

Notice that

$$\begin{aligned}
\mathbb{P}(\mathcal{B}(\mathcal{V})) &\geq 1 - \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}} \frac{\hat{\sigma}_{X_{j,\text{RB}}}^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}\right) - \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}} \frac{u_j^2 - \sigma_{X_{j,\text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}\right) \\
&\quad - \mathbb{P}\left(2\left|\sum_{j \in \mathcal{V}} \frac{\beta_{X_j} u_j}{\sigma_{Y_j}^2}\right| \geq \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}\right) - \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}} \frac{u_j \nu_j}{\sigma_{Y_j}^2}\right| \geq \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}\right) - \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}} \frac{\beta_{X_j} \nu_j}{\sigma_{Y_j}^2}\right| \geq \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}\right).
\end{aligned}$$

Under Condition 1, we have

$$\begin{aligned}
& \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}} \frac{\hat{\sigma}_{X_{j, \text{RB}}}^2 - \sigma_{X_{j, \text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}\right) \leq 2 \cdot e^{-c \cdot \min\left\{\frac{(\sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2})^2}{16 \cdot v}, \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}\right\}} = 2 \cdot e^{-c \cdot \min\left\{\frac{1}{16} \cdot n^2 \cdot v \beta_0^4, \frac{1}{4} \cdot n \cdot v \beta_0^2\right\}} \\
& \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}} \frac{u_j^2 - \sigma_{X_{j, \text{RB}}}^2}{\sigma_{Y_j}^2}\right| \geq \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}\right) \leq 2 \cdot e^{-c \cdot \min\left\{\frac{(\sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2})^2}{16 \cdot v}, \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}\right\}} = 2 \cdot e^{-c \cdot \min\left\{\frac{1}{16} \cdot n^2 \cdot v \beta_0^4, \frac{1}{4} \cdot n \cdot v \beta_0^2\right\}} \\
& \mathbb{P}\left(2 \left|\sum_{j \in \mathcal{V}} \frac{\beta_{X_j} u_j}{\sigma_{Y_j}^2}\right| \geq \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}\right) \leq 2 \cdot e^{-64 \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}} = 2 \cdot e^{-\frac{c}{64} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}} = 2 \cdot e^{-\frac{c}{64} n \cdot \sum_{j \in \mathcal{V}} \beta_{X_j}^2} = 2 \cdot e^{-\frac{c}{64} n \cdot v \cdot \beta_0^2} \\
& \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}} \frac{\beta_{X_j} \nu_j}{\sigma_{Y_j}^2}\right| \geq \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}\right) \leq 2 \cdot e^{-\frac{c \cdot (\sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2})^2}{16 \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}}} = 2 \cdot e^{-\frac{c}{16} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}} = 2 \cdot e^{-\frac{c}{16} n \cdot \sum_{j \in \mathcal{V}} \beta_{X_j}^2} = 2 \cdot e^{-\frac{c}{16} n \cdot v \cdot \beta_0^2} \\
& \mathbb{P}\left(\left|\sum_{j \in \mathcal{V}} \frac{u_j \nu_j}{\sigma_{Y_j}^2}\right| \geq \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}\right) \leq 2 \cdot e^{-c \cdot \min\left\{\frac{(\sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2})^2}{16 \cdot v}, \frac{1}{4} \cdot \sum_{j \in \mathcal{V}} \frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}\right\}} = 2 \cdot e^{-c \cdot \min\left\{\frac{1}{16} \cdot n^2 \cdot v \beta_0^4, \frac{1}{4} \cdot n \cdot v \beta_0^2\right\}}.
\end{aligned}$$

Thus we have $\mathbb{P}(\mathcal{B}(\mathcal{V})) \geq 1 - 2 \cdot e^{-\frac{c}{16} n \cdot v \cdot \beta_0^2} - 2 \cdot e^{-\frac{c}{64} n \cdot v \cdot \beta_0^2} - 6 \cdot e^{-c \cdot \min\left\{\frac{1}{16} \cdot n^2 \cdot v \beta_0^4, \frac{1}{4} \cdot n \cdot v \beta_0^2\right\}}$.

Under the event $\bigcap_{\mathcal{V} \subseteq \mathcal{S}_\lambda} \mathcal{B}(\mathcal{V})$, we can show that

$$\hat{\theta}(\mathcal{V}) \leq \frac{9}{2} \cdot (\theta_0 + \left| \frac{r_3}{\beta_0} \right|) + 2.$$

holds uniformly for all subset $\mathcal{V} \subseteq \mathcal{S}_\lambda$. We also notice that

$$\mathbb{P}\left(\bigcap_{\mathcal{V} \subseteq \mathcal{S}_\lambda} \mathcal{B}(\mathcal{V})\right) = 1 - \mathbb{P}\left(\bigcup_{\mathcal{V} \subseteq \mathcal{S}_\lambda} \mathcal{B}^c(\mathcal{V})\right) \geq 1 - \sum_{\mathcal{V} \subseteq \mathcal{S}_\lambda} \mathbb{P}(\mathcal{B}^c(\mathcal{V})).$$

Here $\mathcal{B}^c(\mathcal{V})$ is the complement of the event $\mathcal{B}(\mathcal{V})$.

To prove $\mathbb{P}\left(\bigcap_{\mathcal{V} \subseteq \mathcal{S}_\lambda} \mathcal{B}(\mathcal{V})\right) \rightarrow 1$, we only need to show

$$\sum_{\mathcal{V} \subseteq \mathcal{S}_\lambda} \mathbb{P}(\mathcal{B}^c(\mathcal{V})) \leq e^{(s_\lambda + 1) \cdot \ln(s_\lambda)} \max_{\mathcal{V} \subseteq \mathcal{S}_\lambda} \mathbb{P}(\mathcal{B}^c(\mathcal{V})) \rightarrow 0.$$

We have shown $\mathbb{P}(\mathcal{B}(\mathcal{V})) \geq 1 - 2 \cdot e^{-\frac{c}{16}n \cdot v \cdot \beta_0^2} - 2 \cdot e^{-\frac{c}{64}n \cdot v \cdot \beta_0^2} - 6 \cdot e^{-c \cdot \min\{\frac{1}{16} \cdot n^2 \cdot v \beta_0^4, \frac{1}{4} \cdot n \cdot v \beta_0^2\}}$, thus $\mathbb{P}(\mathcal{B}^c(\mathcal{V})) < 2 \cdot e^{-\frac{c}{16}n \cdot v \cdot \beta_0^2} + 2 \cdot e^{-\frac{c}{64}n \cdot v \cdot \beta_0^2} + 6 \cdot e^{-c \cdot \min\{\frac{1}{16} \cdot n^2 \cdot v \beta_0^4, \frac{1}{4} \cdot n \cdot v \beta_0^2\}}$.

With these results, we can prove $e^{(s_\lambda+1) \cdot \ln(s_\lambda)} \max_{\mathcal{V} \subseteq \mathcal{S}_\lambda} \mathbb{P}(\mathcal{B}^c(\mathcal{V})) \rightarrow 0$ if we have $\frac{s_\lambda \ln(s_\lambda)}{n \beta_0^2} \rightarrow 0$.

Thus we have

$$\widehat{\theta}(\mathcal{V}) \leq \frac{9}{2} \cdot (\theta_0 + |\frac{r_3}{\beta_0}|) + 2.$$

holds uniformly for all subset $\mathcal{V} \subseteq \mathcal{S}_\lambda$ with probability approaching one.

If there exists a constant $C > 0$ such that $|\frac{r_3}{\beta_0}| < C$, we then can verify that

$$\widehat{\theta}(\mathcal{V}) \leq \frac{9}{2} \cdot (\theta_0 + C) + 2.$$

for all subset $\mathcal{V} \subseteq \mathcal{S}_\lambda$ with probability going to one.

S.7 Connections and differences with [4]

A summary of the proposed method in [4]. The authors consider a setup with an initial GWAS scan and a replication study. In the initial scan, they let the $\{X_1, \dots, X_K\}$ be the estimated effect size for K SNPs. They also assume these effect sizes follow normal distributions:

$$X_i \sim \mathcal{N}(\mu_i, \sigma_{1,i}^2).$$

This initial scan is used for selecting the strong SNPs. They ordered these effect size as $X_{(1)}, \dots, X_{(K)}$ and denote the corresponding means and variances as $\mu_{(1)}, \dots, \mu_{(K)}$ and $\sigma_{1,(1)}^2, \dots, \sigma_{1,(K)}^2$, and then perform the following selection:

$$\frac{|X_{(1)}|}{\sigma_{1,(1)}} \geq \frac{|X_{(2)}|}{\sigma_{1,(2)}} \geq \dots \frac{|X_{(1,k)}|}{\sigma_{(k)}} \geq \Phi^{-1}(1 - c_{crit}) = \lambda.$$

Due to the selection step, the distribution of $X_{(i)}$ becomes truncated normal, and therefore, $X_{(i)}$ is a biased estimator of $\mu_{(i)}$. To get an unbiased estimation, the authors leverage the replication

study, where Y_i is the effect size of the i -th ranked SNP such that

$$Y_i \sim \mathcal{N}(\mu_{(i)}, \sigma_{2,i}^2)$$

For simplicity, we let $\sigma_{1,(i)}^2 = \sigma_{2,i}^2 = \sigma_i^2$. Obviously, Y_i is an unbiased estimator of $\mu_{(i)}$. However, it often has a large variance. For this reason, the authors proposed a weighed version of estimator,

$$\hat{\mu}_{(i)} = \frac{\sigma_{1,(i)}^2 Y_i + \sigma_{2,i}^2 X_{(i)}}{\sigma_{1,(i)}^2 + \sigma_{2,i}^2} = \frac{Y_i + X_{(i)}}{2},$$

which can effectively combine the data from both the initial scan and the replication study.

Based on this estimator $\hat{\mu}_{(i)}$, they then construct an unbiased estimator of $\mu_{(i)}$ and further use the Rao-Blackwellization to obtain an unbiased estimator with the minimum variance.

This estimator takes the form:

$$\tilde{\mu}_{(i)} = \hat{\mu}_{(i)} - \frac{\sigma_i}{\sqrt{2}} \cdot \frac{\phi(W_{i,i+1}^{(0)}) - \phi(W_{i,i-1}^{(0)}) - \phi(W_{i,i+1}^{(1)}) + \phi(W_{i,i-1}^{(1)})}{\Phi(W_{i,i+1}^{(0)}) - \Phi(W_{i,i-1}^{(0)}) - \Phi(W_{i,i+1}^{(1)}) + \Phi(W_{i,i-1}^{(1)})},$$

where $W_{s,t}^{(p)} = \frac{\sqrt{2}}{\sigma_s} \cdot (\hat{\mu}_{(s)} - (-1)^p \frac{\sigma_s |X_{(t)}|}{\sigma_t})$, $\frac{|X_{(0)}|}{\sigma_{(0)}} = \infty$, and $\frac{|X_{(k+1)}|}{\sigma_{(k+1)}} = \Phi^{-1}(1 - c_{crit}) = \lambda$.

Connections and differences with our approach. When customizing the proposed approach in [4] to our problem for SNP selection, we may perform the selection following:

$$\frac{|X_{(1)}|}{\sigma_1} \geq \Phi^{-1}(1 - c_{crit}) = \lambda.$$

Then, the corresponding unbiased estimator for $\mu_{(1)}$ can then be given by

$$\begin{aligned} \tilde{\mu}_{(1)} &= \hat{\mu}_{(1)} - \frac{\sigma_1}{\sqrt{2}} \cdot \frac{\phi(W_{1,2}^{(0)}) - \phi(W_{1,0}^{(0)}) - \phi(W_{1,2}^{(1)}) + \phi(W_{1,0}^{(1)})}{\Phi(W_{1,2}^{(0)}) - \Phi(W_{1,0}^{(0)}) - \Phi(W_{1,2}^{(1)}) + \Phi(W_{1,0}^{(1)})} \\ &= \hat{\mu}_{(1)} - \frac{\sigma_1}{\sqrt{2}} \cdot \frac{\phi(W_{1,2}^{(0)}) - \phi(W_{1,2}^{(1)})}{\Phi(W_{1,2}^{(0)}) + 1 - \Phi(W_{1,2}^{(1)})} \\ &= \hat{\mu}_{(1)} - \frac{\sigma_1}{\sqrt{2}} \cdot \frac{\phi(\frac{\sqrt{2}}{\sigma_1} \cdot (\hat{\mu}_{(1)} - \lambda)) - \phi(\frac{\sqrt{2}}{\sigma_1} \cdot (\hat{\mu}_{(1)} + \lambda))}{\Phi(\frac{\sqrt{2}}{\sigma_1} \cdot (\hat{\mu}_{(1)} - \lambda)) + 1 - \Phi(\frac{\sqrt{2}}{\sigma_1} \cdot (\hat{\mu}_{(1)} + \lambda))}. \end{aligned}$$

Although this estimator appears very similar to the one proposed in our manuscript, it requires

summary statistics from a replication study (as $W_{s,t}^{(p)}$ depends on Y_i). In other words, to provide an unbiased estimator of $\mu_{(1)}$, [4] requires two GWAS: one initial GWAS for selection and another replication study for unbiased estimator construction. Therefore, the key difference between [4] and our approach is that our method can construct an unbiased estimator of $\mu_{(1)}$ without a replication study. In other words, the setting considered in [4] is aligned with the three-sample MR setting, where SNP selection is performed on a third independent exposure GWAS sample. In contrast, we focus on the two-sample MR, where SNP selection and parameter estimation are carried out in the same exposure GWAS sample. From a different perspective, our approach is indeed connected to [4] as a Rao-Blackwellization step is applied to improve the estimation efficiency.

S.8 Simulation settings and additional simulation results

S.8.1 Simulation settings

Note that the total effects of SNP j on exposure X and outcome Y can be written as:

$$\beta_{X_j} = \gamma_j + \beta_{XU}\phi_j; \quad \beta_{Y_j} = \theta\beta_{X_j} + \alpha_j + \beta_{YU}\phi_j,$$

where γ_j , ϕ_j , and α_j is the true direct effect of SNP j on X , confounding factor U , and Y , respectively (see Figure 1). Following [39], we simulate summary-level association statistics $\widehat{\beta}_{X_j}$ and $\widehat{\beta}_{Y_j}$ directly. Specifically, we generate

$$\begin{bmatrix} \widehat{\beta}_{X_j} \\ \widehat{\beta}_{Y_j} \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \beta_{X_j} \\ \beta_{Y_j} \end{bmatrix}, \begin{bmatrix} \sigma_{X_j} & 0 \\ 0 & \sigma_{Y_j} \end{bmatrix} \right),$$

where $\sigma_{X_j} = \sqrt{1/n_X}$ and $\sigma_{Y_j} = \sqrt{1/n_Y}$.

To save space and make the simulations representative of real GWAS data, we focus on general simulation settings where both directional correlated pleiotropy and balanced uncorrelated pleiotropy are considered simultaneously. Other specific simulation settings have also been briefly considered. Specifically, we generate the underlying parameters from a mixture of distributions, a

setup that has been widely used for modeling the effect sizes of complex traits in GWAS [7, 47, 55]:

$$\begin{pmatrix} \gamma_j \\ \alpha_j \\ \phi_j \end{pmatrix} \sim \underbrace{\pi_1 \begin{pmatrix} \mathcal{N}(0, \sigma_x^2) \\ \delta_0 \\ \delta_0 \end{pmatrix}}_{\text{Valid IVs}} + \underbrace{\pi_2 \begin{pmatrix} \mathcal{N}(0, \sigma_x^2) \\ \mathcal{N}(0.015, \sigma_u^2) \\ \mathcal{N}(0, \sigma_u^2) \end{pmatrix}}_{\text{correlated pleiotropy}} + \underbrace{\pi_3 \begin{pmatrix} \mathcal{N}(0, \sigma_x^2) \\ \mathcal{N}(0, \sigma_y^2) \\ \delta_0 \end{pmatrix}}_{\text{uncorrelated pleiotropy}} + \underbrace{\pi_4 \begin{pmatrix} \delta_0 \\ \mathcal{N}(0, \sigma_y^2) \\ \delta_0 \end{pmatrix}}_{\text{IVs fail the relevance assumption}} + \underbrace{\pi_5 \begin{pmatrix} \delta_0 \\ \delta_0 \\ \delta_0 \end{pmatrix}}_{\text{IVs fail the relevance assumption}}, \quad (\text{S9})$$

where δ_0 is a Dirac measure centered on zero (i.e., the point mass at 0), π_1 controls the proportion of valid IVs (where $\gamma_j \neq 0$ and both α_j and ϕ_j are equal to zero), π_2 controls the proportion of invalid IVs due to correlated pleiotropy, π_3 controls the proportion of invalid IVs due to uncorrelated pleiotropy, π_4 controls the proportion of SNPs that are only associated with Y , and $\pi_5 = 1 - \sum_{j=1}^3 \pi_j$ controls the proportion of SNPs that have no association with both X and Y . Note that when $\phi_j \neq 0$, the Instrument Strength Independent on Direct Effect (InSIDE) assumption is violated for SNP j because the exposure effect is correlated with their pleiotropic effects on the come due to mediation by common confounding factor U . InSIDE assumption is popular in MR literature and requires that the exposure effects of individual SNPs are independent of their pleiotropic effects on the outcome [10].

Following [39], we generate 200,000 independent SNPs to represent all underlying common variants and set $\sigma_x^2 = \sigma_y^2 = \sigma_u^2 = 1 \times 10^{-5}$, $\beta_{XU} = \beta_{YU} = 0.3$. We set $n_X = n_Y = 500,000$ to reflect the sample size of a typical GWAS in our real data analyses. We further set $\pi_1 + \pi_2 + \pi_3 = 0.02$, $\pi_2 = \pi_3$, $\pi_4 = 0.01$, and $\pi_5 = 0.97$. We vary the proportion of invalid IVs, which is defined as $(\pi_2 + \pi_3)/(\pi_1 + \pi_2 + \pi_3)$, to simulate different situations.

