*Submitted to the Annals of Statistics*
arXiv: arXiv:0000.0000

# DEBIASED INVERSE-VARIANCE WEIGHTED ESTIMATOR IN TWO-SAMPLE SUMMARY-DATA MENDELIAN RANDOMIZATION*

By Ting Ye[†], Jun Shao[‡,§] and Hyunseung Kang[§]

*University of Pennsylvania*[†] *East China Normal University*[‡]
*University of Wisconsin-Madison*[§]

Mendelian randomization (MR) has become a popular approach to study the effect of a modifiable exposure on an outcome by using genetic variants as instrumental variables. A challenge in MR is that each genetic variant explains a relatively small proportion of variance in the exposure and there are many such variants, a setting known as many weak instruments. To this end, we provide a theoretical characterization of the statistical properties of two popular estimators in MR, the inverse-variance weighted (IVW) estimator and the IVW estimator with screened instruments using an independent selection dataset, under many weak instruments. We then propose a debiased IVW estimator, a simple modification of the IVW estimator, that is robust to many weak instruments and doesn't require screening. Additionally, we present two instrument selection methods to improve the efficiency of the new estimator when a selection dataset is available. An extension of the debiased IVW estimator to handle balanced horizontal pleiotropy is also discussed. We conclude by demonstrating our results in simulated and real datasets.

## 1. Introduction.

1.1. *Motivation: Many Weak Instruments in MR.* Instrumental variable (IV) is a well-known method to estimate the effect of a treatment, policy, or an exposure on an outcome in observational studies with unmeasured confounding [5, 24, 4]. Mendelian randomization (MR), a type of IV method, utilizes genetic variants as instruments to study the effect of a modifiable exposure or potential risk factor on an outcome in the presence of unmeasured confounding [18, 31, 26, 25, 11, 17, 42, 28]. A distinct feature of MR

*The research of Hyunseung Kang was supported in part by the U.S. National Science Foundation Grant DMS-1811414. The research of Jun Shao was partially supported by the National Natural Science Foundation of China Grant 11831008 and the U.S. National Science Foundation Grant DMS-1914411.

*MSC 2010 subject classifications:* Primary 62E30, 60K35; secondary 46N60, 62P10

*Keywords and phrases:* causal inference, inverse variance weighted estimator, many weak instruments, Mendelian randomization, summary data

is that there can be a large number of genetic variants, specifically single nucleotide polymorphisms (SNPs) from pre-existing large genome-wide association studies (GWASs), and many or possibly all SNPs are weak IVs; the setting is also referred to as many weak instruments in econometrics [15]. In particular, these genetic instruments/SNPs can be weak for the following three reasons. First, many SNPs may have zero/null effects on the exposure. Second, when SNPs are *common genetic variants*, i.e., their minor allele frequencies (MAF) are greater than 0.05 [20, 16], they may have small effects on the exposure. Third, when SNPs are *rare variants*, i.e., their MAF are less than 0.05, they may have small or modest effects on the exposure, but their genetic variances are small so that their total contribution to the variation of the exposure is small.

In this article, we focus on a popular setup in MR known as two-sample summary-data MR, where two sets of summary statistics are obtained from two GWASs [27]. The first set from one GWAS consists of $\hat{\gamma}_j$, the estimated marginal association between the $j$th SNP and the exposure, and its standard error (SE) $\hat{\sigma}_{Xj}$, $j = 1, ..., p$. The second set from another GWAS consists of $\hat{\Gamma}_j$, the estimated marginal association between the $j$th SNP and the outcome, and its SE $\hat{\sigma}_{Yj}$, $j = 1, ..., p$. In MR, the main parameter of interest is the exposure effect on the outcome, denoted as $\beta_0$, which can be estimated by $\hat{\beta}_j = \hat{\Gamma}_j/\hat{\gamma}_j$ for each $j$. However, $\hat{\beta}_j$ may be seriously biased and unstable when SNP $j$ is weak because $\hat{\gamma}_j$ is close to zero [30]. This leads to several modern MR methods that aggregate many possibly unstable estimators $\hat{\beta}_j$s using a meta-analysis strategy [10, 6, 7, 21]. The most popular among them is the inverse-variance weighted (IVW) estimator considered in [10],

$$\hat{\beta}_{\mathrm{IVW}} = \frac{\sum_{j=1}^{p} \hat{w}_j \hat{\beta}_j}{\sum_{j=1}^{p} \hat{w}_j}, \qquad \hat{\beta}_j = \frac{\hat{\Gamma}_j}{\hat{\gamma}_j}, \qquad \hat{w}_j = \frac{\hat{\gamma}_j^2}{\hat{\sigma}_{Yj}^2}. \tag{1.1}$$

A variant of the typical IVW estimator (1.1) is to only include SNPs that pass the genome-wide significance threshold in a third independent GWAS, known as the selection dataset, inside the IVW estimator (1.1); we call this the IVW estimator with screening and we formally define it in equation (2.3). Despite their popularity and widespread usage, very little is known about the theoretical properties of the IVW estimator, with or without screening. Specifically, in common MR setups where there are many weak IVs, it's unknown whether the IVW estimator $\hat{\beta}_{\mathrm{IVW}}$ in (1.1) or the screening counterpart are consistent or asymptotically normal.

1.2. *Prior Work and Our Contributions.* Prior work on weak IVs in MR is vast, but mostly limited to numerical studies [12, 13, 14, 27, 10]. In econometrics, the issue of weak IVs has been studied, but the results are limited to one-sample individual-data settings; see Stock, Wright and Yogo [34] and Andrews and Stock [3] for surveys. Recent papers by Zhao et al. [40, 41] and Bowden et al. [9] proposed new two-sample summary-data MR estimators that are robust to many weak IVs. Also, Wang and Kang [38] proposed new tests for two-sample summary-data MR when the number of instruments is fixed, but the instruments are arbitrarily weak. To the best of our knowledge, however, no work has addressed the theoretical properties of the IVW estimator in (1.1) or the IVW estimator with screening in (2.3), arguably the most popular estimators in MR, under a typical MR setting with many weak IVs.

Our overarching goal is to characterize the properties of the IVW estimators and to propose some improvements over them. The main contributions can be divided into four parts.

1. We provide an asymptotic phase transition analysis of the IVW estimator (1.1) in terms of IVs' average strength defined in (2.4). We conduct a similar exercise for the IVW estimator (2.3) with screening.
2. We propose a simple way to improve the IVW estimator (1.1) under many weak IVs, which we call the debiased IVW (dIVW) estimator. It is explicitly formulated as the IVW estimator multiplied by a bias correction factor; see equation (4.1). Unlike the IVW estimator, the dIVW estimator is robust to many weak IVs. In fact, even without screening for strong IVs, the dIVW estimator is generally consistent and asymptotically normal. As such, the dIVW estimator does not need a third independent GWAS to select instruments to mitigate the "winner's curse" bias [41, 10]. Finally, our dIVW estimators stand in contrast to recent optimization-based estimators (e.g., [40, 41]) that are robust to many weak IVs, but are arguably more complex than the dIVW estimators. As an example, the optimization-based estimator in [41] may not have unique estimates in every data generating scenario.
3. We depart from past theoretical studies in MR by considering the case where $\hat{\sigma}^2_{Xj}$ and $\hat{\sigma}^2_{Yj}$ are estimates, not respectively equal to $\sigma^2_{Xj}$ and $\sigma^2_{Yj}$, the true but unknown variances of $\hat{\gamma}_j$ and $\hat{\Gamma}_j$. We assess the impact of using estimated variances in the properties of the dIVW and the IVW estimators.
4. To improve the efficiency of the dIVW estimator, we propose two methods to select "efficiency-increasing" SNPs for the dIVW estimator, when a third GWAS dataset is available for screening. The first one is

straightforward and capable of eliminating IVs with no association to the exposure. The second one is data-driven and iteratively selects a threshold, leading to the most efficient estimator in a given class.

The rest of this paper is organized as follows. Section 2 introduces notation, setup, and assumptions. Section 3 characterizes the consistency and asymptotic normality of the IVW estimator and the IVW estimator with screening. Section 4 proposes the dIVW estimator and dIVW estimator with screening for improving efficiency, and establishes their asymptotic properties. Also included in Section 4 are two methods for selecting a threshold for screening, and an extension of the dIVW estimator to balanced horizontal pleiotropy [32, 36, 22]. Results from simulation studies and a real data analysis are presented in Sections 5 and 6, respectively. We conclude with a summary and discussion. Technical proofs and some additional results are in the Supplementary Material.

**2. Notation, Setup, and Assumption.** As part of a common data cleaning and pre-processing step in MR studies, millions of SNPs are decorrelated through linkage disequilibrium pruning or clumping via software [23]. We assume that this initial step produces $p$ independent SNPs, represented by bounded and mutually independent random variables $Z_1, ..., Z_p$.

Let $X$ be the exposure and $Y$ be the continuous outcome. Following the two-sample summary-data MR literature [27, 8, 40], we assume models

$$X = \sum_{j=1}^{p} \gamma_j Z_j + \eta_X U + E_X, \tag{2.1}$$

$$Y = \beta_0 X + \eta_Y U + E_Y, \tag{2.2}$$

where $\eta_X$, $\eta_Y$, $\beta_0$, $\gamma_1, ..., \gamma_p$ are unknown parameters, $U$ is an unmeasured confounder independent of $Z_1, ..., Z_p$, $E_X$ and $E_Y$ are mutually independent random noises that are also independent of $(Z_1, \ldots, Z_p, U)$, and $U$, $E_X$ and $E_Y$ have finite 4th order moments.

The goal in an MR analysis is to estimate the effect of the exposure $X$ on the outcome $Y$, which is represented by $\beta_0$. Since the unobserved $U$ is related with $X$, estimating $\beta_0$ using only model (2.2) with ordinary least squares leads to biased estimates. Instead, an MR approach to estimating $\beta_0$ typically assumes a model for $X$ like (2.1) and makes three core assumptions [18, 31, 26]. The first assumption is that instruments are associated with the exposure $X$, which amounts to $\gamma_j$'s in model (2.1) not simultaneously being zero. We call an instrument with $\gamma_j \neq 0$ to be a relevant or non-null IV and an instrument with $\gamma_j = 0$ to be an irrelevant or null IV. The

second assumption is that instruments are independent of the unmeasured confounder $U$; this is encoded by assuming $U$ is independent of $Z_1, \ldots, Z_p$ in (2.1)-(2.2). The third and last core assumption is that instruments affect the outcome $Y$ only through the exposure $X$; this is true under (2.1)-(2.2) since (2.2) does not involve $Z_j$'s. However, this last assumption may be violated in some studies; see Section 4.4 for one example based on balanced horizontal pleiotropy. For more detailed discussions on the core assumptions, models, and their implication in MR, see [19] and [8].