Our proposed CARE estimator is compared with widely used IVW method [8] and seven other popular, recently proposed robust MR methods, including cML and cML-DP [53], MR-Egger [2], Weighted-Median [3], MR-mix [38], Weighted-Mode [22], MR-APSS [25], RAPS [56], contamination mixture [ContMix; 9], and MR-Lasso [40]. For IVW, we use the random effects version, which accounts for invalid IVs by allowing over-dispersion in the regression model. For CARE and MR-APSS, we set the significance threshold at 5×10^{-5} . Following common practice, for other benchmark methods, we set the cut-off value λ at 5.45 (corresponding to the significance threshold 5×10^{-8}). In our numerical studies, we used $\eta = 0.5$ as the default value in the winner's curse

removal step. We simulate 500 Monte Carlo repetitions to evaluate empirical statistical power ($\theta \neq 0$) and 1,000 Monte Carlo repetitions to evaluate Type 1 error rates ($\theta = 0$).

We report our simulation results with five measures: Type 1 error rates (proportions of mistaken rejection under $\theta = 0$), power (proportions of p-values less than the significance threshold 0.05 under $\theta \neq 0$), absolute bias (the absolute difference between the estimated $\hat{\theta}$ and the true θ), mean squared error (the average squared difference between the estimated $\hat{\theta}$ and the true θ), and coverage probability (average coverage probability of the 95% confidence interval).

S.8.2 Different proportions of invalid IVs, CARE without winner's curse, and running time

We conduct several additional simulations. We generate the parameters using the following distribution:

$$\begin{pmatrix} \gamma_j \\ \alpha_j \\ \phi_j \end{pmatrix} \sim \underbrace{\pi_1 \begin{pmatrix} N(0, \sigma_x^2) \\ \delta_0 \\ \delta_0 \end{pmatrix}}_{\text{Valid IVs}} + \underbrace{\pi_2 \begin{pmatrix} N(0, \sigma_x^2) \\ N(0.015, \sigma_u^2) \\ N(0, \sigma_u^2) \end{pmatrix}}_{\text{correlated pleiotropy}} + \underbrace{\pi_3 \begin{pmatrix} N(0, \sigma_x^2) \\ N(0, \sigma_y^2) \\ \delta_0 \end{pmatrix}}_{\text{uncorrelated pleiotropy}} + \underbrace{\pi_4 \begin{pmatrix} \delta_0 \\ N(0, \sigma_y^2) \\ \delta_0 \end{pmatrix}}_{\text{IVs fail the relevance assumption}} + \underbrace{\pi_5 \begin{pmatrix} \delta_0 \\ \delta_0 \\ \delta_0 \end{pmatrix}}_{\text{IVs fail the relevance assumption}}, \quad (S10)$$

We follow the main simulation setting and set $\pi_1 + \pi_2 + \pi_3 = 0.02$, $\pi_4 = 0.01$, and $\pi_5 = 0.97$. We vary the proportion of invalid IVs, which is defined as $(\pi_2 + \pi_3)/(\pi_1 + \pi_2 + \pi_3)$, to simulate different situations. Figure S1 summarizes the result to compare the performance of the CARE estimator and CARE estimator without winner's curse bias correction under the setting with 50% invalid IVs. Figures S2 and S3 summarize the results for the settings with 30% and 70% invalid IVs. Figure S4 summarizes the runtime of the CARE estimator and several robust MR methods for the setting with 50% invalid IVs.

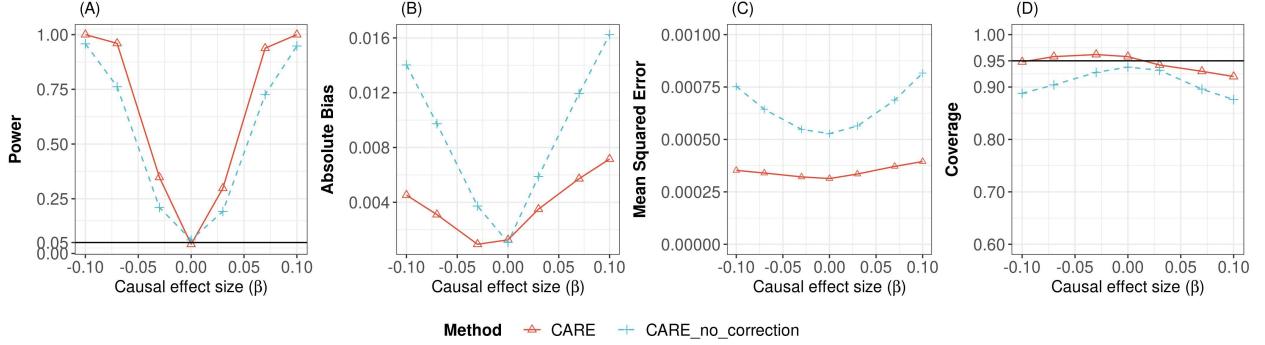


Figure S1: Type 1 error rates, power, absolute bias, mean squared error, and coverage of the CARE estimator and CARE estimator without winner's curse bias correction (CARE_no_correction) under the setting with 50% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

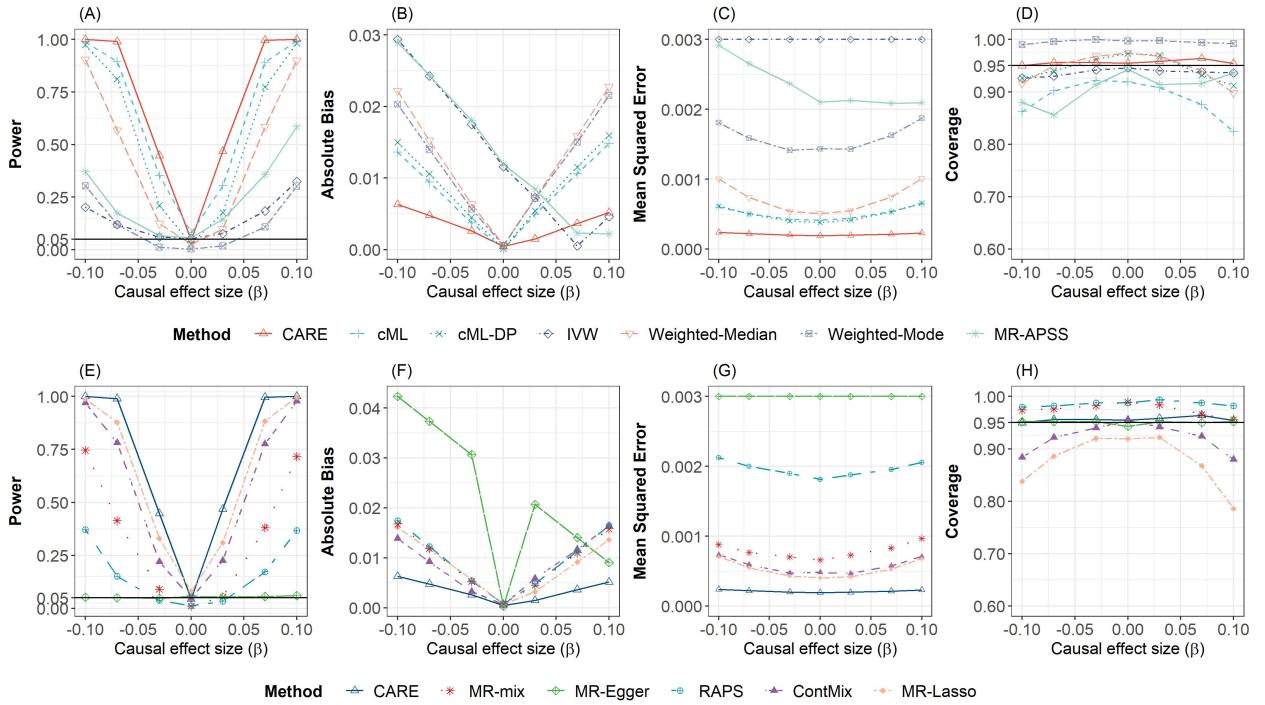


Figure S2: Power, absolute bias, mean squared error, and coverage of the CARE estimator and several robust MR methods under the main setting with 30% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

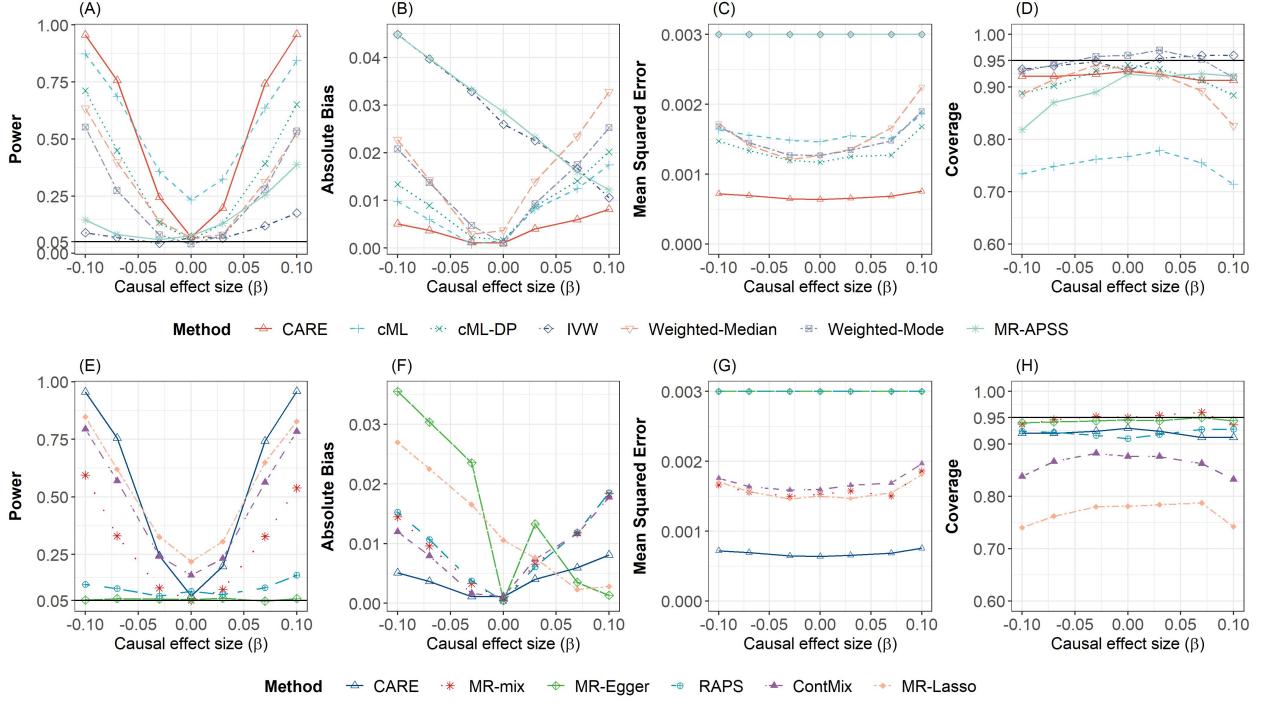


Figure S3: Power, absolute bias, mean squared error, and coverage of the CARE estimator and several robust MR methods under the main setting with 70% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

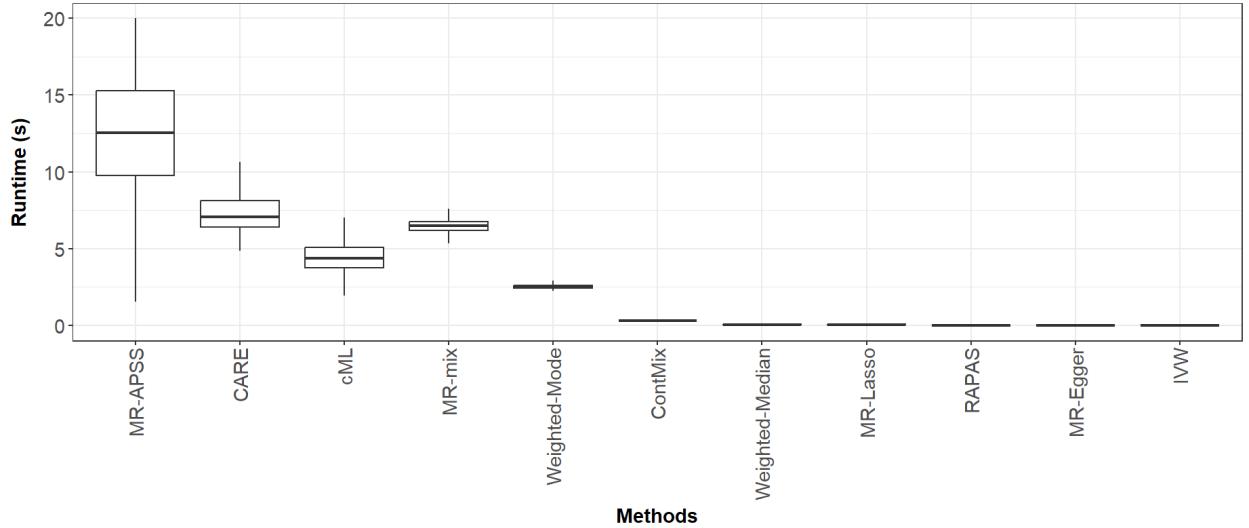


Figure S4: Runtime (in seconds) of the CARE estimator and several robust MR methods under the main setting (12,000 simulations in total). The box limits represent the lower and upper quartiles, the central line represents the median, and the whiskers represent all samples lying within 1.5 times the interquartile range (IQR).

S.8.3 Uniform distributed effects in correlated pleiotropy

Under the setting using uniform distributed effects in correlated pleiotropy, α_j follows the uniform distribution. We generate γ_j, α_j using the following distribution:

$$\begin{pmatrix} \gamma_j \\ \alpha_j \\ \phi_j \end{pmatrix} \sim \underbrace{\pi_1 \begin{pmatrix} N(0, \sigma_x^2) \\ \delta_0 \\ \delta_0 \end{pmatrix}}_{\text{Valid IVs}} + \underbrace{\pi_2 \begin{pmatrix} N(0, \sigma_x^2) \\ U(0.01, 0.03) \\ N(0, \sigma_u^2) \end{pmatrix}}_{\text{correlated pleiotropy}} + \underbrace{\pi_3 \begin{pmatrix} N(0, \sigma_x^2) \\ N(0, \sigma_y^2) \\ \delta_0 \end{pmatrix}}_{\text{uncorrelated pleiotropy}} + \underbrace{\pi_4 \begin{pmatrix} \delta_0 \\ N(0, \sigma_y^2) \\ \delta_0 \end{pmatrix}}_{\text{IVs fail the relevance assumption}} + \underbrace{\pi_5 \begin{pmatrix} \delta_0 \\ \delta_0 \\ \delta_0 \end{pmatrix}}_{\text{IVs fail the relevance assumption}}, \quad (S11)$$

We follow the main simulation setting and set $\pi_1 + \pi_2 + \pi_3 = 0.02$, $\pi_4 = 0.01$, and $\pi_5 = 0.97$. We vary the proportion of invalid IVs, which is defined as $(\pi_2 + \pi_3)/(\pi_1 + \pi_2 + \pi_3)$, to simulate different situations. Figures S5 to S7 summarize the results for the settings with 30%, 50%, and 70% invalid IVs.

S.8.4 Balanced horizontal pleiotropy with InSIDE assumption satisfied

Under the setting of balanced horizontal pleiotropy with the InSIDE assumption satisfied, we allow the InSIDE assumption to be satisfied by setting $\phi_j = 0$. We generate γ_j, α_j using the following distribution:

$$\begin{pmatrix} \gamma_j \\ \alpha_j \end{pmatrix} \sim \underbrace{\pi_1 \begin{pmatrix} N(0, \sigma_x^2) \\ \delta_0 \end{pmatrix}}_{\text{Valid IVs}} + \underbrace{\pi_3 \begin{pmatrix} N(0, \sigma_x^2) \\ N(0, \sigma_y^2) \end{pmatrix}}_{\text{uncorrelated pleiotropy}} + \underbrace{\pi_4 \begin{pmatrix} \delta_0 \\ N(0, \sigma_y^2) \end{pmatrix}}_{\text{IVs fail the relevance assumption}} + \underbrace{\pi_5 \begin{pmatrix} \delta_0 \\ \delta_0 \end{pmatrix}}_{\text{IVs fail the relevance assumption}}.$$

We follow the main simulation setting and set $\pi_1 + \pi_3 = 0.02$, $\pi_4 = 0.01$, and $\pi_5 = 0.97$. We vary the proportion of invalid IVs, which is defined as $(\pi_3)/(\pi_1 + \pi_3)$, to simulate different situations. Figures S8 to S10 summarize the results for the settings with 30%, 50%, and 70% invalid IVs.