In classic IV settings, estimation of $\beta_0$ is based on $n$ independent and identically distributed (i.i.d.) observations of $(Z_1, \ldots, Z_p, X, Y)$. In two-sample MR, estimation is based on $n_X$ i.i.d. observations of $(X, Z_1, \ldots, Z_p)$ from the exposure dataset and $n_Y$ i.i.d. observations of $(Y, Z_1, \ldots, Z_p)$ from the outcome dataset. The two datasets are assumed to be independent of each other and we never jointly observe $Y$ and $X$.

In two-sample summary-data MR, which is the most popular data setting in MR and the setting considered in this paper, only summary statistics from the exposure and outcome datasets are available for analysis, not the individual-level data. Specifically, from the exposure dataset, we have $\hat{\gamma}_j$, the ordinary least square estimate from a linear regression of $X$ on $Z_j$, and its SE $\hat{\sigma}_{Xj}$, $j = 1, ..., p$. From the outcome dataset, we obtain $\hat{\Gamma}_j$, the ordinary least square estimate from a linear regression of $Y$ on $Z_j$, and its SE $\hat{\sigma}_{Yj}$, $j = 1, ..., p$. Note that models (2.1)-(2.2) and the independence of instruments imply that $\hat{\gamma}_j$ consistently estimates $\gamma_j$ and $\hat{\Gamma}_j$ consistently estimates $\beta_0\gamma_j$ for each $j$.

Many of the $p$ SNPs in (2.1), produced by the de-correlation step, could be potentially weak IVs with zero or small values of $\gamma_j^2 \mathrm{Var}(Z_j)$. It is therefore common in MR studies to screen IVs and include only selected IVs in the IVW estimator. To avoid selection bias or the "winner's curse", it is usually recommended to use a third independent dataset of size $n_{X^*}$ under model (2.1), called the selection dataset, solely for screening IVs [10, 41]. Typically, because only summary statistics are available from the selection dataset, thresholding is applied to screen out SNPs in (2.1) with the smallest marginal z-scores calculated from the summary statistics in the selection dataset. Future research will analyze a more sophisticated IV selection that simultaneously incorporates de-correlation and IV strength and its effects on estimation.

Formally, the IVW estimator with screening is a hard-thresholding esti-

mator with a z-score threshold $\lambda \geq 0$,

$$\hat{\beta}_{\lambda,\mathrm{IVW}} = \frac{\sum_{j\in S_\lambda} \hat{w}_j \hat{\beta}_j}{\sum_{j\in S_\lambda} \hat{w}_j} = \frac{\sum_{j\in S_\lambda} \hat{\Gamma}_j \hat{\gamma}_j \hat{\sigma}_{Yj}^{-2}}{\sum_{j\in S_\lambda} \hat{\gamma}_j^2 \hat{\sigma}_{Yj}^{-2}}, \quad S_\lambda = \{j : |\hat{\gamma}_j^*| > \lambda \hat{\sigma}_{Xj}^*\} \tag{2.3}$$

where $\hat{\gamma}_j^*$ and $\hat{\sigma}_{Xj}^*$ are counterparts of $\hat{\gamma}_j$ and $\hat{\sigma}_{Xj}$ computed from the selection dataset. If $\lambda = 0$, then $\hat{\beta}_{\lambda,\mathrm{IVW}}$ reduces to the original IVW estimator $\hat{\beta}_{\mathrm{IVW}}$ in (1.1). If $\lambda > 0$, only instruments with absolute value of z-scores higher than $\lambda$ are selected into the IVW estimator. A value for $\lambda$ that is used widely in MR is the genome-wide significance threshold $\lambda \approx 5.45$, which corresponds to screening out IVs whose p-values associated with $\hat{\gamma}_j$'s are above the genome-wide significance level $5 \times 10^{-8}$. More discussions about this genome-wide significance level are given in later sections.

We make the following two assumptions for our asymptotic analysis.

ASSUMPTION 1. *The sample sizes $n_X$ and $n_Y$ (and $n_{X^*}$ of the selection dataset if it exists) diverge to infinity with the same order. The number of SNPs, $p$, diverges to infinity.*

The conditions on sample sizes and $p$ are reasonable in our setup as many modern GWASs involve 10 to 100 thousands of participants and a few thousands of SNPs are typically found to be independent after the de-correlation pre-processing step.

The next assumption about summary statistics is also assumed in [40].

ASSUMPTION 2. *$\{\hat{\gamma}_j, \hat{\Gamma}_j, \hat{\gamma}_j^*, j = 1, ..., p\}$ are mutually independent and, for every $j$, $\hat{\gamma}_j \sim N(\gamma_j, \sigma_{Xj}^2)$, $\hat{\Gamma}_j \sim N(\beta_0 \gamma_j, \sigma_{Yj}^2)$, and $\hat{\gamma}_j^* \sim N(\gamma_j, \sigma_{Xj}^{*2})$. The variance ratios $\sigma_{Xj}^2/\sigma_{Yj}^2$ and $\sigma_{Xj}^2/\sigma_{Xj}^{*2}$ for all $j$ are bounded away from 0 and infinity.*

We briefly assess the plausibility of Assumption 2. With large sample sizes under Assumption 1, the normality of $\hat{\Gamma}_j$, $\hat{\gamma}_j$, and $\hat{\gamma}_j^*$ is plausible. The two-sample MR data structure guarantees the independence of $\hat{\gamma}_j$'s and $\hat{\Gamma}_j$'s (and $\hat{\gamma}_j^*$'s if they exist). Also, two-sample MR prunes/clumps SNPs to be far apart in genetic distance and each SNP only explains a very small proportion of the total variance in the exposure and outcome variables, making the independence within $\hat{\gamma}_j$'s, $\hat{\Gamma}_j$'s (and $\hat{\gamma}_j^*$'s if they exist) as well as the boundedness of variance ratios likely. Furthermore, if $Y$ is binary, Assumption 2 is a first-order local approximation of a logistic outcome model [40, 41].

We define the average strength of $p$ IVs as

$$\kappa = \frac{1}{p}\sum_{j=1}^{p}\frac{\gamma_j^2}{\sigma_{Xj}^2} \tag{2.4}$$

where $\gamma_j/\sigma_{Xj}$ is a normalized effect of SNP $j$ on $X$. If $\kappa$ is small, SNPs are, on average, weakly associated with the exposure. If $\kappa$ is large, SNPs are, on average, strongly associated with the exposure. We also define the average strength of IVs for IVW estimators with screening,

$$\kappa_\lambda = \frac{1}{p_\lambda}\sum_{j=1}^{p}\frac{\gamma_j^2}{\sigma_{Xj}^2}q_{\lambda,j}, \tag{2.5}$$

where $q_{\lambda,j} = P(|\hat{\gamma}_j^*| > \lambda\sigma_{Xj}^*)$ and $p_\lambda = \sum_{j=1}^{p} q_{\lambda,j}$. Clearly, if $\lambda = 0$ so that all the IVs are included in the IVW estimator, $q_{\lambda,j}$, $\kappa_\lambda$, and $p_\lambda$ become 1, $\kappa$, and $p$, respectively. As we will see in Sections 3-4, the limiting values of $\kappa$ and $\kappa_\lambda$ play a key role in characterizing the asymptotic properties of the IVW estimators.

The parameters $\kappa_\lambda$, $\kappa$, and $p_\lambda$ can be estimated by $\hat{\kappa}_\lambda$, $\hat{\kappa}$, and $\hat{p}_\lambda$, respectively, where

$$\hat{\kappa}_\lambda = \frac{1}{\hat{p}_\lambda}\sum_{j\in S_\lambda}\frac{\hat{\gamma}_j^2}{\hat{\sigma}_{Xj}^2} - 1, \qquad \hat{\kappa} = \hat{\kappa}_0, \qquad \hat{p}_\lambda = \text{the size of } S_\lambda. \tag{2.6}$$

We later show how to use these estimators in practice to check the theoretical conditions underlying the properties of the IVW estimator with or without screening.

**3. Properties of the IVW Estimators.** We study the consistency and asymptotic normality of the IVW estimators described in Sections 1-2 under different limiting values of $\kappa$ and $\kappa_\lambda$ defined in (2.4) and (2.5).

In the MR literature, it is common to assume that the standard deviations (SDs) $\sigma_{Xj}$, $\sigma_{Yj}$, and $\sigma_{Xj}^*$ (in Assumption 2) are known (e.g., [10, 6, 40, 41, 29]) so that $\hat{\sigma}_{Yj} = \sigma_{Yj}$ and $\hat{\sigma}_{Xj}^* = \sigma_{Xj}^*$ are used in (1.1) and (2.3). This is motivated by the fact that the sample sizes $n_X$, $n_Y$, and $n_X^*$ are usually very large in modern GWASs and the aforementioned references show empirically that such approximation works well in practice. In this section, we confine ourselves to the situation where the SDs are known and $\hat{\sigma}_{Yj} = \sigma_{Yj}$ and $\hat{\sigma}_{Xj}^* = \sigma_{Xj}^*$. The study of more general and realistic case where $\hat{\sigma}_{Yj} \neq \sigma_{Yj}$ and $\hat{\sigma}_{Xj}^* \neq \sigma_{Xj}^*$ is deferred to Section 4.

In what follows, $\xrightarrow{P}$ denotes convergence in probability and $\xrightarrow{D}$ denotes convergence in distribution.

THEOREM 3.1. *Assume models (2.1)-(2.2) and Assumptions 1-2. Also, assume that* $\hat{\sigma}_{Yj} = \sigma_{Yj}$ *and* $\hat{\sigma}^*_{Xj} = \sigma^*_{Xj}$ *in (1.1) and (2.3). When* $\beta_0 \neq 0$, *we have the following conclusions for either* $\lambda = 0$ *or* $\lambda > 0$.

*(a) If* $\kappa_\lambda/p_\lambda \to \infty$, $\max_j(\gamma_j^2\sigma_{Xj}^{-2}q_{\lambda,j})/(\kappa_\lambda p_\lambda) \to 0$, *and when* $\lambda \neq 0$, $\kappa_\lambda\sqrt{p_\lambda}/\lambda^2 \to \infty$, *then* $\hat{\beta}_{\lambda,\mathrm{IVW}}$ *is consistent and asymptotically normal, i.e.,*

$$V_{\lambda,\mathrm{IVW}}^{-1/2}\left(\hat{\beta}_{\lambda,\mathrm{IVW}} - \beta_0\right) \xrightarrow{D} N(0,1), \tag{3.1}$$

*where*

$$V_{\lambda,\mathrm{IVW}} = \frac{\sum_{j=1}^p[(w_j+v_j)q_{\lambda,j} + \beta_0^2 v_j(w_j+3v_j)q_{\lambda,j} - \beta_0^2 v_j^2 q_{\lambda,j}^2]}{[\sum_{j=1}^p(w_j+v_j)q_{\lambda,j}]^2},$$

$w_j = \gamma_j^2/\sigma_{Yj}^2$, *and* $v_j = \sigma_{Xj}^2/\sigma_{Yj}^2$, $j = 1, ..., p$.

*(b) If* $\kappa_\lambda \to \infty$ *and when* $\lambda \neq 0$, $\kappa_\lambda\sqrt{p_\lambda}/\lambda^2 \to \infty$, *then* $\hat{\beta}_{\lambda,\mathrm{IVW}} \xrightarrow{P} \beta_0$.

*(c) If* $\kappa_\lambda \to c > 0$ *and* $\sqrt{p_\lambda}/\max(1,\lambda^2) \to \infty$, *then*

$$\hat{\beta}_{\lambda,\mathrm{IVW}} - \beta_0\frac{\sum_{j=1}^p w_j q_{\lambda,j}}{\sum_{j=1}^p(w_j+v_j)q_{\lambda,j}} \xrightarrow{P} 0.$$

*(d) If* $\kappa_\lambda \to 0$ *and* $\sqrt{p_\lambda}/\max(1,\lambda^2) \to \infty$, *then* $\hat{\beta}_{\lambda,\mathrm{IVW}} \xrightarrow{P} 0$.

*When* $\beta_0 = 0$, *we have the following conclusion for either* $\lambda = 0$ *or* $\lambda > 0$.

*(e) If* $\max_j(\gamma_j^2\sigma_{Xj}^{-2}q_{\lambda,j})/(\kappa_\lambda p_\lambda + p_\lambda) \to 0$ *and when* $\lambda \neq 0$, $(\kappa_\lambda\sqrt{p_\lambda} + \sqrt{p_\lambda})/\max(1,\lambda^2) \to \infty$, *then* $\hat{\beta}_{\lambda,\mathrm{IVW}} \xrightarrow{P} 0$ *and* $V_{\lambda,\mathrm{IVW}}^{-1/2}\hat{\beta}_{\lambda,\mathrm{IVW}} \xrightarrow{D} N(0,1)$, *where* $V_{\lambda,\mathrm{IVW}}$ *is the same as that in (3.1) with* $\beta_0 = 0$.

We now elaborate the results in Theorem 3.1.

First, consider the case of $\lambda = 0$, i.e., the IVW estimator $\hat{\beta}_{\mathrm{IVW}}$ in (1.1) without screening. Part (a) of Theorem 3.1 is the only regime where $\hat{\beta}_{\mathrm{IVW}}$ is consistent and asymptotically normal when $\beta_0 \neq 0$. The main condition in Theorem 3.1(a) under $\lambda = 0$, $\kappa/p \to \infty$, means that the average IV strength $\kappa$ diverges to infinity at a rate faster than $p$, which is unlikely in MR studies. In fact, it is shown in the Supplementary Material that, approximately, $\kappa/p \leq n_X/p^2$ and thus, $\kappa/p \to \infty$ implies $n_X/p^2 \to \infty$. This rate is unrealistic in a typical MR study where the number of de-correlated SNPs $p$ is one thousand and, even when $n_X$ is as large as one million, we still have $n_X/p^2 = 1$. To explain why $\kappa/p \to \infty$ is needed for $\hat{\beta}_{\mathrm{IVW}}$ to be

asymptotically normal with mean $\beta_0$, consider

$$\hat{\beta}_{\mathrm{IVW}} - \beta_0 = \frac{\sum_{j=1}^{p}(\hat{\Gamma}_j\hat{\gamma}_j - \beta_0\hat{\gamma}_j^2)\sigma_{Yj}^{-2}}{\sum_{j=1}^{p}\hat{\gamma}_j^2\sigma_{Yj}^{-2}}$$

whose numerator and denominator have expectations $-\beta_0\sum_{j=1}^{p}v_j$ and $\sum_{j=1}^{p}(w_j+v_j)$, respectively, as $E(\hat{\Gamma}_j\hat{\gamma}_j) = E(\hat{\Gamma}_j)E(\hat{\gamma}_j) = \beta_0\gamma_j^2$ and $E(\hat{\gamma}_j^2) = \gamma_j^2 + \sigma_{Xj}^2$. Thus, as noticed by [40], the asymptotic bias of $\hat{\beta}_{\mathrm{IVW}}$ is

$$\mathrm{abias}(\hat{\beta}_{\mathrm{IVW}}) = \frac{-\beta_0\sum_{j=1}^{p}v_j}{\sum_{j=1}^{p}(w_j+v_j)}.$$

It is shown in the Supplementary Material that

$$V_{0,\mathrm{IVW}}^{-1/2}\{\hat{\beta}_{\mathrm{IVW}} - \beta_0 - \mathrm{abias}(\hat{\beta}_{\mathrm{IVW}})\} \xrightarrow{D} N(0,1),$$

where $V_{0,\mathrm{IVW}}$ is the asymptotic variance of $\hat{\beta}_{\mathrm{IVW}}$ given in (3.1) with $\lambda = 0$. When $\beta_0 \neq 0$, this means that $V_{0,\mathrm{IVW}}^{-1/2}(\hat{\beta}_{\mathrm{IVW}} - \beta_0) \xrightarrow{D} N(0,1)$ cannot hold unless

$$\frac{\mathrm{abias}(\hat{\beta}_{\mathrm{IVW}})}{V_{0,\mathrm{IVW}}^{1/2}} = \frac{-\beta_0\sum_{j=1}^{p}v_j}{[\sum_{j=1}^{p}\{(w_j+v_j)+\beta_0^2v_j(w_j+2v_j)\}]^{1/2}} \to 0 \tag{3.2}$$

i.e., the asymptotic bias of $\hat{\beta}_{\mathrm{IVW}}$ tends to 0 faster than the standard deviation of $\hat{\beta}_{\mathrm{IVW}}$. Since the numerator of the right side of (3.2) has order $p$ and the denominator of the right side has order $\{\max(p,\kappa p)\}^{1/2}$, (3.2) holds if $\kappa/p \to \infty$. We can easily construct an example in which (3.2) does not hold when $\kappa/p \not\to \infty$; in fact, the quantity in (3.2) diverges to infinity when $\kappa$ is bounded.

In short, our theory explains the numerical observations in the literature concerning poor normal approximation to $\hat{\beta}_{\mathrm{IVW}} - \beta_0$ in the presence of weak IVs.

Second, consider the case where $\lambda > 0$ in part (a) of Theorem 3.1. Screening with $\lambda > 0$ is a way to relax the condition required for the asymptotic normality of IVW estimator. Specifically, if $\lambda > 0$, $\kappa/p \to \infty$ in part (a) is replaced by $\kappa_\lambda/p_\lambda \to \infty$ and $\kappa_\lambda\sqrt{p_\lambda}/\lambda^2 \to \infty$. The Supplementary Material shows that $\kappa_\lambda$ is approximately increasing in $\lambda$ and thus $\kappa_\lambda/p_\lambda \to \infty$ is weaker than $\kappa/p \to \infty$. A similar analysis (Supplementary Material) shows that the counterpart of the asymptotic bias and standard deviation ratio in (3.2) for $\hat{\beta}_{\lambda,\mathrm{IVW}}$ is of the order $p_\lambda^{1/2}/(1+\kappa_\lambda)^{1/2}$, which tends to 0 as

$\kappa_\lambda/p_\lambda \to \infty$. In short, screening reduces the bias but increases the standard deviation of the IVW estimator (1.1), which is how $\hat{\beta}_{\lambda,\mathrm{IVW}}$ becomes consistent and asymptotically normal.

Third, if we forgo asymptotic normality, part (b) of Theorem 3.1 shows that the IVW estimator, with or without screening, is consistent for non-zero $\beta_0$ if the average IV strength $\kappa_\lambda$ diverges to infinity. Although this condition is weaker than $\kappa_\lambda/p_\lambda \to \infty$ in part (a) of Theorem 3.1, it is still unlikely to be satisfied in typical MR studies, not to mention that the consistency of IVW estimators is not enough for assessing variability or making statistical inference on $\beta_0$. To complement part (b), parts (c) and (d) of Theorem 3.1 show that if the average IV strength $\kappa_\lambda$ does not diverge to infinity, a common scenario in MR studies with many weak and null IVs, the IVW estimators are inconsistent and biased towards 0.

Finally, the last part (e) of Theorem 3.1 is for the special case of $\beta_0 = 0$, in which the weak IV bias of $\hat{\beta}_{\mathrm{IVW}}$ is not an issue because the asymptotic bias in (3.2) equals zero when $\beta_0 = 0$ and $\hat{\beta}_{\mathrm{IVW}}$ is consistent and asymptotically normal under reasonable conditions.

Result (3.1) still holds if we replace $V_{\lambda,\mathrm{IVW}}$ by a plug-in consistent estimator

$$\hat{V}_{\lambda,\mathrm{IVW}} = \frac{\sum_{j\in S_\lambda}[\hat{w}_j + \hat{\beta}^2_{\lambda,\mathrm{IVW}}\hat{v}_j(\hat{w}_j + \hat{v}_j)]}{(\sum_{j\in S_\lambda}\hat{w}_j)^2}, \tag{3.3}$$

where $\hat{w}_j = \hat{\gamma}_j^2/\hat{\sigma}^2_{Yj}$, $\hat{v}_j = \hat{\sigma}^2_{Xj}/\hat{\sigma}^2_{Yj}$, and $S_\lambda = \{j : |\hat{\gamma}^*_j| > \lambda\hat{\sigma}^*_{Xj}\}$.