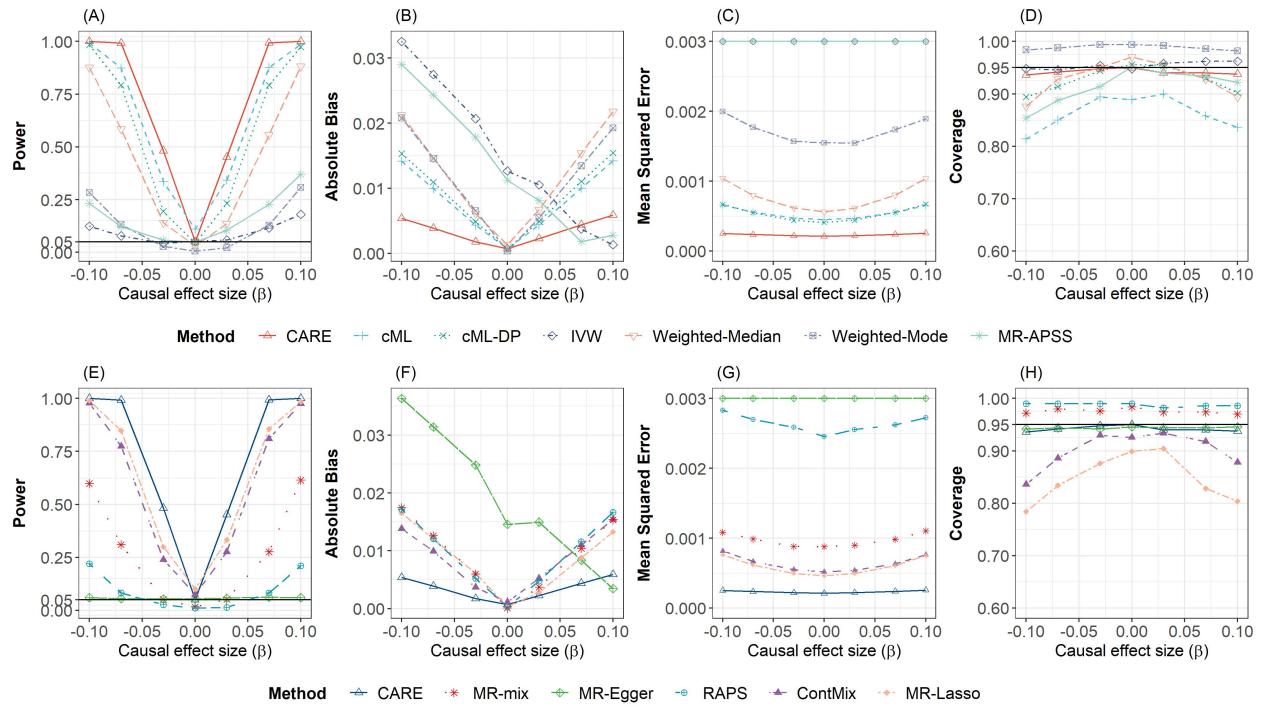


Figure S5: Power, absolute bias, mean squared error, and coverage of the CARE estimator and several robust MR methods under the setting of uniformly distributed effects in correlated pleiotropy with 30% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

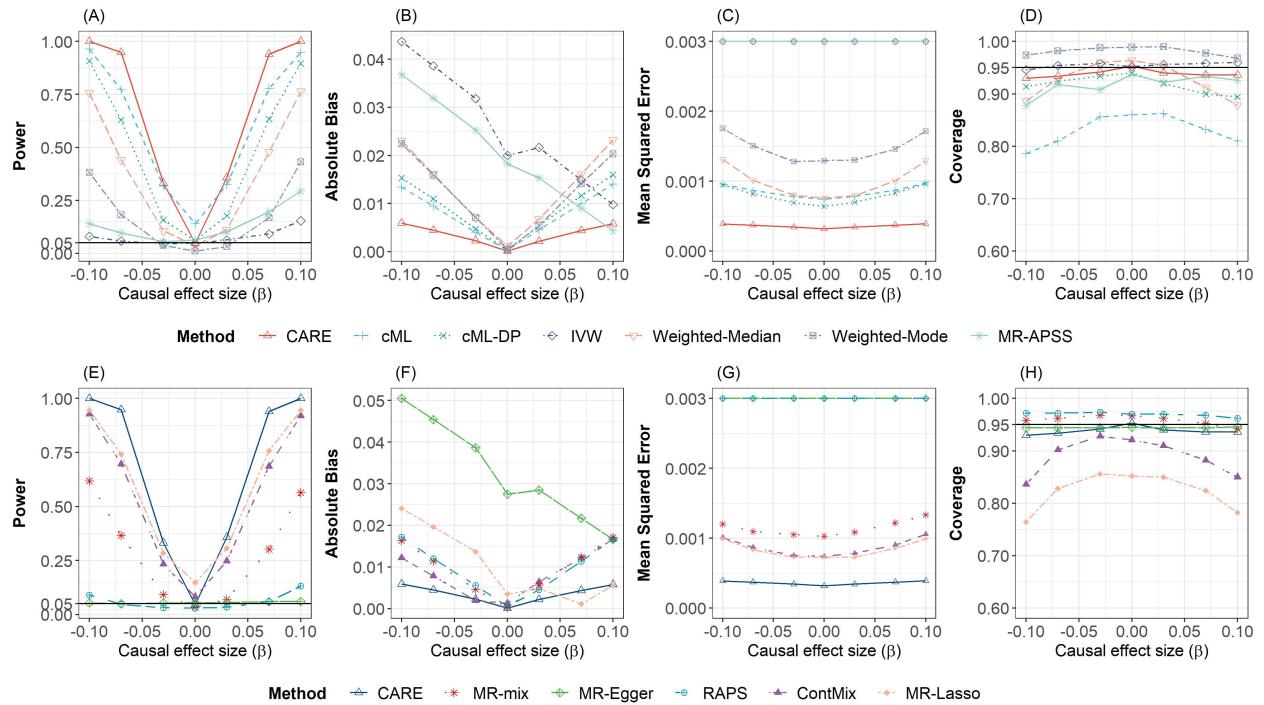


Figure S6: Power, absolute bias, mean squared error, and coverage of the CARE estimator and several robust MR methods under the setting of uniform distributed effects in correlated pleiotropy with 50% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

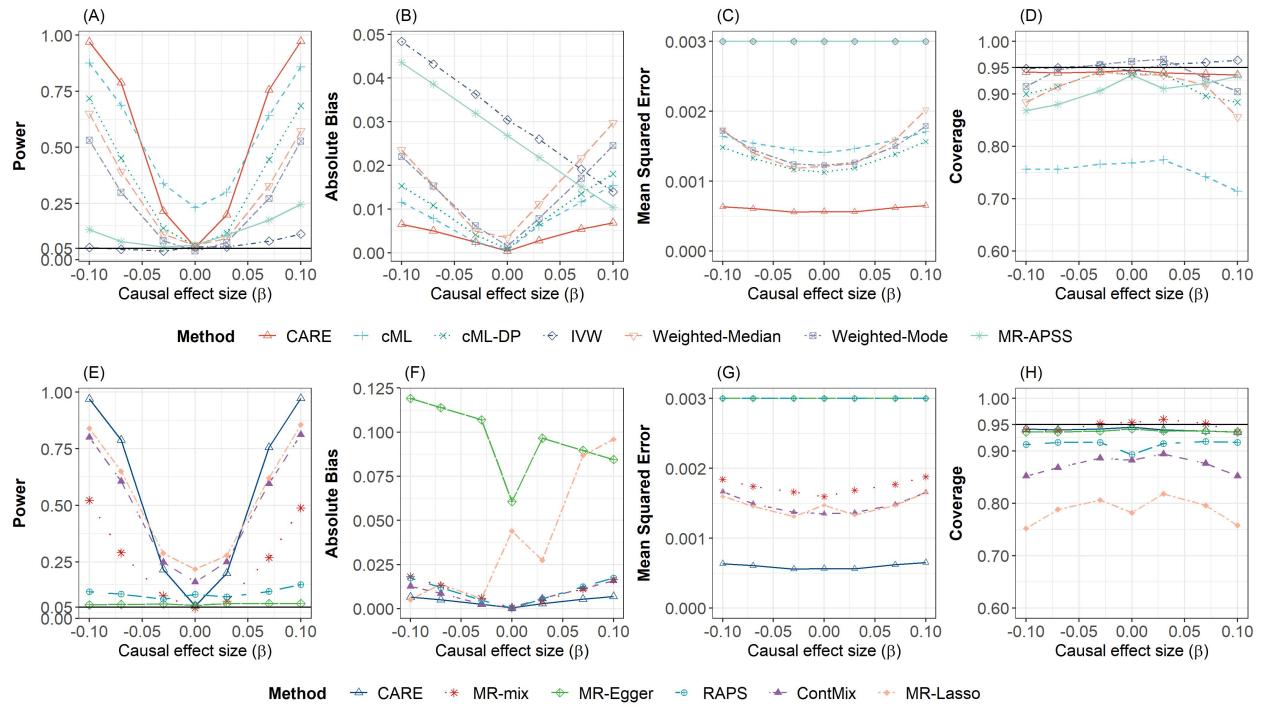


Figure S7: Power, absolute bias, mean squared error, and coverage of the CARE estimator and several robust MR methods under the setting of uniformly distributed effects in correlated pleiotropy with 70% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

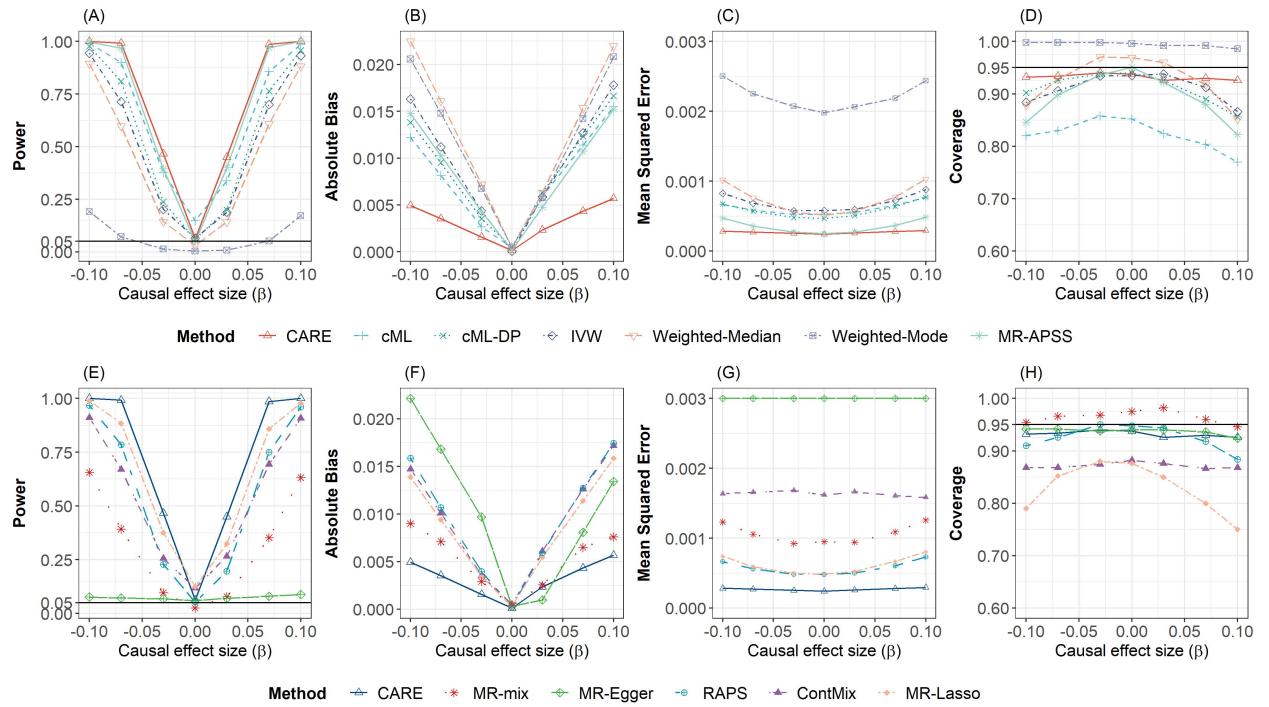


Figure S8: Power, absolute bias, mean squared error, and coverage of the CARE estimator and several robust MR methods under the setting of balanced horizontal pleiotropy with InSIDE assumption satisfied with 30% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

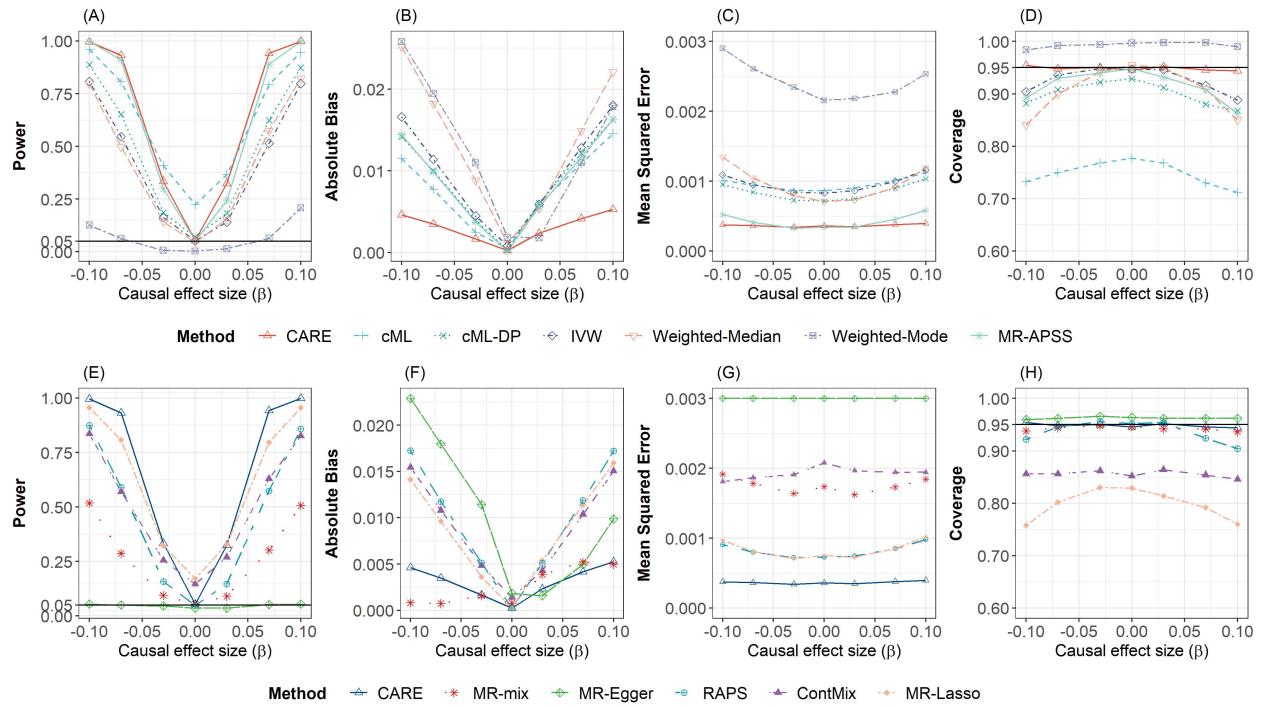


Figure S9: Power, absolute bias, mean squared error, and coverage of the CARE estimator and several robust MR methods under the setting of balanced horizontal pleiotropy with InSIDE assumption satisfied with 50% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

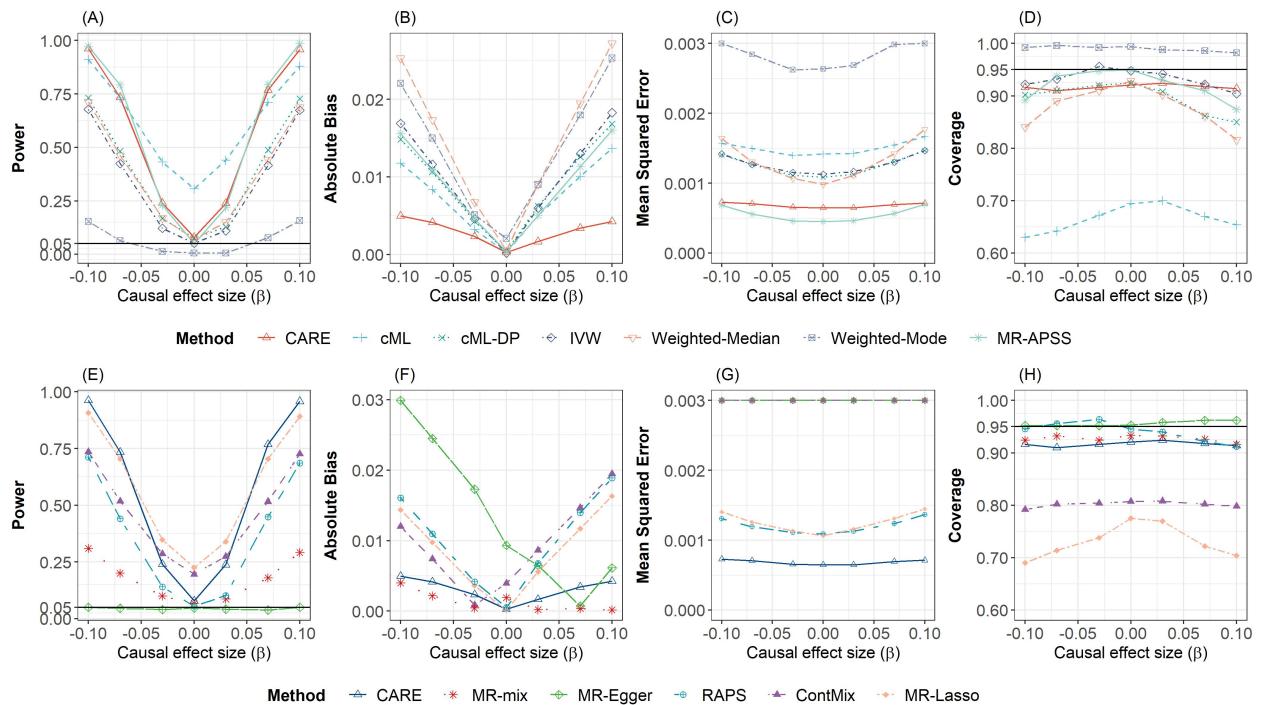


Figure S10: Power, absolute bias, mean squared error, and coverage of the CARE estimator and several robust MR methods under the setting of balanced horizontal pleiotropy with InSIDE assumption satisfied with 70% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

S.8.5 Directional horizontal pleiotropy with InSIDE assumption violated

Under the setting of directional horizontal pleiotropy with InSIDE assumption violated, we generate the underlying parameters using the following distribution:

$$\begin{pmatrix} \gamma_j \\ \alpha_j \\ \phi_j \end{pmatrix} \sim \underbrace{\pi_1 \begin{pmatrix} N(0, \sigma_x^2) \\ \delta_0 \\ \delta_0 \end{pmatrix}}_{\text{Valid IVs}} + \underbrace{\pi_2 \begin{pmatrix} N(0, \sigma_x^2) \\ U(0.01, 0.03) \\ N(0, \sigma_u^2) \end{pmatrix}}_{\text{correlated pleiotropy}} + \underbrace{\pi_4 \begin{pmatrix} \delta_0 \\ N(0, \sigma_y^2) \\ \delta_0 \end{pmatrix}}_{\text{IVs fail the relevance assumption}} + \underbrace{\pi_5 \begin{pmatrix} \delta_0 \\ \delta_0 \\ \delta_0 \end{pmatrix}}_{\text{IVs fail the relevance assumption}},$$

We follow the main simulation setting and set $\pi_1 + \pi_2 = 0.02$, $\pi_4 = 0.01$, and $\pi_5 = 0.97$. We vary the proportion of invalid IVs, which is defined as $(\pi_2)/(\pi_1 + \pi_2)$, to simulate different situations. Figures S11 to S13 summarize the results for the settings with 30%, 50%, and 70% invalid IVs.

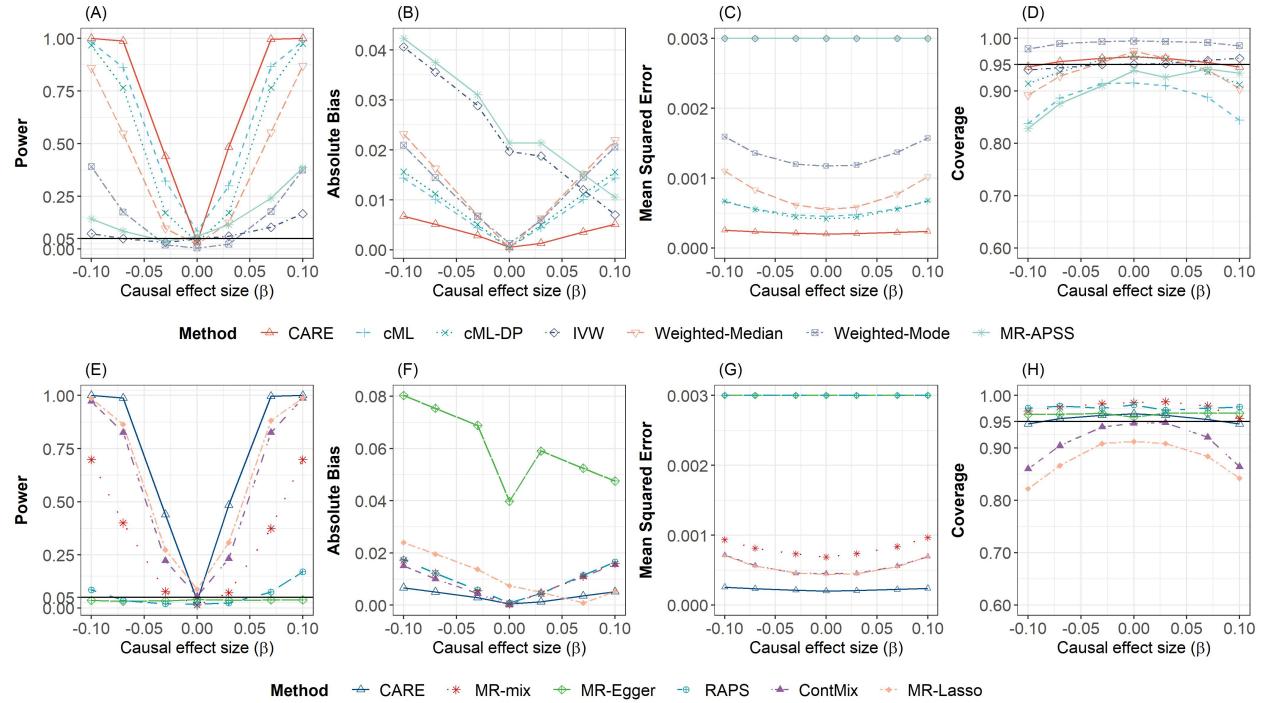


Figure S11: Power, absolute bias, mean squared error, and coverage of the CARE estimator and several robust MR methods under the setting of directional horizontal pleiotropy with InSIDE assumption violated with 30% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

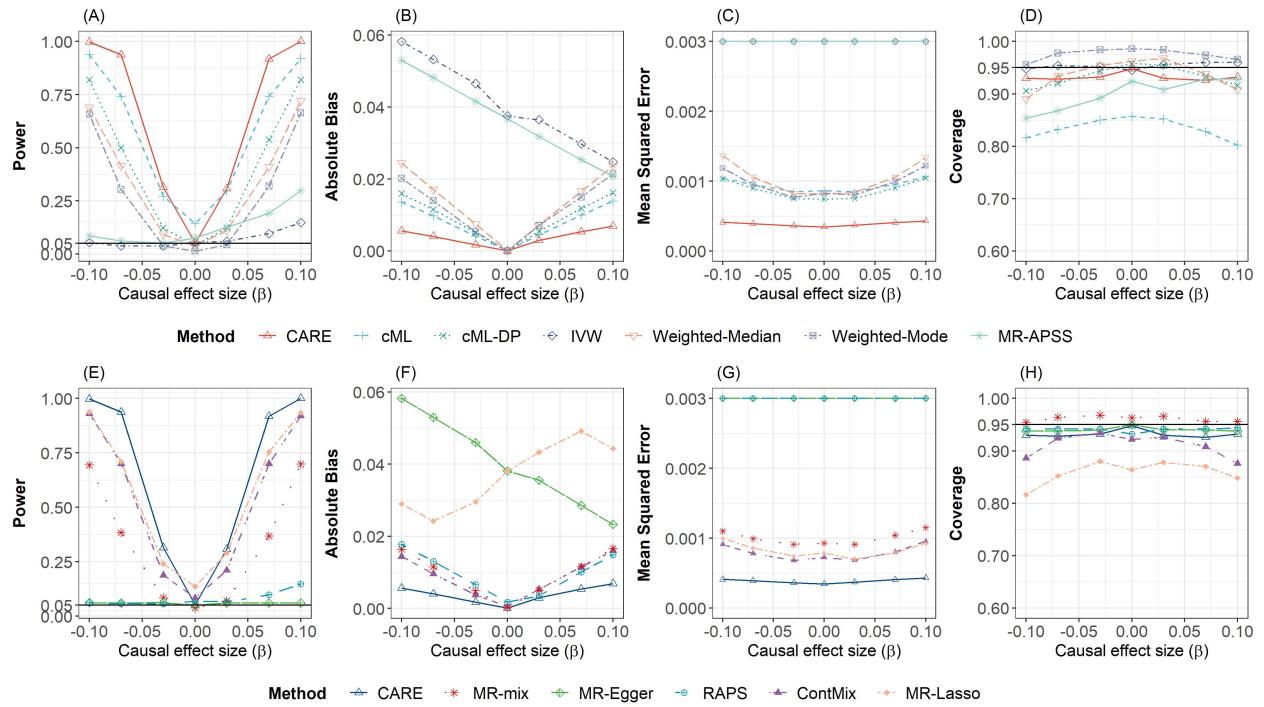


Figure S12: Power, absolute bias, mean squared error, and coverage of the CARE estimator and several robust MR methods under the setting of directional horizontal pleiotropy with InSIDE assumption violated with 50% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

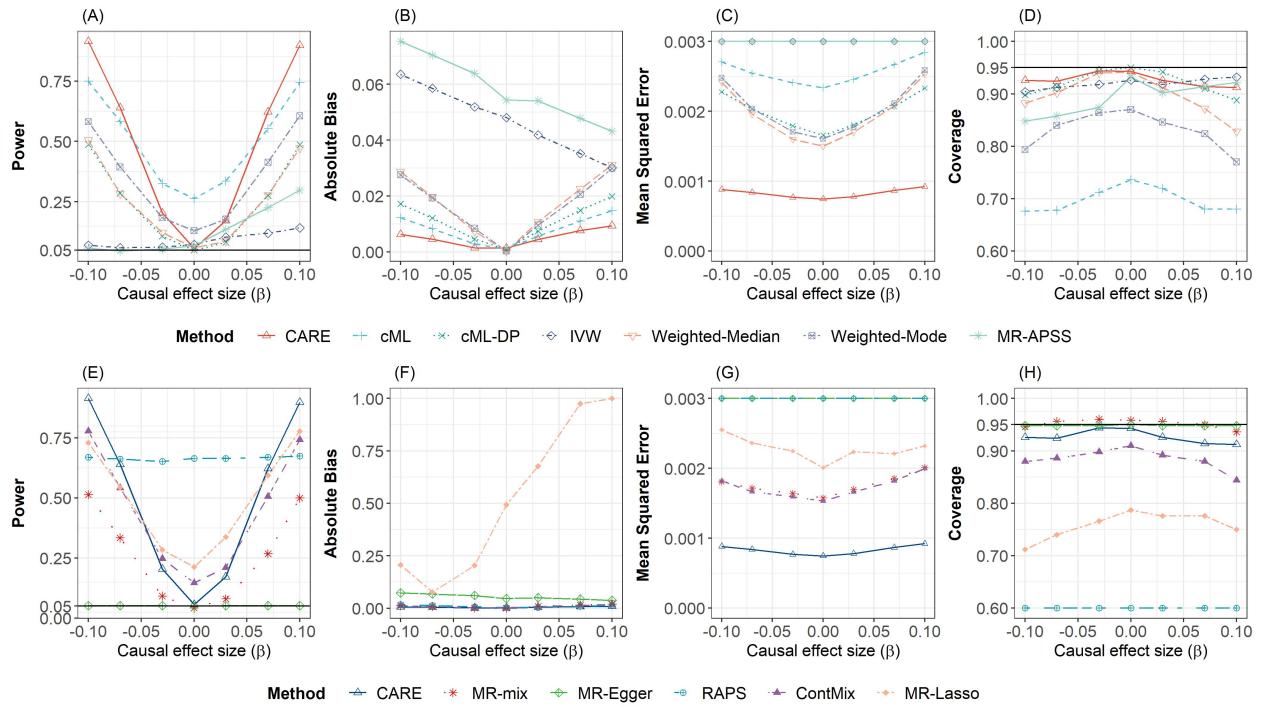


Figure S13: Power, absolute bias, mean squared error, and coverage of the CARE estimator and several robust MR methods under the setting of directional horizontal pleiotropy with InSIDE assumption violated with 70% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

S.8.6 Sensitivity analysis using different values of η

We conducted sensitivity analyses using different values of η (0.1, 0.3, 0.5, 0.7, 0.9) in our main setting. We generate the underlying parameters using the following distribution:

$$\begin{pmatrix} \gamma_j \\ \alpha_j \\ \phi_j \end{pmatrix} \sim \underbrace{\pi_1 \begin{pmatrix} N(0, \sigma_x^2) \\ \delta_0 \\ \delta_0 \end{pmatrix}}_{\text{Valid IVs}} + \underbrace{\pi_2 \begin{pmatrix} N(0, \sigma_x^2) \\ N(0.015, \sigma_u^2) \\ N(0, \sigma_u^2) \end{pmatrix}}_{\text{correlated pleiotropy}} + \underbrace{\pi_3 \begin{pmatrix} N(0, \sigma_x^2) \\ N(0, \sigma_y^2) \\ \delta_0 \end{pmatrix}}_{\text{uncorrelated pleiotropy}} + \underbrace{\pi_4 \begin{pmatrix} \delta_0 \\ N(0, \sigma_y^2) \\ \delta_0 \end{pmatrix}}_{\text{IVs fail the relevance assumption}} + \underbrace{\pi_5 \begin{pmatrix} \delta_0 \\ \delta_0 \\ \delta_0 \end{pmatrix}}_{\text{IVs fail the relevance assumption}},$$

We follow the main simulation setting and set $\pi_1 + \pi_2 + \pi_3 = 0.02$, $\pi_4 = 0.01$, and $\pi_5 = 0.97$.

Figures S14 summarize the results for the settings with 50% invalid IVs.

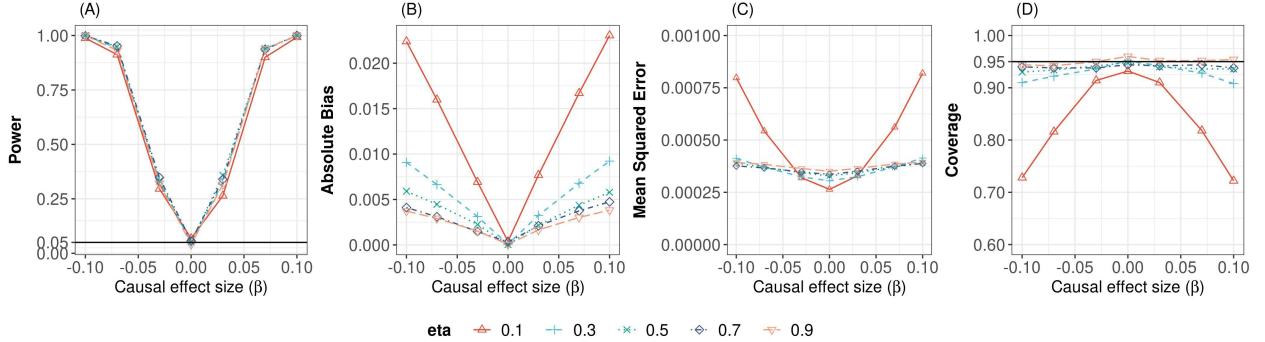


Figure S14: Power, absolute bias, mean squared error, and coverage of the CARE estimator with different η under the main setting. Power is the empirical power estimated by the proportion of p-values less than the significance threshold 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

S.8.7 Consistency of using GBIC with different choices of κ_n as model selection methods

We discuss the adjustment of BIC when s_λ tends to infinity with generalized BIC (GBIC) of the following form:

$$\text{GBIC}(v) = -2\widehat{l}(\widehat{\theta}(v), \{\widehat{r}_j(v)\}_{j \in \widehat{\mathcal{V}}} + \kappa_n \cdot (s_\lambda - v), \quad s_\lambda = |\mathcal{S}_\lambda|.$$

We tested two choices of κ_n : (i) $\kappa_n = \log n$ and (ii) $\kappa_n = \log(s_\lambda) \cdot \log(\log(n))$, both satisfying $\kappa_n \gg \log(s_\lambda)$. We generate the underlying parameters using the following distribution:

$$\begin{pmatrix} \gamma_j \\ \alpha_j \\ \phi_j \end{pmatrix} \sim \underbrace{\pi_1 \begin{pmatrix} N(0, \sigma_x^2) \\ \delta_0 \\ \delta_0 \end{pmatrix}}_{\text{Valid IVs}} + \underbrace{\pi_2 \begin{pmatrix} N(0, \sigma_x^2) \\ N(0.015, \sigma_u^2) \\ N(0, \sigma_u^2) \end{pmatrix}}_{\text{correlated pleiotropy}} + \underbrace{\pi_3 \begin{pmatrix} N(0, \sigma_x^2) \\ N(0, \sigma_y^2) \\ \delta_0 \end{pmatrix}}_{\text{uncorrelated pleiotropy}} + \underbrace{\pi_4 \begin{pmatrix} \delta_0 \\ N(0, \sigma_y^2) \\ \delta_0 \end{pmatrix}}_{\text{IVs fail the relevance assumption}} + \underbrace{\pi_5 \begin{pmatrix} \delta_0 \\ \delta_0 \\ \delta_0 \end{pmatrix}}_{\text{IVs fail the relevance assumption}},$$

We follow the main simulation setting and set $\pi_1 + \pi_2 + \pi_3 = 0.02$, $\pi_4 = 0.01$, and $\pi_5 = 0.97$.

Figures S15 and S16 summarize the results for the settings with 50% invalid IVs.