Comparing the IVW estimator (1.1) with the IVW estimator (2.3) with screening, the former requires far more stringent conditions on IV strength to guarantee its consistency or asymptotic normality, whereas the latter requires finding a threshold $\lambda$ and checking whether $\lambda$ satisfies conditions in Theorem 3.1(a), which is cumbersome and not always successful. We highlight some examples below.

1. Consider the common practice of selecting IVs that pass the p-value threshold of $5\times10^{-8}$, which as mentioned earlier is equivalent to setting $\lambda \approx 5.45$. This $\lambda$ may or may not satisfy the conditions for consistency or asymptotic normality of $\hat{\beta}_{\lambda,\mathrm{IVW}}$ in Theorem 3.1(a)-(b).
2. If all IVs are very weak, for example $\gamma_j^2 \le c\sigma^2_{Xj}$ for all $j$ and some positive constant $c$, then $\kappa_\lambda$ is bounded regardless of the choice of $\lambda$.
3. If every IV strength equals $\lambda$, then $q_{\lambda,j} \approx 1/2$ and $\kappa_\lambda \approx \lambda^2$. But, $\kappa_\lambda/p_\lambda \approx 2\lambda^2/p$ may be small, implying that the asymptotic normality in Theorem 3.1(a) may not hold.

## 4. Debiased IVW Estimators.

4.1. *Debiased IVW Estimator.* Motivated by the stringent assumptions underlying the asymptotic normality of the IVW estimator without screening and the need of a third dataset and a carefully chosen threshold $\lambda$ in the IVW estimator with screening, we propose a simple estimator of $\beta_0$ that relies on neither. We name the new estimator as the debiased IVW (dIVW) estimator. It is the original IVW estimator multiplied by an explicit bias correction factor, i.e.,

$$\hat{\beta}_{\text{dIVW}} = \hat{\beta}_{\text{IVW}} \cdot \frac{\sum_{j=1}^{p} \hat{w}_j}{\sum_{j=1}^{p} (\hat{w}_j - \hat{v}_j)} = \frac{\sum_{j=1}^{p} \hat{\Gamma}_j \hat{\gamma}_j \hat{\sigma}_{Yj}^{-2}}{\sum_{j=1}^{p} (\hat{\gamma}_j^2 - \hat{\sigma}_{Xj}^2) \hat{\sigma}_{Yj}^{-2}}, \tag{4.1}$$

where $\hat{w}_j = \hat{\gamma}_j^2 / \hat{\sigma}_{Yj}^2$ and $\hat{v}_j = \hat{\sigma}_{Xj}^2 / \hat{\sigma}_{Yj}^2$. The bias correction factor amplifies the IVW estimator that is biased towards 0 according to Theorem 3.1(c)-(d).

Surprisingly, this simple correction makes the resulting estimator dramatically more robust to many weak IVs. To explain why, recall that the asymptotic normality of $\hat{\beta}_{\text{IVW}}$ requires that its asymptotic bias tend to 0 fast enough, i.e., (3.2) or the stringent condition $\kappa/p \to \infty$ holds. For the dIVW estimator,

$$\hat{\beta}_{\text{dIVW}} - \beta_0 = \frac{\sum_{j=1}^{p} (\hat{\Gamma}_j \hat{\gamma}_j - \beta_0 \hat{\gamma}_j^2 + \beta_0 \hat{\sigma}_{Xj}^2) \hat{\sigma}_{Yj}^{-2}}{\sum_{j=1}^{p} (\hat{\gamma}_j^2 - \hat{\sigma}_{Xj}^2) \hat{\sigma}_{Yj}^{-2}},$$

the numerator has mean zero under Assumption 2. This indicates that, $\hat{\beta}_{\text{dIVW}}$ has a negligible asymptotic bias, and hence its asymptotic normality does not need a stringent condition such as $\kappa/p \to \infty$ to ensure (3.2).

As we show in the next section, the dIVW estimator (4.1) is consistent and asymptotically normal if $\kappa\sqrt{p} \to \infty$ and $\max_j(\gamma_j^2 \sigma_{Xj}^{-2})/(\kappa p + p) \to 0$. The condition $\kappa\sqrt{p} \to \infty$ is considerably weaker than $\kappa/p \to \infty$ required by the original IVW estimator (1.1). For example, $\kappa\sqrt{p} \to \infty$ holds even when $\kappa \to 0$ but at a slower rate than $1/\sqrt{p}$; in contrast, $\kappa/p \to \infty$ requires $\kappa \to \infty$. Also, when IVs are common variants, but are weak in the sense of Staiger and Stock [33] (i.e., $\gamma_j$ and $\sigma_{Xj}$ are both of the order $n_X^{-1/2}$), the dIVW estimator still remains consistent and asymptotically normal if the number of such weak IVs is large. Finally, the condition $\kappa\sqrt{p} \to \infty$ is also related to conditions imposed by the limited information maximum likelihood (LIML) estimator in the one-sample individual-level data setting [15] and the robust adjusted profile score (MR-raps) estimator [40].

The quantity $\kappa\sqrt{p}$ can be interpreted as an effective sample size for the dIVW estimator and can be estimated by $\hat{\kappa}\sqrt{p}$ with $\hat{\kappa}$ defined in (2.6). In

our simulation studies (i.e., Figure 1), we provide some guidelines on what would be considered a large value of $\hat{\kappa}\sqrt{p}$ for the asymptotics promised to kick in. This is akin to qualitative guidelines on what would be a large enough sample size for a normal approximation of an estimator to hold.

4.2. *Improving Efficiency With Screening.* While the dIVW estimator $\hat{\beta}_{\text{dIVW}}$ (4.1) without screening is consistent and asymptotically normal even if many IVs are weak, its asymptotic variance, $V_{0,\text{dIVW}}$ defined in (4.3) with $\lambda = 0$, is larger than $V_{0,\text{IVW}}$, the asymptotic variance of the IVW estimator $\hat{\beta}_{\text{IVW}}$ (1.1) with $\lambda = 0$; we remark that both $V_{0,\text{IVW}}$ and $V_{0,\text{dIVW}}$ have order $(\kappa p)^{-1}$ when $\hat{\beta}_{\text{dIVW}}$ and $\hat{\beta}_{\text{IVW}}$ are asymptotically normal. The increase in variance of the dIVW estimator is due to a bias-variance trade off between the IVW estimator and the dIVW estimator and the bias due to weak IVs in MR studies tends to dominate the SD of the estimator.

When summary statistics from an independent selection dataset are available, we explore how to make the dIVW estimator more efficient by screening. We remark here that screening in the dIVW estimator is solely for efficiency improvement since $\hat{\beta}_{\text{dIVW}}$ without screening remains asymptotically normal under weak conditions. In contrast, the IVW estimator uses screening to reduce bias and to achieve asymptotic normality.

Formally, consider the the dIVW estimator using only IVs selected from the selection dataset,

$$\hat{\beta}_{\lambda,\text{dIVW}} = \frac{\sum_{j\in S_\lambda} \hat{\Gamma}_j \hat{\gamma}_j \hat{\sigma}_{Yj}^{-2}}{\sum_{j\in S_\lambda} (\hat{\gamma}_j^2 - \hat{\sigma}_{Xj}^2)\hat{\sigma}_{Yj}^{-2}}, \qquad S_\lambda = \{j : |\hat{\gamma}_j^*| > \lambda \hat{\sigma}_{Xj}^*\} \tag{4.2}$$

Theorem 4.1 establishes the asymptotic normality of $\hat{\beta}_{\lambda,\text{dIVW}}$ in (4.2). The Theorem includes $\hat{\beta}_{\text{dIVW}}$ in (4.1) as a special case of $\hat{\beta}_{\lambda,\text{dIVW}}$ when $\lambda = 0$.

THEOREM 4.1. *Assume models (2.1)-(2.2), Assumptions 1-2, and that* $\kappa_\lambda\sqrt{p_\lambda}/\max(1,\lambda^2) \to \infty$ *and* $\max_j(\gamma_j^2\sigma_{Xj}^{-2}q_{\lambda,j})/(\kappa_\lambda p_\lambda + p_\lambda) \to 0$. *Assume further that either* $\hat{\sigma}_{Xj} = \sigma_{Xj}$, $\hat{\sigma}_{Yj} = \sigma_{Yj}$ *and* $\hat{\sigma}_{Xj}^* = \sigma_{Xj}^*$ *in (4.1)-(4.2) or* $p/n_X \to 0$. *Then,* $\hat{\beta}_{\lambda,\text{dIVW}}$ *is consistent and asymptotically normal, i.e.,*

$$V_{\lambda,\text{dIVW}}^{-1/2}(\hat{\beta}_{\lambda,\text{dIVW}} - \beta_0) \xrightarrow{D} N(0,1)$$

*where*

$$V_{\lambda,\text{dIVW}} = \frac{\sum_{j=1}^p [(w_j + v_j) + \beta_0^2 v_j(w_j + 2v_j)]q_{\lambda,j}}{(\sum_{j=1}^p w_j q_{\lambda,j})^2} \tag{4.3}$$

*for* $\lambda = 0$ *or* $\lambda > 0$, $w_j = \gamma_j^2/\sigma_{Yj}^2$, *and* $v_j = \sigma_{Xj}^2/\sigma_{Yj}^2$, $j = 1, ..., p$.

For consistency and asymptotic normality, the stringent condition $\kappa_\lambda/p_\lambda \to \infty$ required in Theorem 3.1(a) for IVW estimators is not needed in Theorem 4.1.

Theorem 4.1 shows that, if $p/n_X \to 0$, then using the estimates SEs leads to the same asymptotic result as using the true SDs in (4.1)-(4.2). A parallel result can also be established for the IVW estimators but is omitted here. Thus, our result provides a theoretical justification for safely treating SEs as SDs, a commonly adopted approach in MR studies.