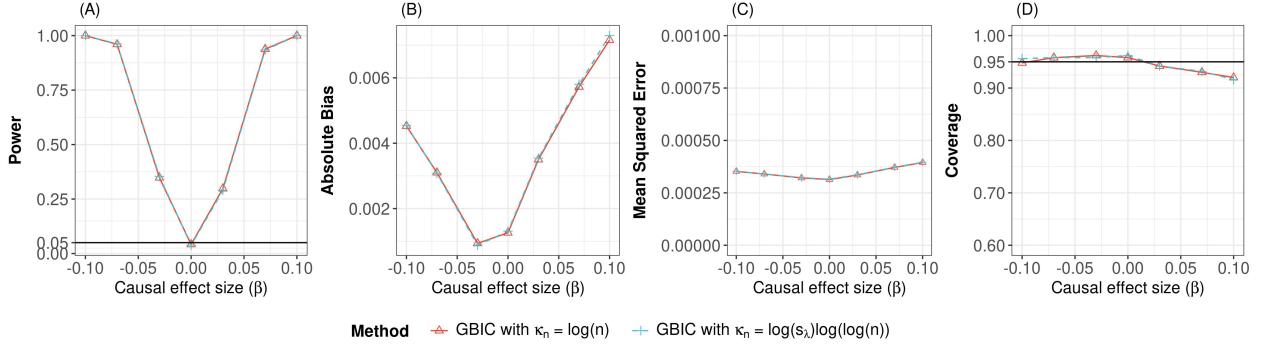


Figure S15: Power, absolute bias, mean squared error, and coverage of the CARE estimator using GBIC with $\kappa_n = \log n$ and $\kappa_n = \log(s_\lambda) \cdot \log(\log(n))$ under the main setting. Power is the empirical power estimated by the proportion of p-values less than the significance threshold 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

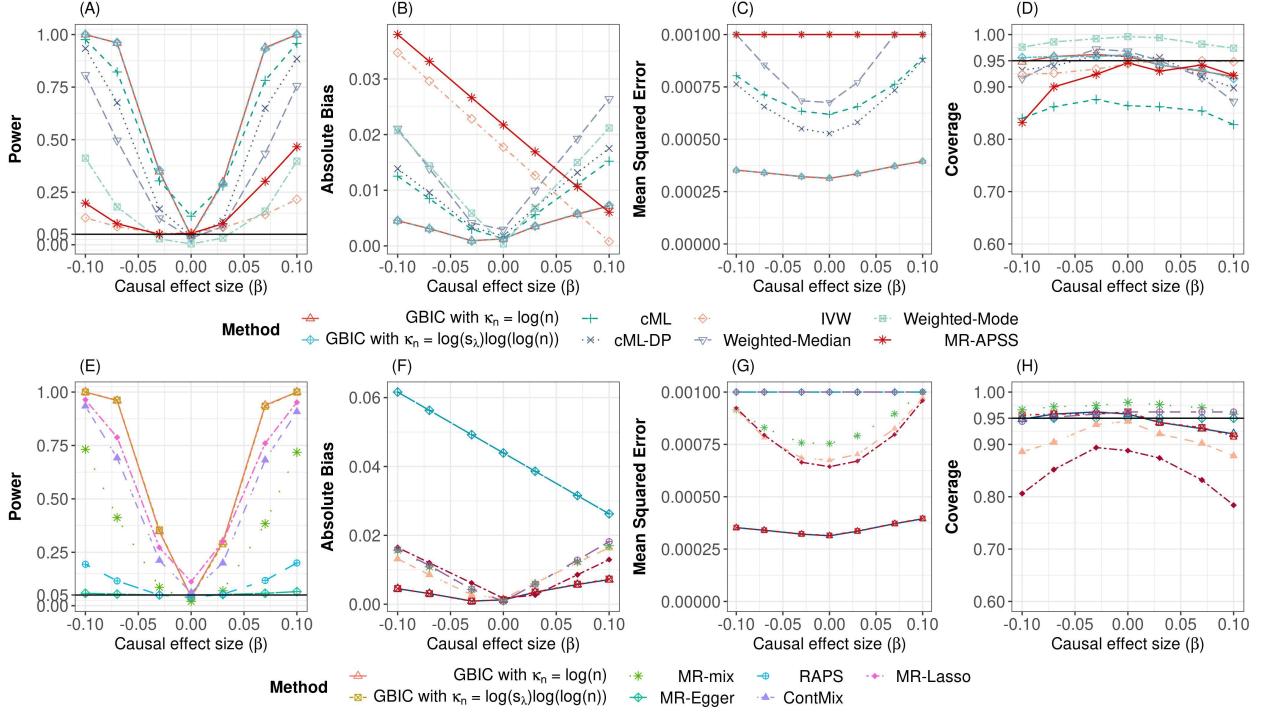


Figure S16: Comparison of Power, absolute bias, mean squared error, and coverage of the CARE estimator using GBIC with $\kappa_n = \log n$ and $\kappa_n = \log(s_\lambda) \cdot \log(\log(n))$ and other benchmark methods under the main setting with 50% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

S.8.8 Nonlinear settings

In our non-linear simulation settings, we implement a four-step process to model complex genetic relationships. First, we simulate p mutually independent single nucleotide polymorphisms (SNPs), denoted as $\mathbf{G} = (G_1, \dots, G_p)^T$. Each SNP G_j follows a $\text{Binomial}(2, \text{MAF}_j)$ distribution, where MAF_j represents the minor allele frequency drawn from a $\text{Uniform}(0.01, 0.5)$ distribution. Next, we simulate an unmeasured confounder U as $U = \sum_{j=1}^p \phi_j G_j + E_U$. The risk factor X is then simulated as $X = \sum_{j=1}^p f(G_j) + \beta_{XU} U + E_X$, and finally, the outcome Y is modeled as $Y = \theta X + \beta_{YU} U + \sum_{j=1}^p \alpha_j G_j + E_Y$. In these equations, E_U , E_X , and E_Y represent mutually independent random noise terms, distributed as $E_U \sim \mathcal{N}(0, \sigma_U^2)$, $E_X \sim \mathcal{N}(0, \sigma_X^2)$, and $E_Y \sim \mathcal{N}(0, \sigma_Y^2)$, respectively. These distributions are consistent with the main setting. Similarly, the coefficients γ_j , α_j , and ϕ_j are generated from the same mixture of distributions as described in the main setting. To explore different non-linear relationships, we consider three scenarios. In the first, we focus on non-linearity

in X with a linear Y , where $f(G_j) = \gamma_{1j}G_j^2 + \gamma_{2j}G_j$, with $\gamma_{1j} = \gamma_{2j} = \gamma_j$. The second scenario introduces additional complexity by incorporating interaction terms between SNPs in the model for X , such that $X = \sum_{j=1}^p f(G_j) + \sum_{i,j \in S} \gamma_{ij}G_iG_j + \beta_{XU}U + E_X$, where $f(G_j)$ remains as in the first scenario, and S represents a randomly selected set of 20 SNP pairs for which interaction effects are modeled. The third scenario introduces non-linearity in Y with $Y = \theta^2 X + \beta_{YU}U + \sum_{j=1}^p \alpha_jG_j + E_Y$.

Supplementary Figures S17 to S19 summarize the results for the three scenarios with 50% invalid IVs. In the first two scenarios with non-linear X on G , CARE showed slightly inflated Type 1 error rates, larger bias, and worse coverage (Supplementary Figures S17 and S18). The third scenario revealed that CARE demonstrated diminished power, larger bias, and poor coverage (Supplementary Figure S19).

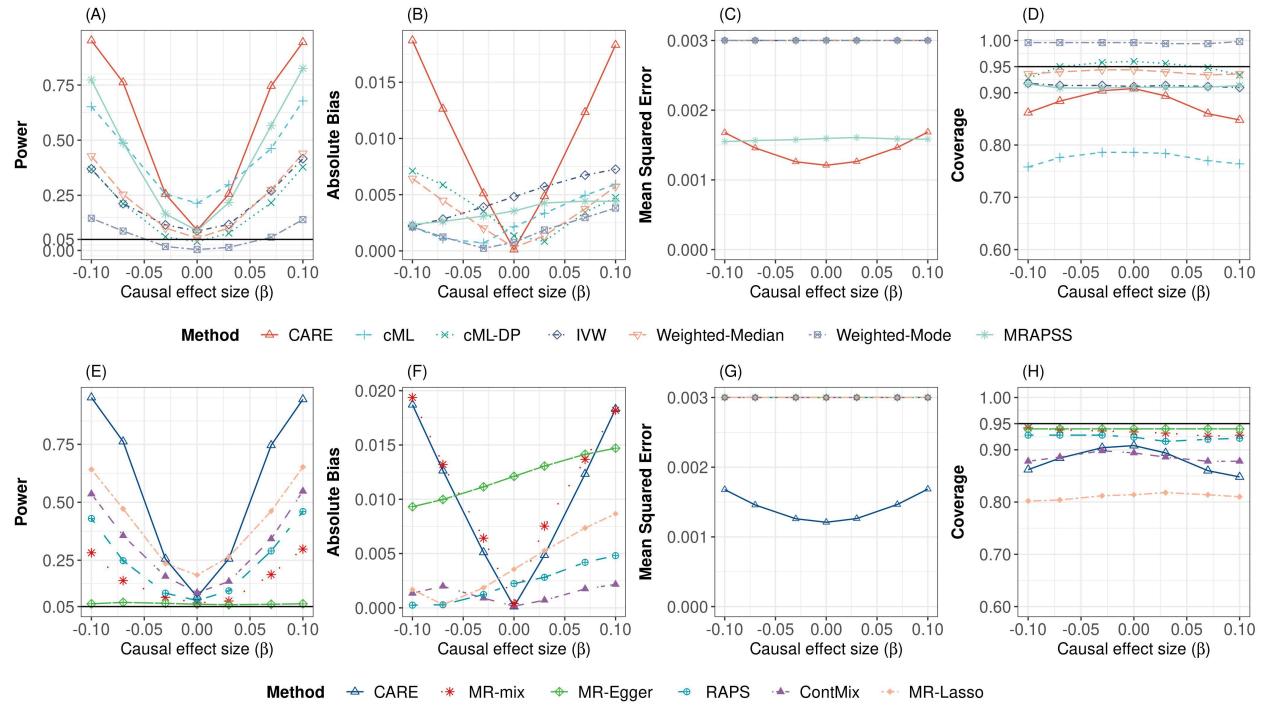


Figure S17: Power, absolute bias, mean squared error, and coverage of the CARE estimator and several robust MR methods under the setting of non-linearity in exposure without interaction terms with 50% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

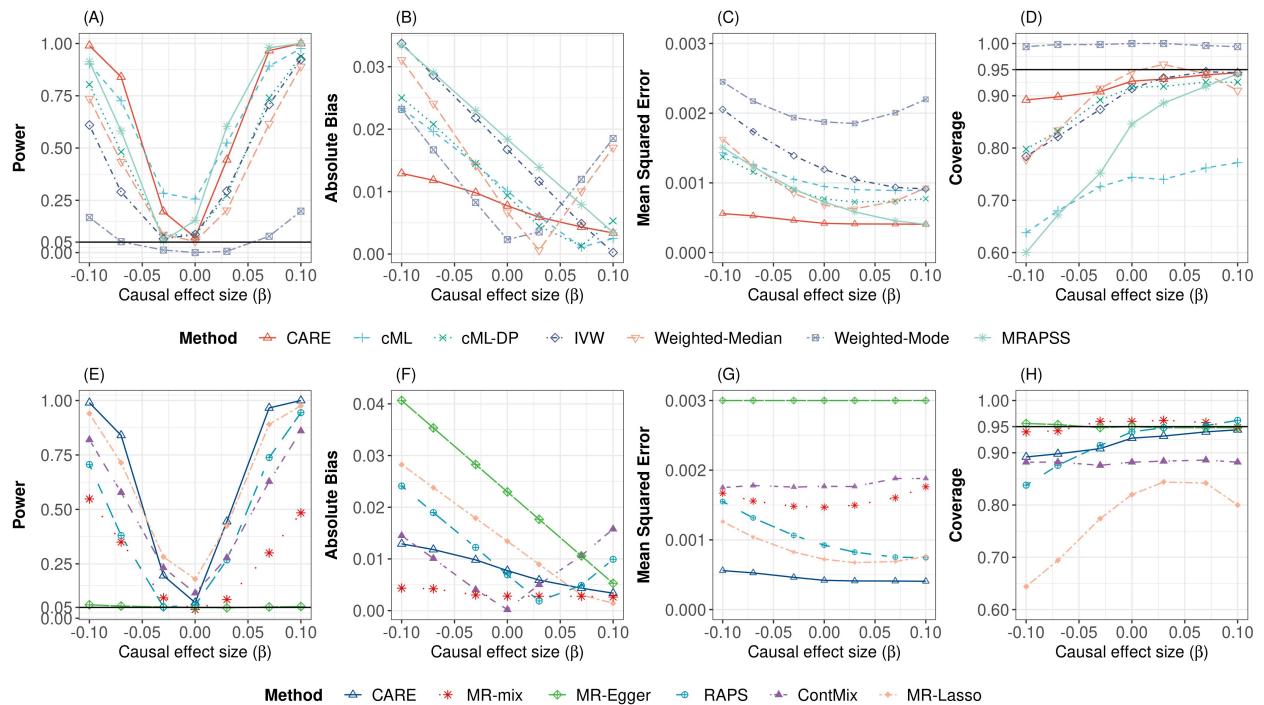


Figure S18: Power, absolute bias, mean squared error, and coverage of the CARE estimator and several robust MR methods under the setting of non-linearity in exposure with interaction terms with 50% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

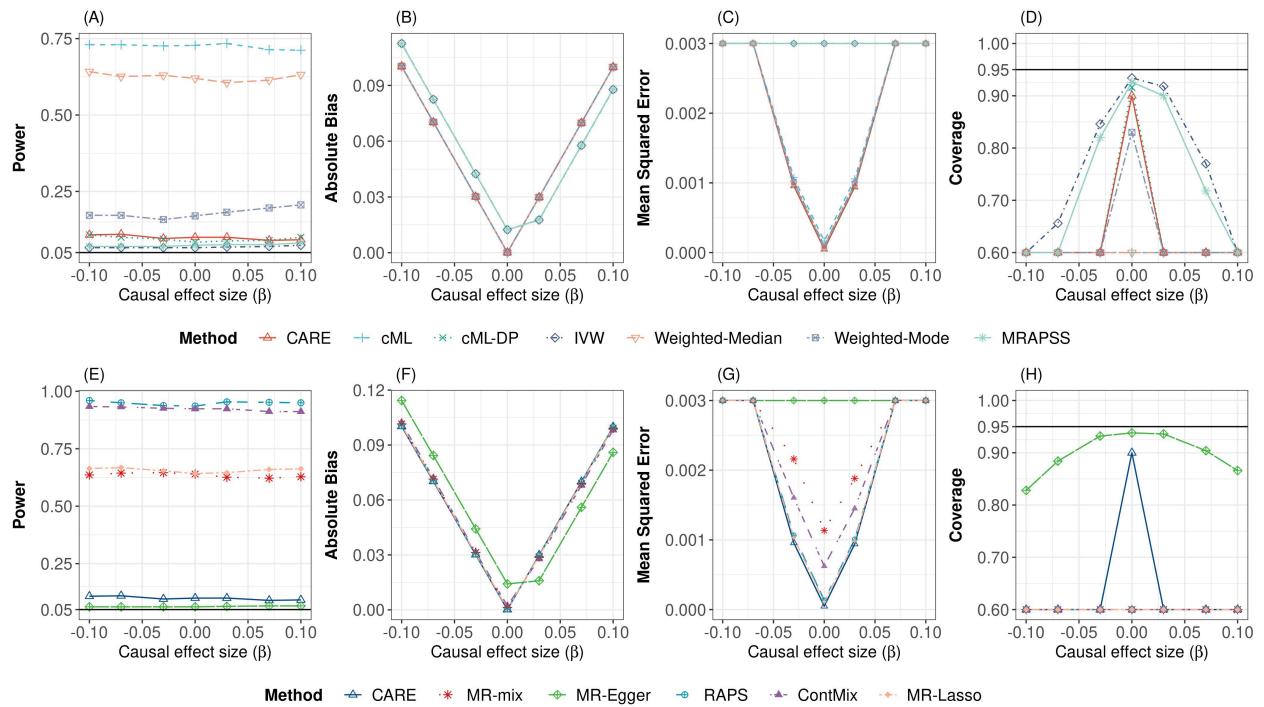


Figure S19: Power, absolute bias, mean squared error, and coverage of the CARE estimator and several robust MR methods under the setting of non-linearity in both exposure and outcome without interaction terms with 50% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

S.8.9 Sample size variation of GWAS

We evaluate the performance of CARE estimator and benchmark MR methods with different sample sizes of both exposure and outcome GWAS (100000, 50000, 10000, 5000). We generate the underlying parameters using the same distribution as the main setting.

We follow the main simulation setting and set $\pi_1 + \pi_2 + \pi_3 = 0.02$, $\pi_4 = 0.01$, and $\pi_5 = 0.97$. We vary the proportion of invalid IVs, which is defined as $(\pi_2 + \pi_3)/(\pi_1 + \pi_2 + \pi_3)$, to simulate different situations. To maintain heritability within a biologically plausible range, we adjust the variance of the risk factor, denoted as σ_X^2 , across different simulation scenarios to maintain reasonable heritability. Figure S20 to S23 summarize the results for different sample sizes. The findings indicate that CARE's performance deteriorates as the sample size of GWAS decreases.

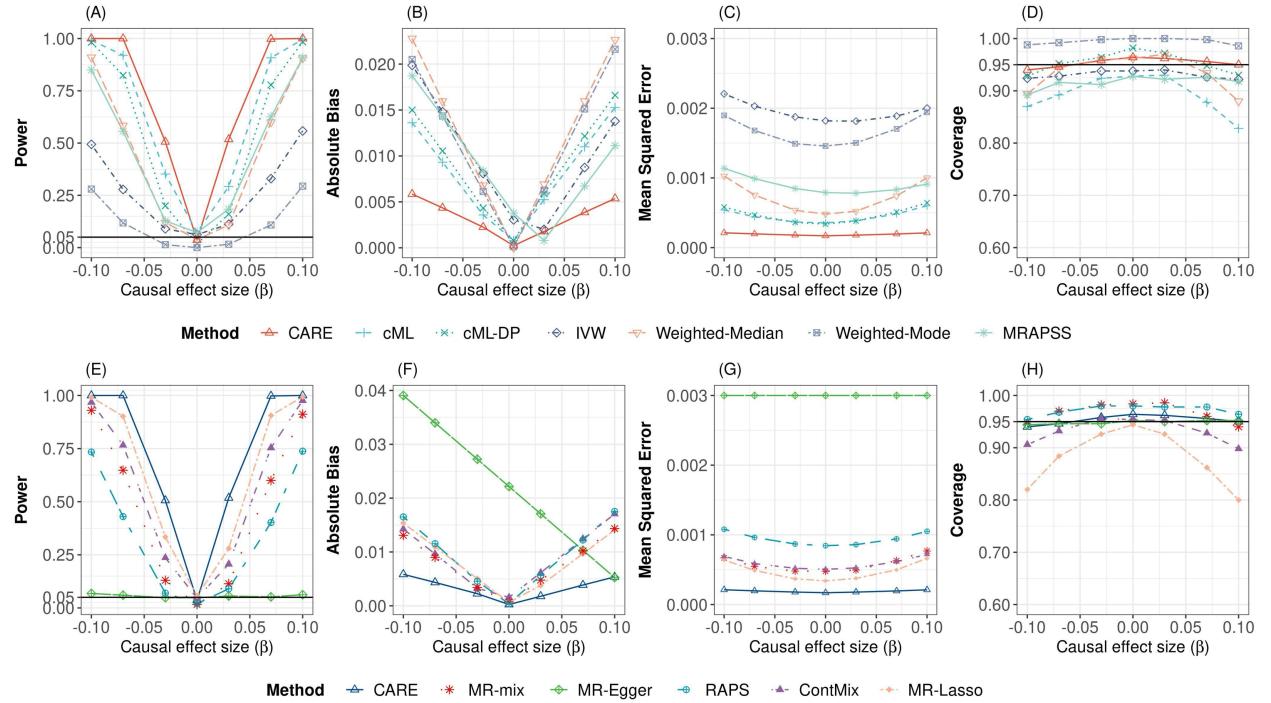


Figure S20: Power, absolute bias, mean squared error, and coverage of the CARE estimator and several robust MR methods with sample size = 100000, $\sigma_x^2 = 1 \times 10^{-5}$ and 50% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

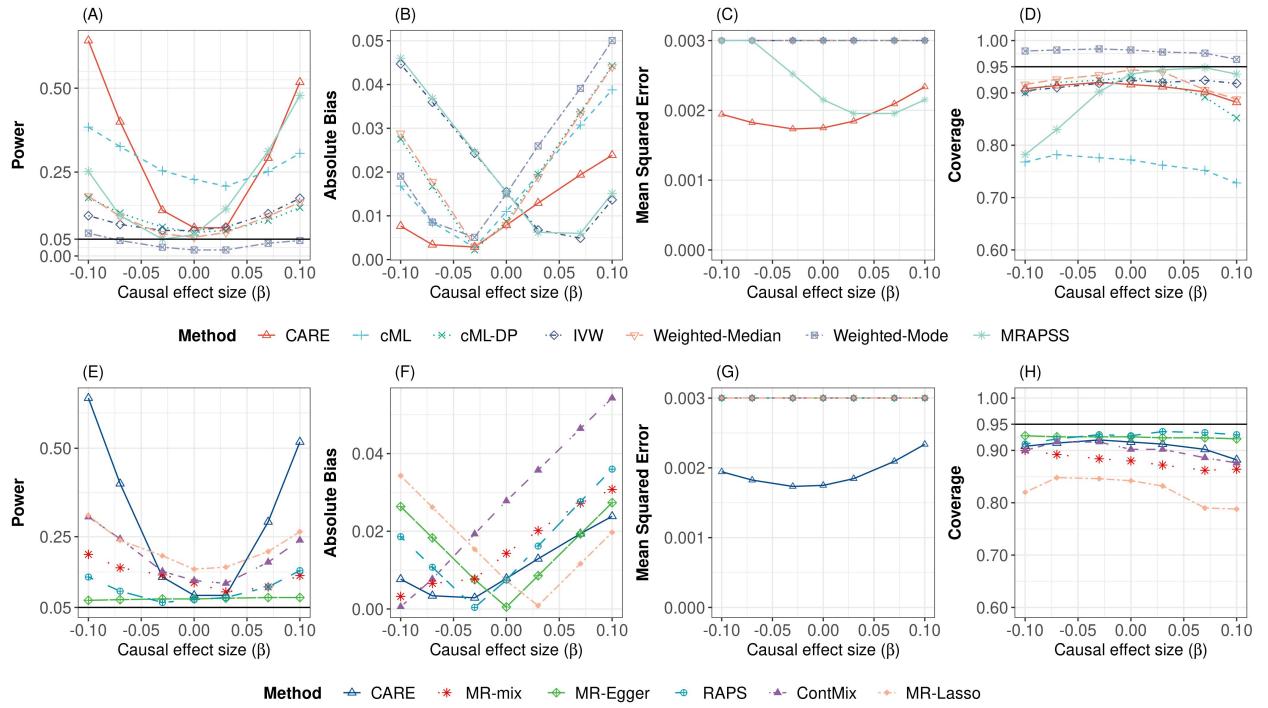


Figure S21: Power, absolute bias, mean squared error, and coverage of the CARE estimator and several robust MR methods with sample size = 50000, $\sigma_x^2 = 5 \times 10^{-5}$ and 50% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

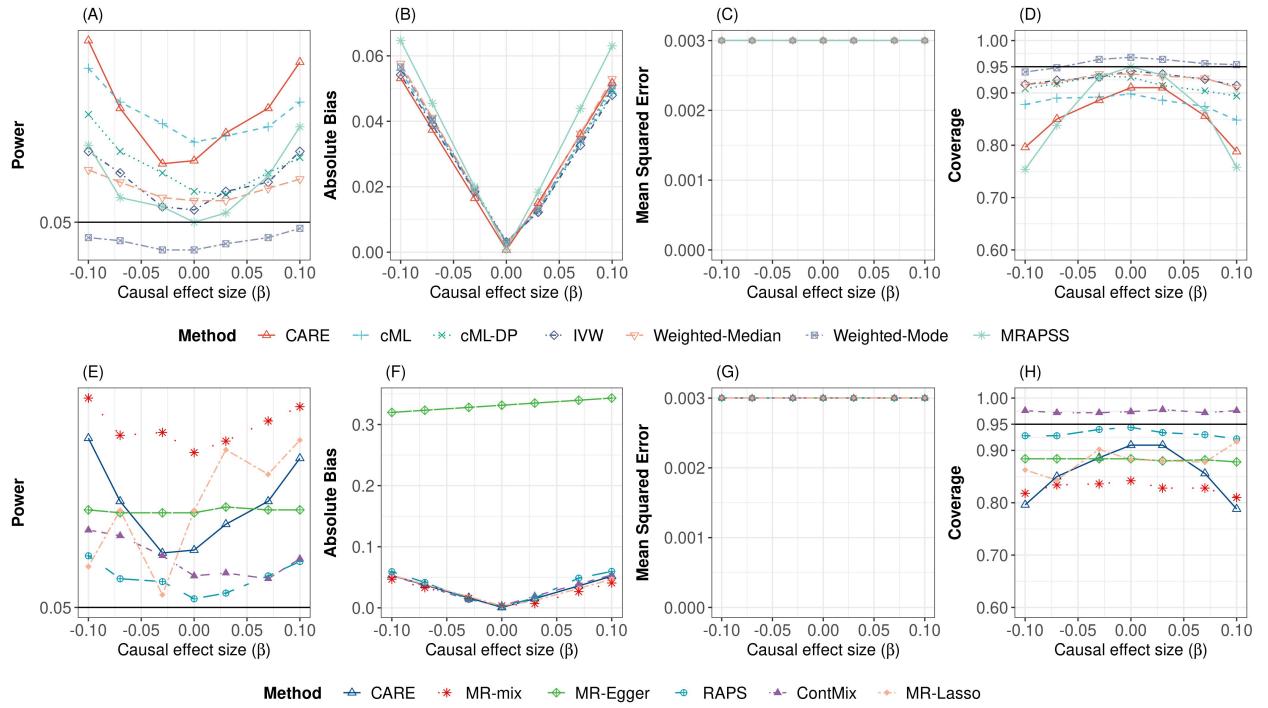


Figure S22: Power, absolute bias, mean squared error, and coverage of the CARE estimator and several robust MR methods with sample size = 10000, $\sigma_x^2 = 8 \times 10^{-5}$ and 50% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

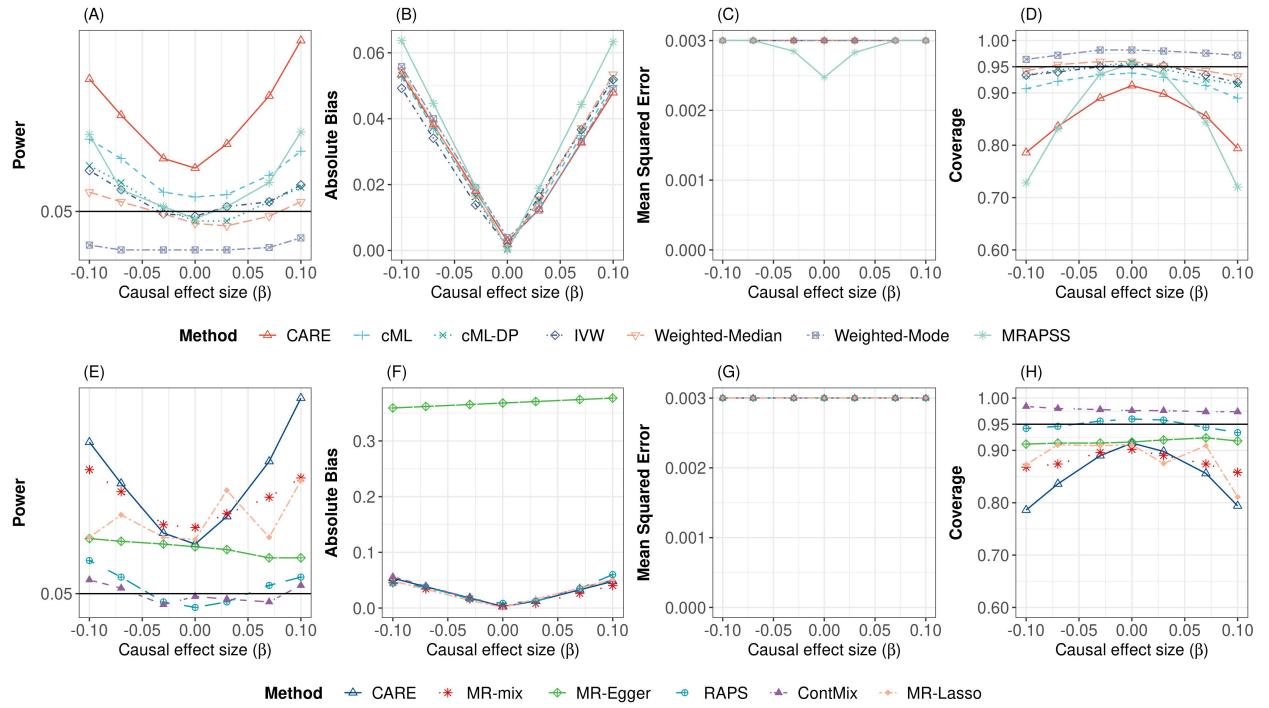


Figure S23: Power, absolute bias, mean squared error, and coverage of the CARE estimator and several robust MR methods with sample size = 5000, $\sigma_x^2 = 1 \times 10^{-4}$ and 50% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

S.8.10 Variations in number of SNPs

We evaluate the performance of CARE estimator and benchmark MR methods with different sample sizes of SNPs (100000, 50000, 10000, 5000, 1000). We generate the underlying parameters using the following distribution:

$$\begin{pmatrix} \gamma_j \\ \alpha_j \\ \phi_j \end{pmatrix} \sim \underbrace{\pi_1 \begin{pmatrix} N(0, \sigma_x^2) \\ \delta_0 \\ \delta_0 \end{pmatrix}}_{\text{Valid IVs}} + \underbrace{\pi_2 \begin{pmatrix} N(0, \sigma_x^2) \\ N(0.015, \sigma_u^2) \\ N(0, \sigma_u^2) \end{pmatrix}}_{\text{correlated pleiotropy}} + \underbrace{\pi_3 \begin{pmatrix} N(0, \sigma_x^2) \\ N(0, \sigma_y^2) \\ \delta_0 \end{pmatrix}}_{\text{uncorrelated pleiotropy}} + \underbrace{\pi_4 \begin{pmatrix} \delta_0 \\ N(0, \sigma_y^2) \\ \delta_0 \end{pmatrix}}_{\text{IVs fail the relevance assumption}} + \underbrace{\pi_5 \begin{pmatrix} \delta_0 \\ \delta_0 \\ \delta_0 \end{pmatrix}}_{\text{IVs fail the relevance assumption}},$$

We follow the main simulation setting and set $\pi_1 + \pi_2 + \pi_3 = 0.02$, $\pi_4 = 0.01$, and $\pi_5 = 0.97$. Supplementary Figure S24 to S27 summarize the results for different sample sizes of SNPs. The findings indicate that CARE's performance deteriorates as the sample size of SNPs decreases.

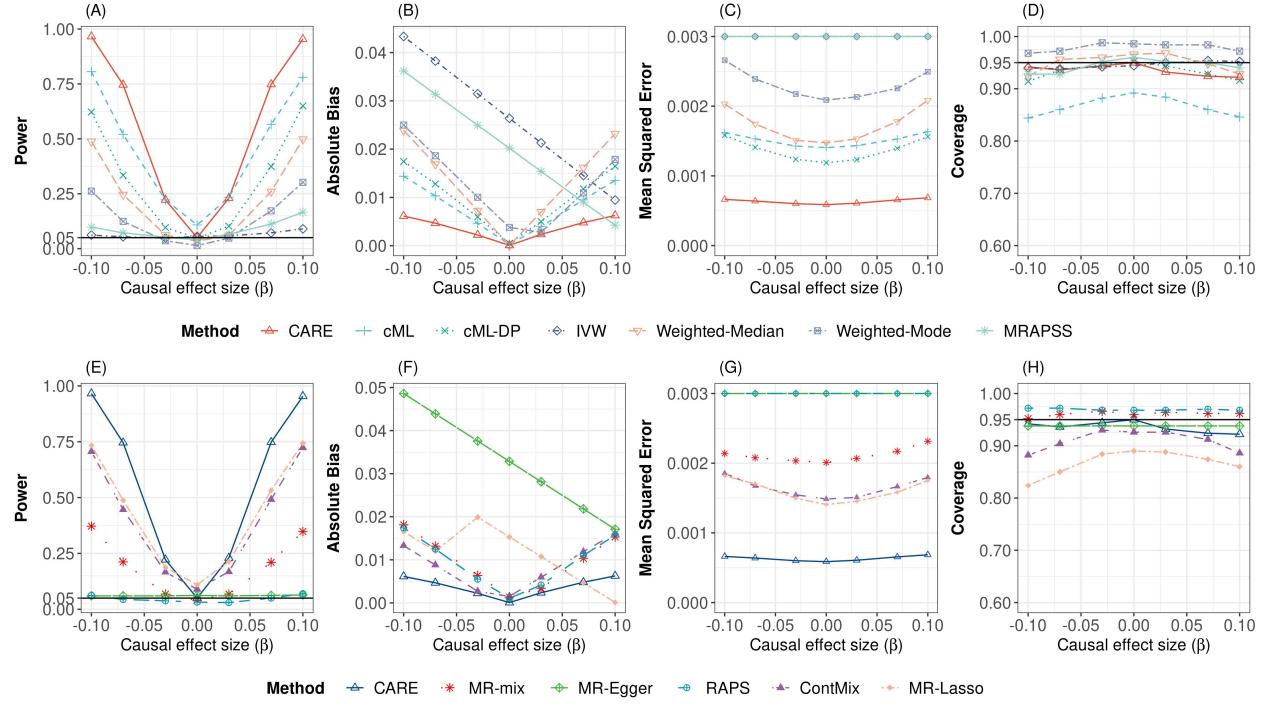


Figure S24: Power, absolute bias, mean squared error, and coverage of the CARE estimator and several robust MR methods with the number of SNPs equal to 100,000, and 50% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

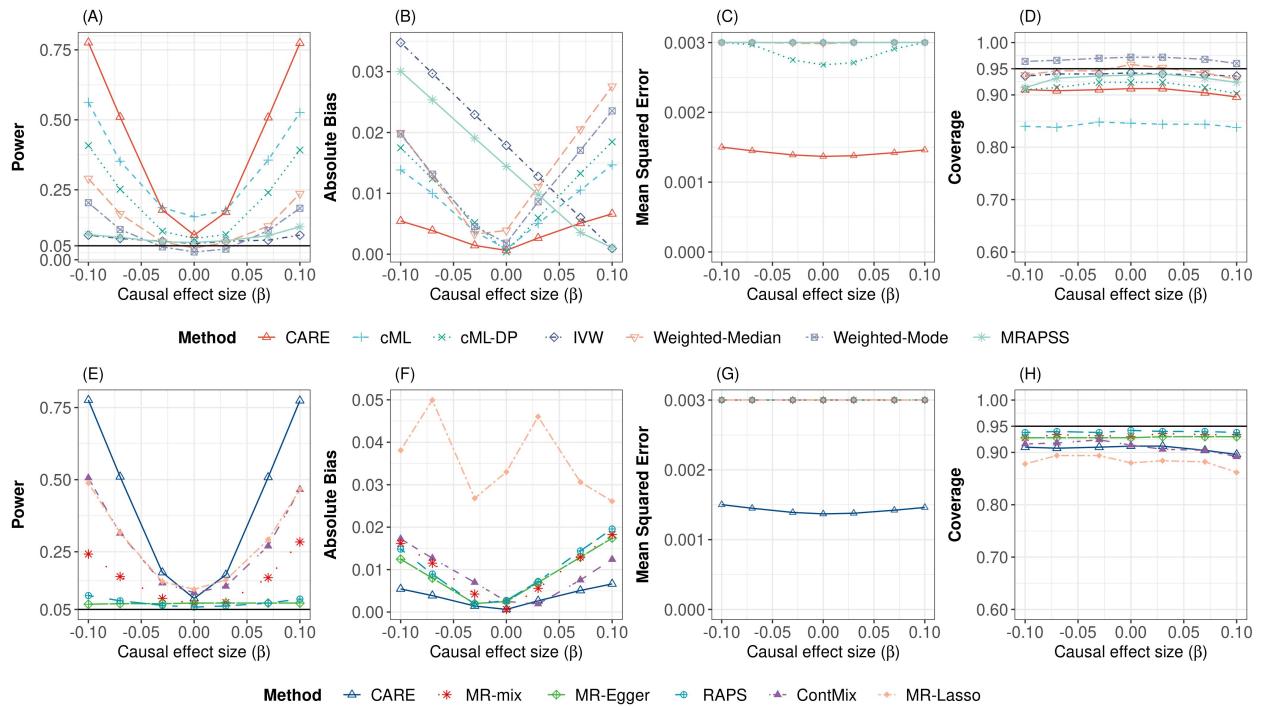


Figure S25: Power, absolute bias, mean squared error, and coverage of the CARE estimator and several robust MR methods with the number of SNPs equal to 50,000, and 50% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