The condition $p/n_X \to 0$ typically holds since after de-correlation, the number of independent SNP is usually around a few thousands and the sample size is around 10 to 100 thousands. Our simulation results in Section 5 shows that the approximation is still very good even when $p/n_X$ is 20%, indicating that $p/n_X \to 0$ is only sufficient rather than necessary.

The quantity $\kappa_\lambda\sqrt{p_\lambda}/\max(1,\lambda^2)$ acts like an effective sample size for the dIVW estimator with screening and can be estimated by $\hat\kappa_\lambda\sqrt{\hat p_\lambda}/\max(1,\lambda^2)$ with $\hat\kappa_\lambda$ and $\hat p_\lambda$ given by (2.6). In our simulation studies, specifically Figure 1, we provide some guidelines on what would be considered a large effective sample size for the asymptotic result to kick in.

The results in Theorem 4.1 still hold if we replace the asymptotic variance $V_{\lambda,\mathrm{dIVW}}$ with a consistent estimator

$$\hat V_{\lambda,\mathrm{dIVW}} = \frac{\sum_{j\in S_\lambda}[\hat w_j + \hat\beta^2_{\lambda,\mathrm{dIVW}}\hat v_j(\hat w_j+\hat v_j)]}{[\sum_{j\in S_\lambda}(\hat w_j - \hat v_j)]^2}, \tag{4.4}$$

where $\hat w_j = \hat\gamma_j^2/\hat\sigma^2_{Yj}$, $\hat v_i = \hat\sigma^2_{Xj}/\hat\sigma^2_{Yj}$, and $S_\lambda = \{j : |\hat\gamma^*_j| > \lambda\hat\sigma^*_{Xj}\}$.

4.3. *Choice of $\lambda$ in Screening.* We consider the choice of $\lambda$ in the dIVW estimator (4.2) with screening. In general, the threshold $\lambda$ should satisfy $\kappa_\lambda\sqrt{p_\lambda}/\max(1,\lambda^2) \to \infty$, as well as increase the efficiency of the dIVW estimator.

One choice is $\lambda = \sqrt{2\log p}$ that diverges to infinity at a very slow rate. This $\lambda$ guarantees that the probability of selecting any null IV is very small, because under Assumptions 1-2 and $p \to \infty$,

$$P(\text{at least one null IV is selected}) \le \frac{2(p-s)}{\lambda\sqrt{2\pi}}e^{-\lambda^2/2} = \frac{p-s}{p\sqrt{\pi\log p}} \to 0$$

where $s$ is the number of non-null IVs.

Another choice of $\lambda$ is motivated by directly studying the asymptotic variance $V_{\lambda,\mathrm{dIVW}}$ in (4.3), which has order $(\kappa_\lambda p_\lambda)^{-1}$ when $\kappa_\lambda \not\to 0$ and $(\kappa^2_\lambda p_\lambda)^{-1}$ when $\kappa_\lambda \to 0$. To illustrate the idea, we focus on the situation

where $\kappa \not\to 0$ so that the asymptotic variances of $\hat{\beta}_{\text{dIVW}}$ and $\hat{\beta}_{\lambda,\text{dIVW}}$ have orders $(\kappa p)^{-1}$ and $(\kappa_\lambda p_\lambda)^{-1}$, respectively. Since $\kappa_\lambda p_\lambda \le \kappa p$ for any $\lambda > 0$, to screen for efficiency rather for relevant IV selection, we should not screen out too many non-null IVs (even if they are weak) to result in $\kappa_\lambda p_\lambda / \kappa p \to 0$. From this point of view, $\lambda = \sqrt{2 \log p}$ is an improvement over the genome-wide significance p-value threshold $5 \times 10^{-8}$ ($\lambda \approx 5.45$) because $\sqrt{2 \log p} < 5.45$ if $p < 10^6$, and using $\lambda = \sqrt{2 \log p}$ eliminates null IVs with probability close to 1 as the previous discussion indicated. But, if there are many weak IVs, $\hat{\beta}_{\text{dIVW}}$ may be asymptotically more efficient than $\hat{\beta}_{\lambda,\text{dIVW}}$ with any $\lambda > 0$ (see Case 3 of the simulation study in Section 5.1). In short, it is better if we can select $\lambda$ adaptively.

This leads to our approach of choosing $\lambda$ that directly minimizes an estimated asymptotic variance of $\hat{\beta}_{\lambda,\text{dIVW}}$, which we call the Mendelian Randomization Estimation-Optimization (MR-EO) algorithm. In a nutshell, MR-EO considers $\hat{\beta}_{\lambda,\text{dIVW}}$ with $\lambda$ that varies in the interval $[0, \sqrt{2 \log p}]$; it assumes that $\kappa_\lambda \sqrt{p_\lambda} / \max(1, \lambda^2) \to \infty$ holds for every $\lambda$ in the range. It then tries to find the $\lambda$ in this range that minimizes the asymptotic variance. Since we cannot directly use estimated variance in (4.4) because $\hat{\beta}_{\lambda,\text{dIVW}}$ is not available prior to the selection of $\lambda$, MR-EO alternates between estimating the exposure effect by $\hat{\beta}_t$ (i.e., the E-Step) and finding the optimal $\lambda$ given the previous estimate $\hat{\beta}_t$ (i.e., the O-Step); see Algorithm 1 for details.

Initialize $t = 0$, $t_{\max}$, $\lambda_0 = \sqrt{2 \log p}, V = \infty$;
**while** $t \le t_{\max}$ **do**
  E-Step: for a given $\lambda_t$, estimate $\beta_0$ with the dIVW estimator $\hat{\beta}_{\lambda_t,\text{dIVW}}$;
  **if** $V \le \hat{V}_{\lambda_t,\text{dIVW}}(\hat{\beta}_{\lambda_t,\text{dIVW}})$ **then**
    exit the while loop;
  **else**
    $V = \hat{V}_{\lambda_t,\text{dIVW}}(\hat{\beta}_{\lambda_t,\text{dIVW}})$;
  **end**
  O-Step: Plug $\hat{\beta}_{\lambda_t,\text{dIVW}}$ into the variance estimator and find

$$\lambda_{t+1} = \arg \min_{\lambda \in [0, \sqrt{2 \log p}]} \hat{V}_{\lambda,\text{dIVW}}(\hat{\beta}_{\lambda_t,\text{dIVW}})$$

  Set $t = t + 1$;
**end**
Output $\lambda_{t-1}$.

**Algorithm 1:** MR-EO algorithm to determine the optimal $\lambda$

We make some comments regarding the implementation of MR-EO and its final output. First, we initialize MR-EO to $\lambda_0 = \sqrt{2 \log p}$ and force the algorithm to stop at $t = t_{\max}$ with a reasonably large $t_{\max}$, mainly for computa-

tional efficiency. Second, the algorithm assumes that $\kappa_\lambda\sqrt{p_\lambda}/\max(1,\lambda^2) \to \infty$ holds for every $\lambda$ in $[0,\sqrt{2\log p}\,]$ so that $\hat{\beta}_{\lambda_t,\mathrm{dIVW}}$ is consistent for $\beta_0$. To verify this, we can empirically evaluate $\hat{\kappa}_\lambda\sqrt{\hat{p}_\lambda}/\max(1,\lambda^2)$ and check whether this quantity is reasonably large for all $\lambda$ in the range. In our simulation study in Section 5, we find that the range $[0,\sqrt{2\log p}]$ works well. Third, with fixed $t_{\max}$ and range of $\lambda$, MR-EO produces a unique dIVW estimator. Finally, the estimated variance for the dIVW estimator based on MR-EO may be too optimistic due to the "winner's curse"; however, when both $p$ and $n_X$ are large and the ratio $p/n_X$ is bounded, ideally small, this issue will be largely moot. In our simulation studies in Section 5, we find that the estimator chosen by MR-EO performs well and the resulting confidence interval maintains nominal coverage, although Theorem 4.1 does not directly guarantee that the estimator chosen by MR-EO is asymptotically normal.

4.4. *Extension to Balanced Horizontal Pleiotropy.* We extend the dIVW estimator to situations under one type of pleiotropy in MR, balanced horizontal pleiotropy [22, 40, 8]. Briefly, under balanced horizontal pleiotropy, the third core IV assumption described in Section 2 is violated and the model (2.2) is extended to

$$Y = \beta_0 X + \sum_{j=1}^{p} \alpha_j Z_j + \eta_Y U + E_Y, \tag{4.5}$$

where the pleiotropic effects of $p$ SNPs on $Y$, $\alpha_1,...,\alpha_p$, $\alpha_j \sim N(0,\tau_0^2)$, are independent random effects and independent of $X$, $Z_j$'s, $U$, $E_Y$ and $E_X$. To incorporate balanced pleiotropy, we replace Assumption 2 with the following assumption [22, 40, 41].

ASSUMPTION 2′. *Suppose Assumption 2 holds except conditional on* $\alpha_j$, $\hat{\Gamma}_j \sim N(\alpha_j + \beta_0\gamma_j, \sigma_{Yj}^2)$ *for every* $j$. *In addition, for some constant* $c_+$, $\tau_0 \le c_+\sigma_{Yj}$ *for all* $j$.

Under the same conditions in Theorem 4.1 with (2.2) and Assumption 2 replaced by (4.5) and Assumption 2′, respectively, the Supplementary Material shows that the dIVW estimators in (4.1) and (4.2) are still consistent and asymptotically normal. However, the variance of the dIVW estimators is larger due to the random effects $\alpha_j$'s and an estimator of it under balanced pleiotropy is

$$\frac{\sum_{j\in S_\lambda}[\hat{w}_j(1+\hat{\tau}^2\hat{\sigma}_{Yj}^{-2}) + \hat{\beta}_{\lambda,\mathrm{dIVW}}^2\hat{v}_j(\hat{w}_j+\hat{v}_j)]}{[\sum_{j\in S_\lambda}(\hat{w}_j-\hat{v}_j)]^2}, \tag{4.6}$$

where

$$\hat{\tau}^2 = \frac{\sum_{j=1}^{p}[(\hat{\Gamma}_j - \hat{\beta}_{\mathrm{dIVW}}\hat{\gamma}_j)^2 - \hat{\sigma}_{Yj}^2 - \hat{\beta}_{\mathrm{dIVW}}^2\hat{\sigma}_{Xj}^2]\hat{\sigma}_{Yj}^{-2}}{\sum_{j=1}^{p}\hat{\sigma}_{Yj}^{-2}}.$$

We remark that the estimator of $\hat{\tau}^2$ relies on $\hat{\beta}_{\mathrm{dIVW}}$. Also, if $\max_k \sigma_{Yk}^{-2}$ is bounded by a constant times the average $p^{-1}\sum_{j=1}^{p}\sigma_{Yj}^{-2}$, as $\kappa\sqrt{p} \to \infty$, the variance estimator in (4.6) is consistent. We can use the aforementioned methods (e.g., MR-EO) to choose $\lambda$ and improve efficiency.