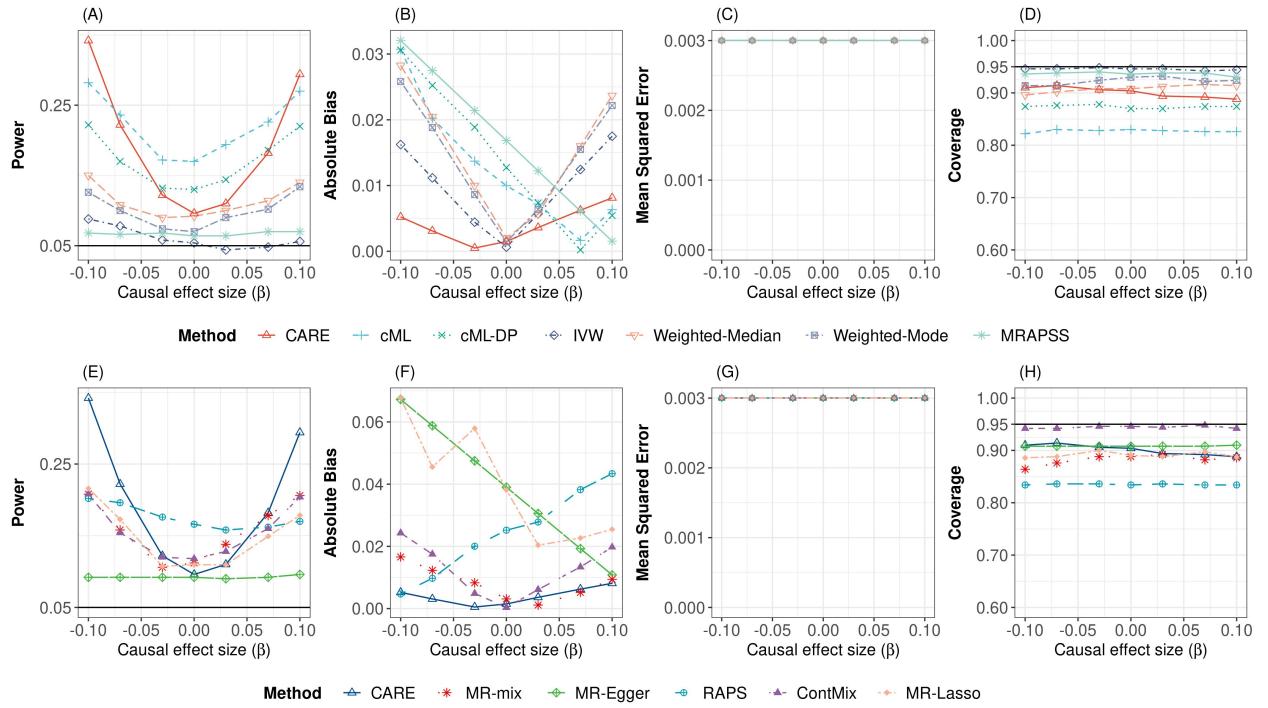


Figure S26: Power, absolute bias, mean squared error, and coverage of the CARE estimator and several robust MR methods with the number of SNPs equal to 10,000, and 50% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

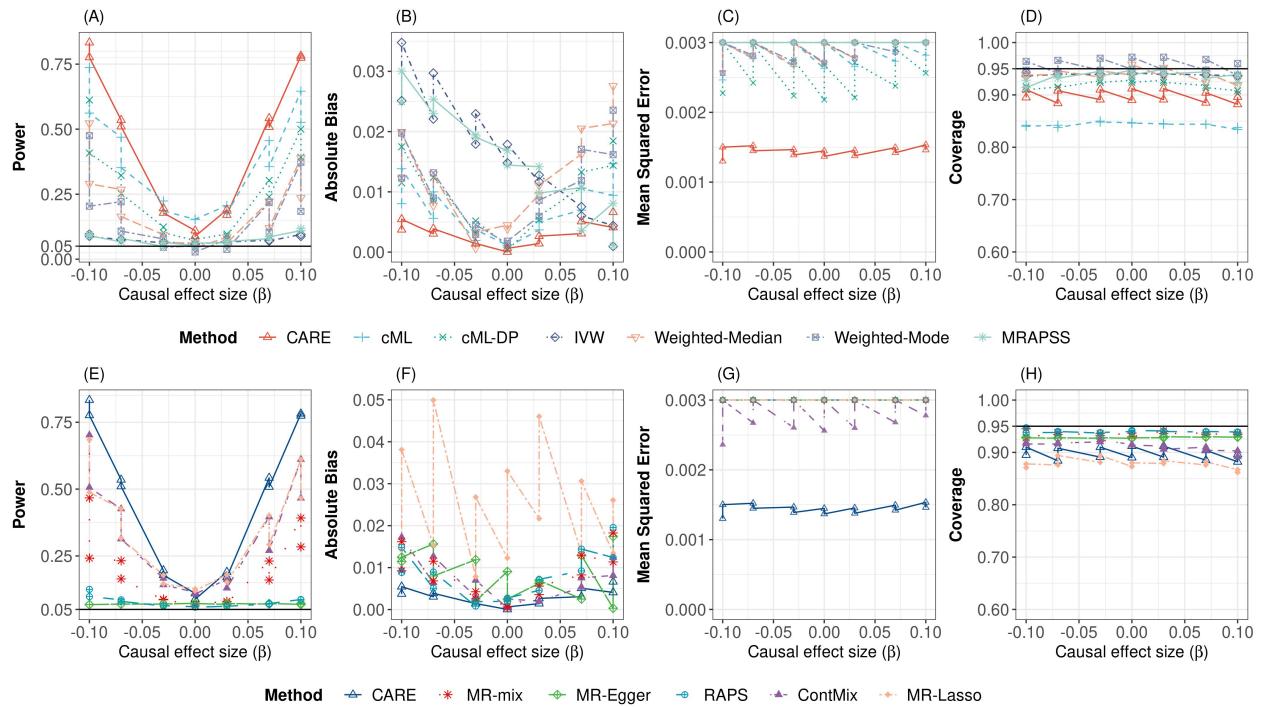


Figure S27: Power, absolute bias, mean squared error, and coverage of the CARE estimator and several robust MR methods with the number of SNPs equal to 5,000, and 50% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

S.8.11 Using the same liberal threshold

To assess the influence of the IV selection threshold, we compared all methods using the same liberal threshold of $p < 5 \times 10^{-5}$ under the main setting. We generate 200,000 independent SNPs to represent all underlying common variants and set $\sigma_x^2 = \sigma_y^2 = \sigma_u^2 = 1 \times 10^{-5}$, $\beta_{XU} = \beta_{YU} = 1$. We set $n_X = n_Y = 500,000$ to reflect the sample size of a typical GWAS in our real data analyses. We further set $\pi_1 + \pi_2 = 0.02$, $\pi_4 = 0.01$, and $\pi_5 = 0.97$. We let the proportion of invalid IVs, which is defined as $\pi_2/(\pi_1 + \pi_2)$ be equal to 50%. While some competing methods showed increased power, this often came at the cost of inflated Type I error rates and poor confidence interval coverage. CARE maintained its advantages in terms of bias, mean squared error, and valid inference. Figure S28 summarize the results for this setting.

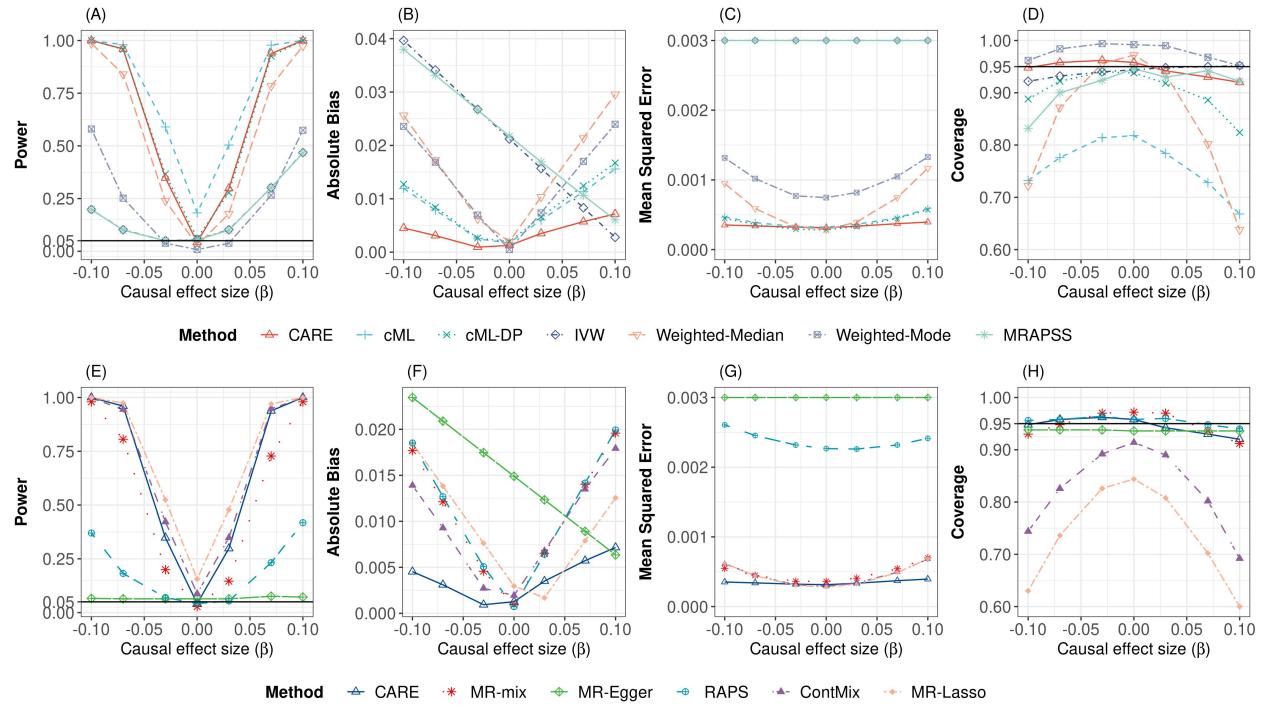


Figure S28: Power, absolute bias, mean squared error, and coverage of the CARE estimator and several robust MR methods under the main setting with 50% invalid IVs. The significant threshold is 5×10^{-5} for all methods. Power is the empirical power estimated by the proportion of p-values less than the significance threshold of 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

S.8.12 Comparison of l_0 and l_1 algorithms

We conduct a series of simulations to compare the performances of these two methods with the l_0 constraint approach adopted in this manuscript. Firstly, we varied the proportion of invalid IVs (30%, 50%). We also tested the performance under the setting of uniform distributed effects in correlated pleiotropy with 50% invalid IVs (See S.8.3 for the details of the setting).

We generate 200,000 independent SNPs to represent all underlying common variants and set $\sigma_x^2 = \sigma_y^2 = \sigma_u^2 = 1 \times 10^{-5}$, $\beta_{XU} = \beta_{YU} = 1$. We set $n_X = n_Y = 500,000$ to reflect the sample size of a typical GWAS in our real data analyses. We further set $\pi_1 + \pi_2 = 0.02$, $\pi_4 = 0.01$, and $\pi_5 = 0.97$. Figure S29 to S31 summarize the results for these settings. Our findings consistently demonstrate that while both approaches maintain comparable Type I error control, absolute bias, mean squared error (MSE), and coverage probability across various scenarios, the l_0 -based CARE method achieves noticeably higher statistical power.

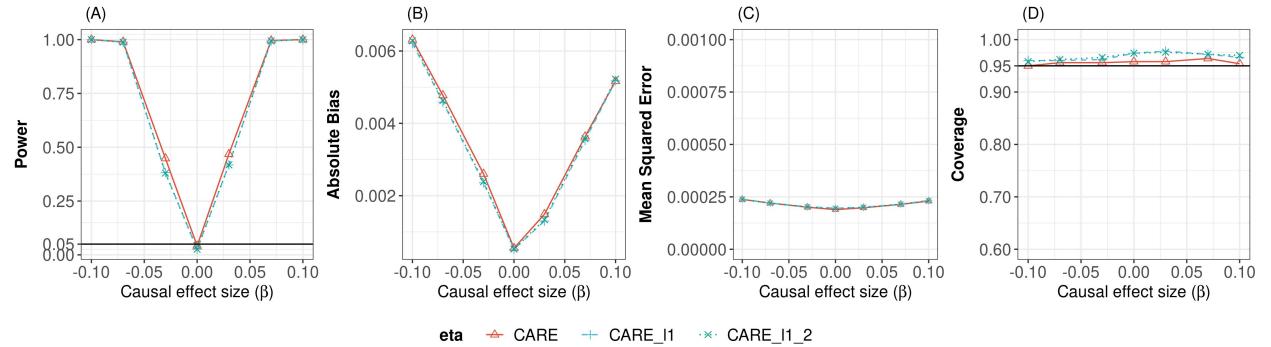


Figure S29: Power, absolute bias, mean squared error, and coverage of the CARE estimator with l_0 and two l_1 algorithms under the main setting with 30% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold of 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

S.8.13 Third sample for selecting IVs

Our method—CARE—effectively integrates winner’s curse correction via Rao-Blackwellization with robust handling of both measurement error and pleiotropy. However, in scenarios where the winner’s curse is no longer a concern—for example, when a third independent sample is available for IV selection based on association strength—some alternative methods may outperform CARE.

To investigate this, we conducted an additional simulation study using a three-sample MR

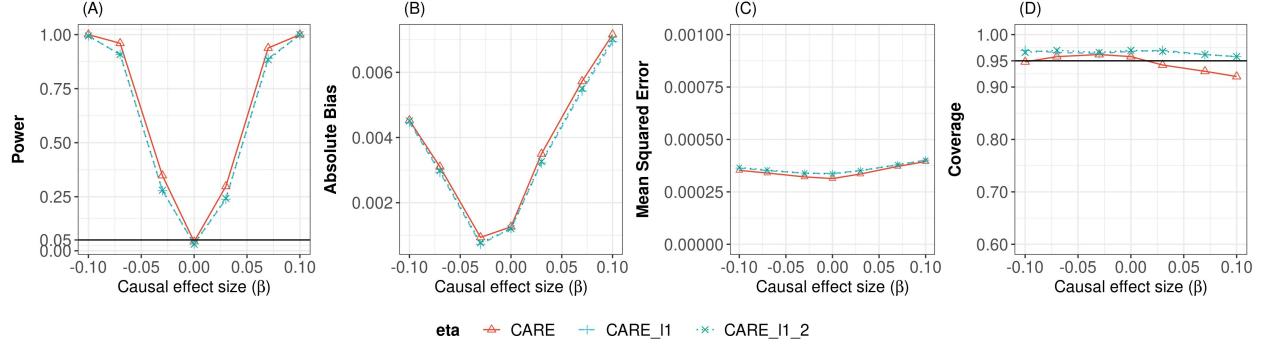


Figure S30: Power, absolute bias, mean squared error, and coverage of the CARE estimator with l_0 and two l_1 algorithms under the main setting with 50% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold of 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

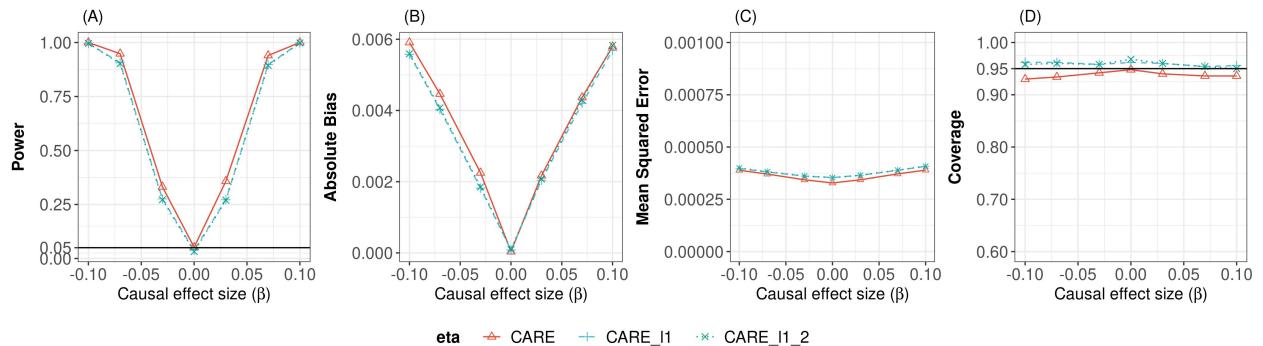


Figure S31: Power, absolute bias, mean squared error, and coverage of the CARE estimator with l_0 and two l_1 algorithms under the setting of uniform distributed effects in correlated pleiotropy with 50% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

design. The data generating process is the same as the main setting in our manuscript, which favors other methods with parametric assumptions (See details in Section S.8.1) In this design, a third independent sample is used exclusively for IV selection based on association strength, thereby eliminating the need for winner's curse correction in all methods. We uniformly apply a liberal IV selection threshold of $p < 5 \times 10^{-5}$ to this third sample across all methods for fair comparison.