Finally, when balanced horizontal pleiotropy does not hold, the dIVW estimator, like other MR estimators built upon this assumption, will be biased. In Section 2 of the Supplementary Material, we show that the dIVW estimator is biased but still asymptotically normal, and we investigate the magnitude of this bias.

## 5. Simulation Studies.

5.1. *A Simulation with the BMI-CAD Dataset as Population.* We conduct a simulation study to compare the finite sample properties of several estimators under different screening thresholds. To closely mirror what is done in practice, we adopt a real two-sample summary-level MR dataset, the BMI-CAD dataset in the *mr.raps* R package (version 0.3.1) of Zhao et al. [41], as the simulation population. The BMI-CAD dataset is used to make inference about the effect of $X$, the body mass index (BMI), on $Y$, the risk of coronary artery disease (CAD). It contains three independent datasets:

1. Exposure dataset: A GWAS for BMI in round 2 of the UK BioBank (sample size: 336,107) [1];
2. Outcome dataset: A GWAS for CAD from the CARDIoGRAMplusC4D consortium (sample size: ≈185,000), with genotype imputation using the 1000 Genome Project [35];
3. Selection dataset: A GWAS for BMI in the Japanese population (sample size: 173,430) [2].

The three datasets have been cleaned so that (i) SNPs appear in all three datasets and (ii) SNPs are far apart in genetic distance; see [41] for details. The initial data cleaning leads to $p = 1119$ SNPs available for analysis. Each GWAS contains publicly available summary statistics that are the estimated coefficients from marginal linear regression and their SEs. We use them as population parameters in our simulation; in Section 6, we use them as data.

To begin, we construct three plausible sets of $\gamma_j$ as follows.

**Case 1** (Some strong IVs, many null IVs): There are $s = 20$ non-null IVs whose $\gamma_j$-values are the 20 marginal regression coefficients with the smallest p-values in the BMI-CAD exposure dataset. The rest $p - s = 1099$ SNPs are null IVs with zero $\gamma_j$'s. Combined, we have a "population" with $\kappa = 2.90$ and $\kappa\sqrt{p} = 97.00$.

**Case 2** (Many weak IVs, many null IVs): This setting is identical to Case 1, except we use the first $s = 100$ marginal regression coefficients in the BMI-CAD exposure dataset as non-null SNPs and set the rest $p-s = 1019$ SNPs as null IVs. This leads to $\kappa = 1.05$ and $\kappa\sqrt{p} = 35.12$.

**Case 3** (Many weak IVs, no null IVs): This setting is identical to Case 1, except we use all $p = s = 1119$ marginal regression coefficients in the BMI-CAD exposure dataset as non-null SNPs and there are no null SNPs. This leads to $\kappa = 7.78$ and $\kappa\sqrt{p} = 260.25$.

Based on the $\gamma_j$'s, we set $\Gamma_j = \beta_0\gamma_j$ with $\beta_0 = 0.4$.

Next, for each simulation run we generate summary statistics $\{\hat{\Gamma}_j, \hat{\gamma}_j, \hat{\gamma}_j^*, j = 1, ..., p\}$ based on Assumption 2 with $\gamma_j$'s as described for each of Cases 1-3 and the SEs in the BMI-CAD dataset as $\sigma_{Xj}$, $\sigma_{Yj}$, and $\sigma_{Xj}^*$, $j = 1, ..., p$. Since we cannot generate SEs (part of summary statistics) from this real-data setting for simulation, in each simulation run we set SEs to be the same as $\sigma_{Xj}$, $\sigma_{Yj}$, and $\sigma_{Xj}^*$, $j = 1, ..., p$. This corresponds to treating SDs as SEs as described in the start of Section 3, i.e., assuming that we know SD values.

We compare seven MR methods: the IVW estimator introduced in Section 3, the dIVW estimator proposed in Section 4, and five other methods in the literature, MR-Egger regression [6], weighted median estimator (MR-median) [7], weighted mode estimator (MR-mode) [21], profile score estimator (MR-raps) [40], and profile score with empirical partially Bayes shrinkage weights (MR-raps-shrink) [41]. MR-Egger, MR-median and MR-mode are implemented in the *MendelianRandomization* R package (version 0.4.1) [39]. To make the comparisons fair, we use the $l_2$ loss for MR-raps as implemented in the *mr.raps* package. For every method except MR-raps, we also use different screening procedures, including $\lambda = 0$ (no screening, all SNPs are included), $\lambda = 5.45$ (p-value cutoff based on the threshold of $5 \times 10^{-8}$), and $\lambda = \sqrt{2 \log p}$ ($\approx 3.75$ when $p = 1119$). We also include the dIVW estimator with $\lambda$ determined by the MR-EO algorithm with the maximum number of iterations set to $t_{\max} = 5$ and used the *optimize* function from R in the O-step. The MR-raps does not have any screening. The default MR-raps-shrink always applies a type of screening through Bayes shrinkage with the independent selection dataset.

Table 1 shows (i) the simulation mean and SD of each estimator, (ii) average of SEs, which are calculated according to (3.3) or (4.4) for IVW

TABLE 1

*Simulation results for Cases 1-3 based on 10,000 repetitions with $\beta_0 = 0.4$; $\lambda$ for MR-EO is the simulation average; SD is the simulation standard deviation; SE is the average of standard errors; CP is the simulation coverage probability of the 95% confidence interval based on normal approximation.*

| Case | Method | $\lambda$ | mean | SD | SE | CP |
|---|---|---|---|---|---|---|
| 1 | IVW | 0 | 0.260 | 0.069 | 0.069 | 46.9 |
| $s = 20$ | IVW | 5.45 | 0.398 | 0.094 | 0.093 | 94.8 |
| $p = 1119$ | IVW | $\sqrt{2 \log p} = 3.75$ | 0.398 | 0.087 | 0.087 | 95.1 |
| | dIVW | 0 | 0.402 | 0.107 | 0.107 | 95.2 |
| | dIVW | 5.45 | 0.401 | 0.095 | 0.094 | 94.9 |
| | dIVW | $\sqrt{2 \log p} = 3.75$ | 0.401 | 0.087 | 0.088 | 95.1 |
| | dIVW | MR-EO $\approx 2.80$ | 0.400 | 0.086 | 0.086 | 95.1 |
| | MR-Egger | 0 | 0.335 | 0.082 | 0.082 | 87.5 |
| | MR-Egger | 5.45 | 0.390 | 0.240 | 0.256 | 96.0 |
| | MR-Egger | $\sqrt{2 \log p} = 3.75$ | 0.389 | 0.205 | 0.214 | 95.6 |
| | MR-median | 0 | 0.371 | 0.110 | 0.122 | 96.3 |
| | MR-median | 5.45 | 0.398 | 0.118 | 0.128 | 96.5 |
| | MR-median | $\sqrt{2 \log p} = 3.75$ | 0.397 | 0.113 | 0.124 | 96.7 |
| | MR-mode | 0 | 0.033 | 74 | 75586 | 100 |
| | MR-mode | 5.45 | 0.395 | 0.139 | 0.151 | 97.1 |
| | MR-mode | $\sqrt{2 \log p} = 3.75$ | 0.395 | 0.142 | 0.157 | 97.2 |
| | MR-raps | 0 | 0.401 | 0.105 | 0.105 | 95.2 |
| | MR-raps-shrink | Bayes shrinkage | 0.400 | 0.086 | 0.086 | 95.1 |
| 2 | IVW | 0 | 0.159 | 0.091 | 0.090 | 23.9 |
| $s = 100$ | IVW | 5.45 | 0.397 | 0.206 | 0.206 | 95.1 |
| $p = 1119$ | IVW | $\sqrt{2 \log p} = 3.75$ | 0.394 | 0.183 | 0.183 | 94.9 |
| | dIVW | 0 | 0.404 | 0.233 | 0.233 | 95.4 |
| | dIVW | 5.45 | 0.400 | 0.207 | 0.207 | 95.1 |
| | dIVW | $\sqrt{2 \log p} = 3.75$ | 0.400 | 0.186 | 0.186 | 94.9 |
| | dIVW | MR-EO $\approx 2.21$ | 0.396 | 0.167 | 0.167 | 95.0 |
| | MR-Egger | 0 | 0.231 | 0.122 | 0.123 | 72.3 |
| | MR-Egger | 5.45 | 0.388 | 0.948 | 0.966 | 96.2 |
| | MR-Egger | $\sqrt{2 \log p} = 3.75$ | 0.385 | 0.359 | 0.385 | 95.8 |
| | MR-median | 0 | 0.276 | 0.152 | 0.170 | 91.3 |
| | MR-median | 5.45 | 0.396 | 0.228 | 0.245 | 96.6 |
| | MR-median | $\sqrt{2 \log p} = 3.75$ | 0.394 | 0.216 | 0.236 | 96.9 |
| | MR-mode | 0 | -1.062 | 130 | 78125 | 100 |
| | MR-mode | 5.45 | 0.387 | 0.267 | 0.291 | 97.0 |
| | MR-mode | $\sqrt{2 \log p} = 3.75$ | 0.390 | 0.237 | 0.286 | 97.1 |
| | MR-raps | 0 | 0.398 | 0.224 | 0.226 | 94.9 |
| | MR-raps-shrink | Bayes shrinkage | 0.399 | 0.160 | 0.159 | 95.2 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 3 | IVW | 0 | 0.352 | 0.047 | 0.047 | 82.6 |
| $s = 1119$ | IVW | 5.45 | 0.395 | 0.086 | 0.087 | 95.4 |
| $p = 1119$ | IVW | $\sqrt{2 \log p} = 3.75$ | 0.392 | 0.068 | 0.069 | 95.0 |
| | dIVW | 0 | 0.400 | 0.054 | 0.054 | 94.7 |
| | dIVW | 5.45 | 0.399 | 0.087 | 0.088 | 95.4 |
| | dIVW | $\sqrt{2 \log p} = 3.75$ | 0.399 | 0.070 | 0.070 | 95.4 |
| | dIVW | MR-EO $\approx 0.03$ | 0.400 | 0.054 | 0.054 | 94.8 |
| | MR-Egger | 0 | 0.372 | 0.066 | 0.067 | 93.1 |
| | MR-Egger | 5.45 | 0.383 | 0.189 | 0.198 | 95.4 |
| | MR-Egger | $\sqrt{2 \log p} = 3.75$ | 0.372 | 0.132 | 0.136 | 95.0 |
| | MR-median | 0 | 0.375 | 0.079 | 0.090 | 96.6 |
| | MR-median | 5.45 | 0.394 | 0.114 | 0.125 | 96.8 |
| | MR-median | $\sqrt{2 \log p} = 3.75$ | 0.391 | 0.100 | 0.111 | 96.9 |
| | MR-mode | 0 | 0.750 | 84 | 23253 | 100 |
| | MR-mode | 5.45 | 0.391 | 0.125 | 0.141 | 96.8 |
| | MR-mode | $\sqrt{2 \log p} = 3.75$ | 0.385 | 0.260 | 0.504 | 97.6 |
| | MR-raps | 0 | 0.400 | 0.054 | 0.054 | 94.9 |
| | MR-raps-shrink | Bayes shrinkage | 0.400 | 0.053 | 0.053 | 94.7 |

or dIVW estimators, and (iii) simulation coverage probability (CP) of 95% confidence intervals from normal approximation. The simulation average $\lambda$ determined by the MR-EO algorithm is also included. Under all scenarios, the IVW estimator without screening (i.e., $\lambda = 0$) is biased towards zero, which agrees with our theoretical result since the average IV strength $\kappa$'s are relatively small. The coverage probabilities based on IVW estimators are far from 95% due to the downward bias and the inaccurate normal approximation. The IVW estimator with screening under the threshold $\lambda = 5.45$ or $\sqrt{2 \log p}$ does substantially better, which again agrees with our theory that the IVW estimator with screening requires less stringent assumptions for consistency and asymptotic normality.

The dIVW estimators with and without screening show negligible bias and nominal coverage across all simulation scenarios. This observation agrees with our theoretical assessment that the dIVW estimator requires far less stringent conditions for consistency and asymptotic normality than the IVW estimator. Also, the dIVW estimator with screening improves the dIVW estimator by having a smaller SD in Cases 1-2 where many IVs are null. However, in Case 3 where all IVs are non-null but many are weak, screening in dIVW does not lead to any improvement. In all cases, our MR-EO algorithm adapts to the underlying data and produces the most efficient estimate of

$\beta_0$ among dIVW estimators, all without losing coverage or large gains in bias. Finally, all SEs based on (3.3) and (4.4) are close to the simulated SDs of IVW and dIVW estimators, even for the biased IVW estimator without screening.

The MR-Egger, MR-median and MR-mode estimators without screening are biased when the average IV strength is small. In particular, the MR-mode without screening can be severely biased with unrealistically large SE. MR-Egger, MR-median and MR-mode with screening thresholds at 5.45 or $\sqrt{2\log p}$ generally have larger SDs compared to the dIVW estimators thresholded at the same level. Also, even with thresholding, these three methods (MR-Egger, MR-median and MR-mode) have larger biases than the dIVW estimator because they inherently rely on using the ratio estimator $\hat{\beta}_j$.

The performance of MR-raps is comparable to that of dIVW estimator (4.1). Both methods do not require the independent selection dataset for screening, but the dIVW estimator without screening has a simple explicit form. The MR-raps-shrink uses an independent selection dataset to improve performance and it is comparable to the dIVW estimator with screening and $\lambda$ chosen by MR-EO. However, MR-raps-shrink is computationally more complicated than dIVW with MR-EO and may not have a unique (or well-defined) solution as mentioned in [41].

Table 2 presents the total number of IVs selected during screening as well as the number of non-null IVs selected. In Case 1 where non-null IVs are strong and a good IV selection procedure should perform well, we see that the number of non-null IVs selected based on genome-wide significance (p-value $\leq 5 \times 10^{-8}$ or $\lambda \approx 5.45$) is much too small compared with $s = 20$, the true number of non-null IVs. On the other hand, both $\lambda = \sqrt{2\log p}$ and MR-EO select close to $s = 20$ non-null IVs. MR-EO selects more null IVs because it aims for efficiency instead of consistent IV selection. In Cases 2-3, there are many weak non-null IVs and all IV selection procedures via thresholding are not adequate (Table 2). However, this is not surprising as screening is used in MR to ultimately improve estimation of $\beta_0$, rather than to consistently select non-null IVs. Finally, comparing the results in Tables 1 and 2 indicates that for the dIVW estimator, removing too many weak IVs may lead to an inefficient estimator of $\beta_0$.

In the Supplementary Material, we conduct a similar simulation study with balanced horizontal pleiotropy added to Case 3. The results are nearly identical to Case 3 without pleiotropy. Also in the Supplementary Material, we conduct a simulation study where balanced horizontal pleiotropy is violated in Case 3 and we assess the bias of our estimator. We generally find

TABLE 2
*Average number of total IVs and non-null IVs selected from the selection dataset of $p = 1119$ IVs under different thresholds.*

| | $\lambda = 5.45$ | | $\lambda = \sqrt{2\log p} = 3.75$ | | MR-EO | |
|---|---|---|---|---|---|---|
| Case | total | non-null | total | non-null | total | non-null |
| 1, $s = 20$ | 12.8 | 12.8 | 18.4 | 18.2 | 27.0 | 19.8 |
| 2, $s = 100$ | 3.9 | 3.9 | 9.2 | 9.0 | 63.2 | 27.8 |
| 3, $s = 1119$ | 23.8 | 23.8 | 84.4 | 84.4 | 1019.6 | 1019.6 |

that the dIVW estimator is biased, but the magnitude of the bias is often mild. This is because the bias term is a weighted average of $\alpha_j/\gamma_j$'s with weights $w_j q_{\lambda,j}$'s, where large $\alpha_j/\gamma_j$ tends to be downweighted by $w_j q_{\lambda,j}$.

Overall, there are four takeaways from this simulation study. First, the dIVW estimator with or without screening always outperforms the IVW estimator. Second, without a selection dataset, the dIVW estimator and MR-raps have comparable performances and both are far better than other MR methods under consideration. Third, with a selection dataset, the dIVW with screening and MR-raps-shrink have comparable performances and outperform other methods. Finally, if a selection dataset is available and used to improve efficiency, we suggest the threshold to be $\lambda = \sqrt{2\log p}$ or $\lambda$ produced by the MR-EO algorithm, instead of the usual cutoff $\lambda \approx 5.45$.

5.2. *Empirical Guidelines for Asymptotics.* In practice, it is important to have some sense of what is "a large enough" sample size for the asymptotic results in the paper to serve as good approximations. Many researchers in MR have conducted such analysis for the IVW estimator, most notably [10]. We conduct a similar simulation-based analysis for the dIVW estimator where we examine what would be a "large" effective sample size, as measured by $\kappa_\lambda\sqrt{p_\lambda}/\max(1,\lambda^2)$, for the asymptotics promised by Theorem 4.1 to be plausible.

The setting is identical to Case 3 in Section 5.1. We choose a grid of 100 equally spaced $\lambda$'s between 0 and 10. For each $\lambda$, we generate 1,000 simulation datasets and calculate the corresponding $\hat{\beta}_{\lambda,\mathrm{dIVW}}$ for each dataset. Figure 1 plots these $\hat{\beta}_{\lambda,\mathrm{dIVW}}$ values against $\hat{\kappa}_\lambda\sqrt{\hat{p}_\lambda}/\max(1,\lambda^2)$ as well as two standard error bands centered at $\beta_0$ (shaded area). Note that the sample size $n_X$ and the number of IVs $p$ are fixed and, therefore, as $\hat{\kappa}_\lambda\sqrt{\hat{p}_\lambda}/\max(1,\lambda^2)$ grows, the confidence band first shrinks and then becomes relatively stable.

We find that for any $\lambda$, the coverage probability for the dIVW estimator with screening ranges from 93.5% to 96.0%. However, as $\hat{\kappa}_\lambda\sqrt{\hat{p}_\lambda}/\max(1,\lambda^2)$ grows larger, we see fewer estimates far from $\beta_0$, an indication that asymptotics have "kicked in". This appears to occur when $\hat{\kappa}_\lambda\sqrt{\hat{p}_\lambda}/\max(1,\lambda^2)$ is

greater than 20. Based on this, we recommend that users of dIVW check to make sure that $\hat{\kappa}_\lambda\sqrt{\hat{p}_\lambda}/\max(1,\lambda^2)$ is at least greater than 20 as part of a diagnostic check for the dIVW estimator.

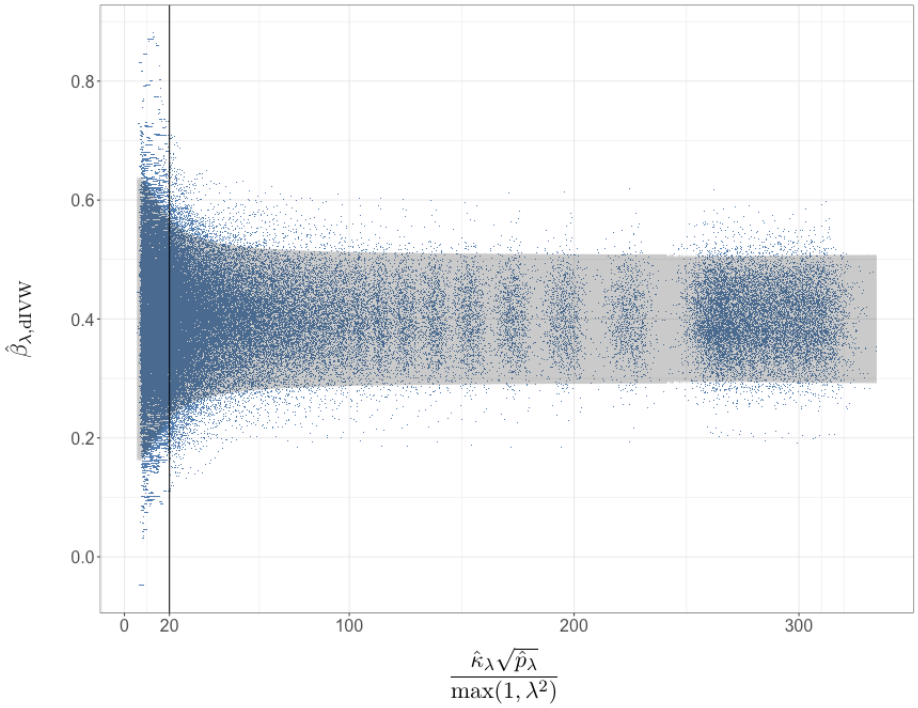


FIG 1. *Evaluation of the consistency and asymptotic normality condition for the dIVW estimator in Theorem 4.1 under Case 3. The x-axis plots the condition that governs the asymptotic rate of the dIVW estimator. The y-axis plots values of the dIVW estimator in 1000 simulations. The shaded area represents two-standard error bands centered at* $\beta_0$.