As shown in Figure S32, cML outperforms CARE in terms of both power and mean squared error (MSE), while maintaining comparable empirical coverage. Other methods, such as cML-DP and IVW, also exhibit competitive performances. These results highlight that when a third sample is available and winner's curse correction is unnecessary, CARE may not be the optimal choice.

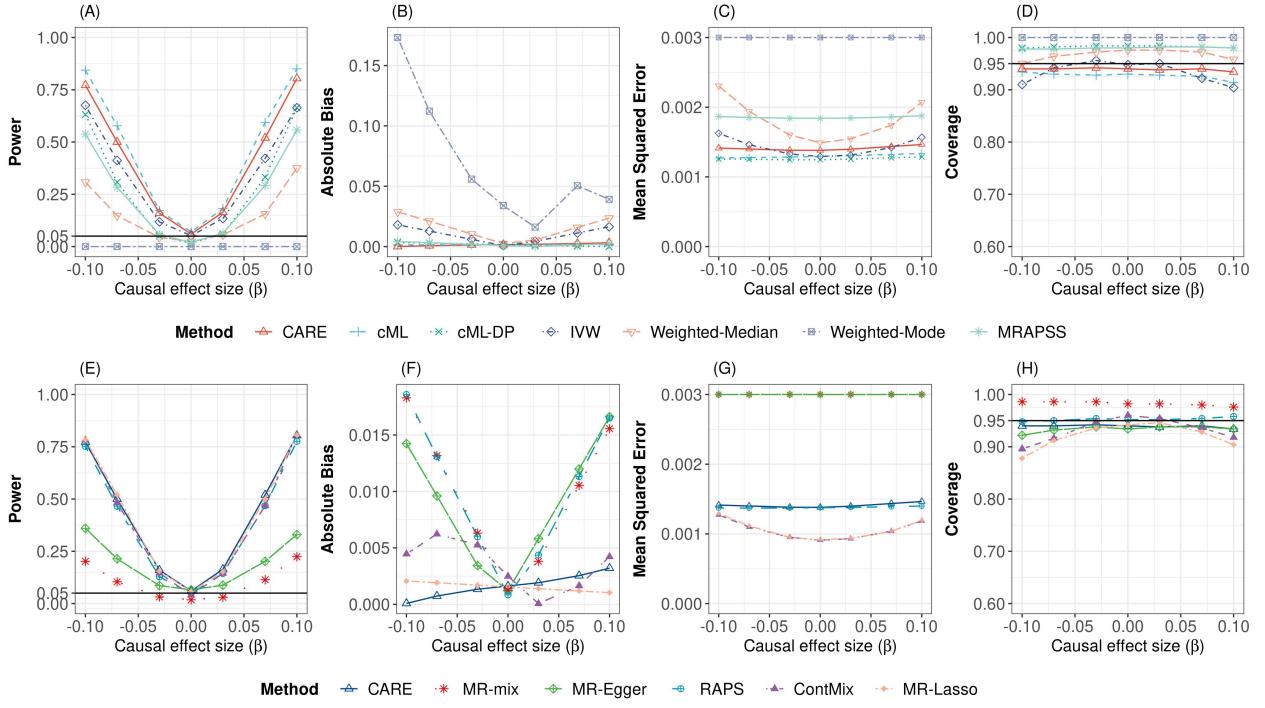


Figure S32: Power, absolute bias, mean squared error, and coverage of the CARE estimator and several robust MR methods under the main setting with 50% invalid IVs, all using a third sample for IV selection based on association strength. Power is the empirical power estimated by the proportion of p-values less than the significance threshold 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

S.9 Additional Real Data Results

S.9.1 Data harmonization

We harmonize GWAS summary data through the following steps. First, we exclude genetic variants that are not available in the outcome GWAS dataset. Second, we select independent genetic variants that have no linkage disequilibrium with other selected genetic variants. No linkage disequilibrium is defined as $R^2 < 0.001$ with an extension of 10,000 Kb in the genome, which has been widely adopted in applied MR studies [24]. For the benchmark methods, in line with the current practice [24], we employ standard clumping, selecting the variant with the smallest p -value of the SNP-exposure association when genetic variants are in linkage disequilibrium. For the proposed method CARE, we employ a revised sigma-based pruning procedure and select the variant with the smallest standard deviation of the SNP-exposure association when genetic variants are in linkage disequilibrium [33]. We employ this revised sigma-based pruning procedure because standard clumping introduces a different type of selection bias; see [41] for related discussion. Third, by leveraging allele frequency information, we infer the strand direction of ambiguous SNPs and harmonize exposure-outcome datasets using the `twosampleMR` package. We use the default setting with $\lambda = 4.06$ and $\eta = 0.5$ for our proposed CARE estimator, and set $\lambda = 4.06$ and $\lambda = 5.45$ for MR-APSS and other benchmark methods, respectively.

S.9.2 Comparative analysis of four MR methods for assessing COVID-19 severity

Second, we focus on four methods with relatively good performance under our negative control outcome analysis to alleviate the concerns of false positives. Figure S33 summarizes the results. First, CARE identifies body mass index (BMI), obesity class 1, obesity class 2, overweight, and extreme BMI are causally associated with COVID-19 severity. According to the Centers for Disease Control and Prevention (CDC), the risk of severe illness (i.e., hospitalization) from COVID-19 increases sharply with higher BMI, indicating that extreme BMI may be a likely causal risk factor for COVID-19 severity. Second, CARE identifies that HDL cholesterol (present in the blood, associated with a lower risk of coronary heart disease) is causally associated with COVID-19 severity. Low HDL level in the blood is reported to be associated with COVID-19 severity and most COVID-19

patients (65%) exhibit severely low HDL levels [36]. In comparison, the competing methods fail to identify HDL. Third, competing methods such as IVW and cML-DP identify childhood obesity, and celiac disease as causally associated with COVID-19 severity, while CARE does not. However, limited evidence supports their roles in COVID-19 severity as these risk factors are not listed on the CDC website, and hard to find support from the literature, suggesting that these two risk factors identified by competing methods may be false positives. Fourth, MR-APSS identifies the waist-to-hip ratio, which has been missed by the other three methods; however, the waist-to-hip ratio has been reported to have no association with COVID-19 severity [16].

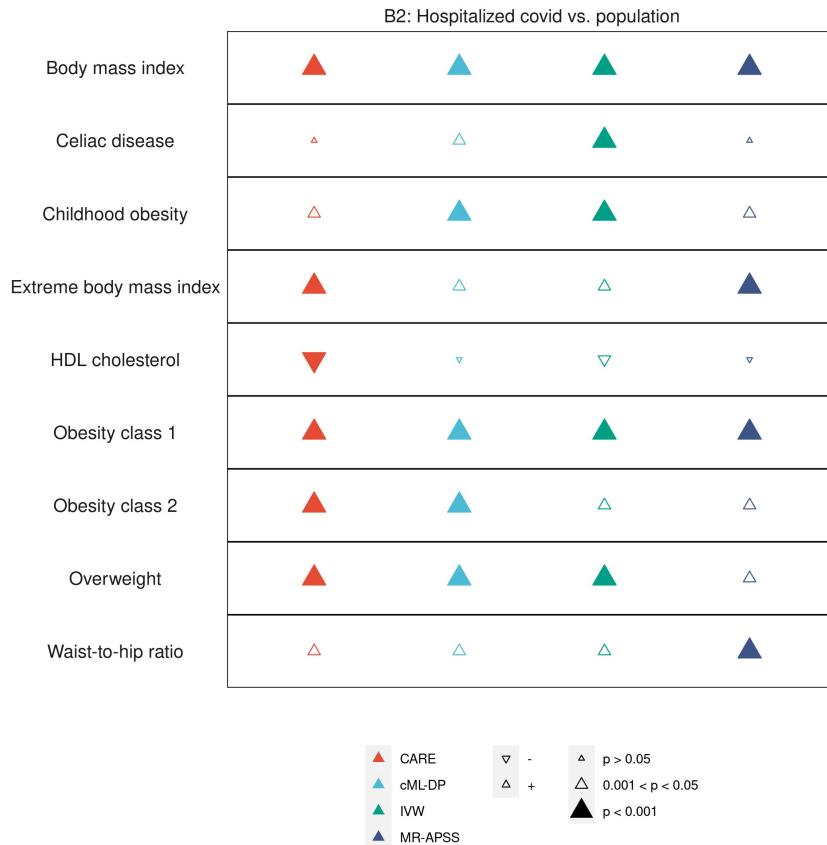


Figure S33: Significant causal exposure COVID-19 severity pairs identified by CARE, cML-DP, IVW, and MR-APSS. We summarize the significant causal exposure identified by at least one method under Bonferroni correction.

S.9.3 Supplementary tables and figures for real data results

GWAS ID	Trait	# SNP	N	PMID
ieu-a-298	Alzheimer's disease*	11,633	74,046	24162737
ieu-a-45	Anorexia nervosa	1,149,254	17,767	24514567
ieu-a-44	Asthma*	546,183	26,475	20860503
ieu-a-806	Autism*	9,499,590	10,263	23453885
ieu-a-801	Bipolar disorder*	2,427,221	16,731	21926972
ieu-a-29	Birth length	2,201,972	28,459	25281659
ieu-a-1083	Birth weight	16,245,524	143,677	27680694
ieu-a-2	Body Mass Index*	2,555,511	339,224	25673413
ieu-a-1109	Cardioembolic stroke	2,421,920	21,185	26935894
ieu-a-1058	Celiac disease	38,037	24,269	22057235
ieu-a-1096	Childhood obesity	2,442,739	13,848	22494627
ieu-a-1102	Chronic kidney disease*	2,191,877	117,165	26831199
ieu-a-7	Coronary heart disease*	9,455,779	123,504	26343387
ieu-a-12	Crohn's disease*	124,888	51,874	26192919
ieu-a-1000	Depressive symptoms*	6,524,475	161,460	27089181
ieu-a-1040	Difference in height between adolescence and adulthood	2,401,290	9,228	23449627
ieu-a-1037	Difference in height between childhood and adulthood	2,384,832	10,799	23449627
ieu-a-85	Extreme body mass index*	1,984,814	16,068	23563607
ieu-a-86	Extreme height	1,966,557	16,196	23563607
ieu-a-87	Extreme waist-to-hip ratio	1,939,901	10,255	23563607
ieu-a-1054	Gout	2,450,548	69,374	23263486
ieu-a-299	HDL cholesterol*	2,447,442	187,167	24097068
ieu-a-89	Height	2,550,859	253,288	25282103
ieu-a-31	inflammatory bowel disease*	12,716,084	34,652	26192919
ieu-a-814	Ischaemic stroke	393,465	517	17434096
ieu-a-300	LDL cholesterol	2,437,752	173,082	24097068
ieu-a-965	Lung adenocarcinoma*	8,881,354	18,336	24880342
ieu-a-966	Lung cancer*	8,945,893	27,209	24880342

GWAS ID	Trait	# SNP	N	PMID
ieu-a-1025	Multiple sclerosis*	156,632	38,589	24076602
ieu-a-798	Mycocardial infarction*	9,289,492	171,875	26343387
ieu-a-1007	Neuroticism	6,524,433	170,911	27089181
ieu-a-90	Obesity class 1*	2,380,428	98,697	23563607
ieu-a-91	Obesity class 2*	2,331,456	72,546	23563607
ieu-a-92	Obesity class 3*	2,250,779	50,364	23563607
ieu-a-93	Overweight*	2,435,045	158,855	23563607
ieu-a-975	Paget's disease	2,479,235	3,440	21623375
ieu-a-812	Parkinson's disease*	453,218	5,691	19915575
ieu-a-833	Rheumatoid arthritis*	9,739,304	80,799	24390342
ieu-a-22	Schizophrenia*	9,444,231	82,315	25056061
ieu-a-967	Squamous cell lung cancer	8,893,750	18,313	24880342
ieu-a-1009	Subjective well being	2,268,675	298,420	27089181
ieu-a-301	Total cholesterol	2,446,982	187,365	24097068
ieu-a-26	Type 2 diabetes*	2,473,442	69,033	22885922
ieu-a-970	Ulcerative colitis	156,116	47,745	26192919
ieu-a-72	Waist-to-hip ratio	2,562,516	224,459	25673412
ukb-b-553	Ease of skin tanning	9,851,867	453,065	
ukb-d-1747_1	Hair colour (natural, before greying): Blonde	13,586,531	360,270	
ukb-d-1747_2	Hair colour (natural, before greying): Red	13,586,531	360,270	
ukb-d-1747_3	Hair colour (natural, before greying): Light brown	13,586,531	360,270	
ukb-d-1747_4	Hair colour (natural, before greying): Dark brown	13,586,531	360,270	
ukb-d-1747_5	Hair colour (natural, before greying): Black	13,586,531	360,270	

Table 1: 45 exposures and six negative control outcomes included in the current study. GWAS ID, Trait, # SNP, N, and PMID stand GWAS ID used in IEU OpenGWAS database, exposure name, number of SNPs in the corresponding full GWAS summary data, sample size of the corresponding study, and PMID used in PubMed, respectively. Traits with stars represent those have been reported by the CDC or in peer-reviewed literature as risk factors for COVID-19 severity.

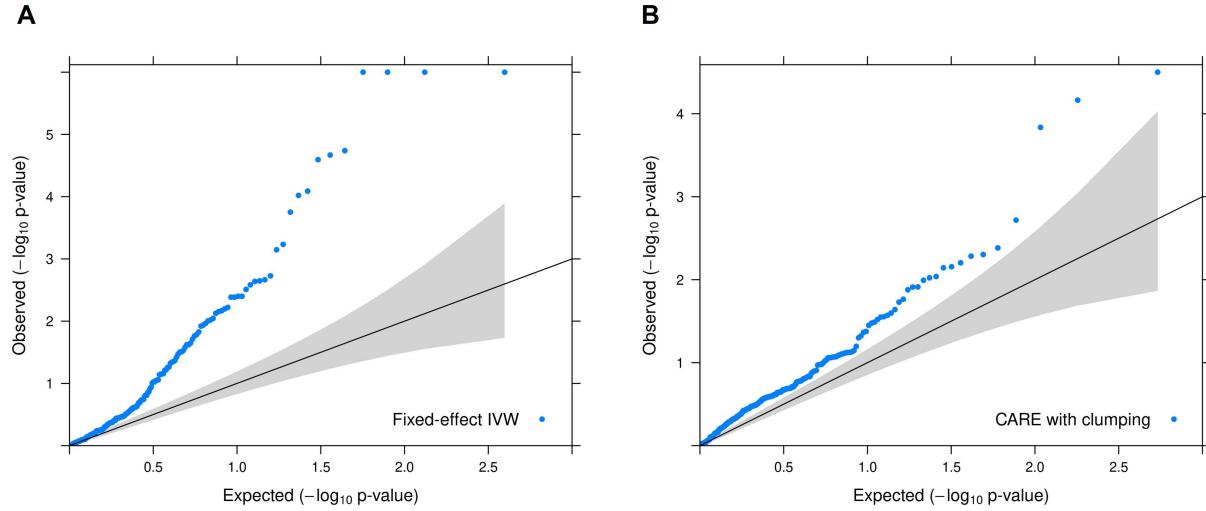


Figure S34: QQ plots of p-values in negative control outcome analysis for fixed-effect IVW (panel A) and CARE using clumping to select candidate IVs (panel B). The gray-shaded part is 95% confidence interval.

Condition	CARE			cML-DP			IVW			MR-APSS		
	β	SE	p-value									
Body mass index	0.3893	0.0595	5.96E-11	0.3952	0.0533	1.22E-13	0.4024	0.0580	4.11E-12	0.4006	0.1008	7.00E-05
Celiac disease	0.0213	0.0189	0.2603	0.0293	0.0089	0.0011	0.0299	0.0086	0.0005	0.0200	0.0166	0.2291
Childhood obesity	0.0749	0.0280	0.0074	0.0915	0.0226	5.49E-05	0.0946	0.0230	3.88E-05	0.0540	0.0231	0.0192
Extreme body mass index	0.0746	0.0194	0.0001	0.0561	0.0212	0.0081	0.0545	0.0191	0.0042	0.0622	0.0180	0.0005
HDL cholesterol	-0.1840	0.0509	0.0003	-0.0598	0.0315	0.0573	-0.0809	0.0359	0.0244	-0.1177	0.0836	0.1591
Obesity class 1	0.1916	0.0379	4.27E-07	0.1312	0.0254	2.47E-07	0.1288	0.0257	5.61E-07	0.1461	0.0307	2.01E-06
Obesity class 2	0.0924	0.0266	0.0005	0.0805	0.0222	0.0003	0.0793	0.0245	0.0012	0.0549	0.0203	0.0069
Overweight	0.2184	0.0602	0.0003	0.1475	0.0407	0.0003	0.1487	0.0443	0.0008	0.1621	0.0514	0.0016
Waist-to-hip ratio	0.3279	0.1139	0.0040	0.1980	0.0841	0.0186	0.2134	0.0842	0.0113	0.4114	0.0990	3.27E-05

Table 2: Association between significant exposure COVID-19 severity pairs using four methods: CARE, cML-DP, IVW, and MR-APSS. Values represent effect sizes (β), standard errors (SE), and p-values.