5.3. *Empirical Evaluation of the Effect of Using SEs not SDs.* In this section, we evaluate the finite sample performance of the proposed dIVW estimators when we don't assume that the SDs of summary statistics are known and instead, we use the estimated SEs, $\hat{\sigma}_{Xj}, \hat{\sigma}_{Yj}$, $\hat{\sigma}^*_{Xj}$'s. We construct a population with parameters $\gamma_j = \varphi_j\sqrt{2h^2/s}$ for $j = 1, ..., s$, and $\gamma_j = 0$ for $j = s+1, ..., p$, where $s$ is the number of non-null SNPs, $h^2$ is the total heritability, i.e., the proportion of variance in $X$ that is attributable to the $s$ non-null SNPs [37], and $\varphi_j$'s are constants that are generated once from a standard normal distribution.

To simulate individual-level data, we first generate $p$ independent SNPs,

$Z_1, ..., Z_p$, from a multinomial distribution with $P(Z_j = 0) = 0.25, P(Z_j = 1) = 0.5, P(Z_j = 2) = 0.25$, and then generate the exposure variable $X$ and outcome variable $Y$ according to models (2.1)-(2.2) with $\eta_X = \eta_Y = 1$, $\beta_0 = 0.4$, $U \sim N(0, 0.6(1-h^2))$, and $E_X, E_Y \sim N(0, 0.4(1-h^2))$. For each simulation repetition, we generate three independent datasets of size $n$ that represent the selection, exposure, and outcome datasets. The summary statistics $\{\hat{\gamma}_j, \hat{\sigma}_{Xj}, j = 1, ..., p\}$, $\{\hat{\Gamma}_j, \hat{\sigma}_{Yj}, j = 1, ..., p\}$, and $\{\hat{\gamma}^*_j, \hat{\sigma}^*_{Xj}, j = 1, ..., p\}$ are obtained from the three datasets through marginal linear regression.

We consider the following combinations of $n$, $p$, $s$, and $h^2$ that reflect what may be found in practice.

**Case 4** $n = 10,000, p = 2,000, s = 200, h^2 = 0.1$, $\kappa = 0.50, \kappa\sqrt{p} = 21.41$.
**Case 5** $n = 10,000, p = 2,000, s = 1,000, h^2 = 0.2$, $\kappa = 0.93, \kappa\sqrt{p} = 41.77$.
**Case 6** $n = 50,000, p = 2,000, s = 1,000, h^2 = 0.2$, $\kappa = 4.67, \kappa\sqrt{p} = 208.84$.
**Case 7** $n = 10,000, p = 2,000, s = 2,000, h^2 = 0.2$, $\kappa = 0.96, \kappa\sqrt{p} = 43.10$.

We consider $n = 10,000$ to be a conservative sample size in modern MR studies and is much smaller than the sample sizes in the BMI-CAD dataset in Section 5.1. Also, $s/p$ takes on values 10%, 50% and 100%.

Table 3 presents the mean, SD, SE, and CP of IVW, dIVW, MR-raps, and MR-raps-shrink estimators, based on 10,000 replications. We omit MR-Egger, MR-median, and MR-mode for conciseness. Overall, a similar trend appears in relation to Table 1 for the case of assuming SDs $\sigma_{Xj}$, $\sigma_{Yj}$, $\sigma^*_{Xj}$ are known: the IVW estimator without screening ($\lambda = 0$) is inconsistent and biased towards 0; the dIVW estimator maintains nominal coverage and outperforms the IVW estimator; and the dIVW estimator with screening by MR-EO performs similarly with the dIVW without screening when $s/p$ is not small. The SEs are generally close to the simulated SDs of point estimators, including the case where the MR-EO is applied. The only exception is in Case 7 when $\lambda = 3.90$. We believe this is because the consistency and asymptotic normality condition, specifically the "effective sample size" value $\kappa_\lambda\sqrt{p_\lambda}/\max(1, \lambda^2)$ is 1.11 when $\lambda = 3.90$ and as Figure 1 illustrates, this value is too small for our asymptotic theory to kick in. Also, upon closer inspection of the numerical results in this case, there are two outlier estimates above 40 or below -20 (out of 10,000 simulation runs). Removing these two simulation runs leads to SD= 0.304 and SE=0.293, which agree more closely with each other. Overall, this observation indicates the importance of performing diagnostic check for the dIVW estimator using $\hat{\kappa}_\lambda\sqrt{\hat{p}_\lambda}/\max(1, \lambda^2)$ and using the MR-EO algorithm to adaptively select $\lambda$ when needed.

We also notice the following observations that were not in Table 1. First, the performance of all estimators tend to improve when $n$ increases (Cases

5-6) even though $s/p = 50\%$. Second, the use of genome-wide significance threshold $\lambda \approx 5.45$ often selects no SNPs in many simulation runs when $s = 1,000$ or $2,000$, another indication that this threshold is too large in MR studies with many weak IVs.

In the Supplementary Material, we also run the same simulations using $\hat{\sigma}_{Xj} = \sigma_{Xj}, \hat{\sigma}_{Yj} = \sigma_{Yj}, \hat{\sigma}^*_{Xj} = \sigma^*_{Xj}$ and obtain almost identical results as Table 3; see Table S3 of the Supplementary Material. This indicates that the effect of assuming known SDs and using them as SEs is negligible, which agrees with many empirical results in the literature as well as our theoretical results in Theorem 4.1.

**6. Real Data Example.** We apply our methods to the BMI-CAD example described in Section 5.1. Table 4 summarizes the results, where $\text{dIVW}_\alpha$ denotes the dIVW estimator developed under balanced horizontal pleiotropy, $\text{MR-raps}_\alpha$ and $\text{MR-raps-shrink}_\alpha$ are MR-raps estimators that account for balanced horizontal pleiotropy by setting the over.dispersion parameter in the *mr.raps* R package to be *TRUE*.

We make the following comments. First, we see that the MR-mode estimator with or without screening is very unstable. Second, in light of our simulation result under Case 3, we suspect that the IVW estimate 0.315 without screening ($\lambda = 0$) is slightly biased towards zero, compared with the dIVW estimate 0.365, although the difference is not statistically significant. Third, except for MR-Egger and MR-mode, selecting IVs based on genome-wide significance (i.e., $\lambda = 5.45$) produces point estimates between 0.278 and 0.287 and larger SEs across all methods, most likely because too many IVs are screened out. Fourth, except for the IVW estimator without screening, the dIVW estimator with MR-EO achieves the smallest SE among all estimates. But, since the dIVW estimate based on MR-EO has the same SE as the dIVW estimate without screening, screening is probably not necessary for this dataset with many weak IVs; this is also supported by the fact that there is not much difference between MR-raps and MR-raps-shrink. Fifth, the estimators accounting for balanced horizontal pleiotropy are similar to those without it, except for an expected increase in SEs due to the random effect terms.

Following [40], we run a diagnostic to assess the plausibility of Assumption 2 by constructing a Quantile-Quantile plot of the standardized residuals, $(\hat{\Gamma}_j - \hat{\beta}_{\text{dIVW}}\hat{\gamma}_j)/(\hat{\sigma}^2_{Yj} + \hat{\beta}^2_{\text{dIVW}}\hat{\sigma}^2_{Xj})^{1/2}$, $j = 1, ..., p$. Figure 2 shows the result. Since the residuals line up close to the 45-degree line, Assumption 2 is likely to hold for this example. In the Supplementary Material, a similar figure is obtained for assessing the plausibility of Assumption $2'$.

TABLE 3

*Simulation results for Cases 4-7 based on 10,000 repetitions with $\beta_0 = 0.4$; $\lambda$ for MR-EO is the simulation average; SD is the simulation standard deviation; SE is the average of standard errors; CP is the simulation coverage probability of the 95% confidence interval based on normal approximation.*

| Case | Method | $\lambda$ | mean | SD | SE | CP |
|---|---|---|---|---|---|---|
| 4 | IVW | 0 | 0.129 | 0.027 | 0.027 | 0 |
| $s = 200$ | IVW | 5.45 | 0.393 | 0.125 | 0.122 | 94.6 |
| $p = 2000$ | IVW | $\sqrt{2 \log p} = 3.90$ | 0.382 | 0.075 | 0.074 | 93.8 |
| $n = 10000$ | | | | | | |
| | dIVW | 0 | 0.402 | 0.090 | 0.089 | 95.0 |
| | dIVW | 5.45 | 0.406 | 0.131 | 0.127 | 95.0 |
| | dIVW | $\sqrt{2 \log p} = 3.90$ | 0.402 | 0.079 | 0.078 | 95.0 |
| | dIVW | MR-EO $\approx 2.17$ | 0.396 | 0.061 | 0.060 | 94.9 |
| | MR-raps | 0 | 0.401 | 0.085 | 0.085 | 94.8 |
| | MR-raps-shrink | Bayes shrinkage | 0.399 | 0.062 | 0.062 | 95.2 |
| 5 | IVW | 0 | 0.193 | 0.024 | 0.023 | 0 |
| $s = 1000$ | IVW | 5.45 | select no IV over 25% of runs | | | |
| $p = 2000$ | IVW | $\sqrt{2 \log p} = 3.90$ | 0.364 | 0.099 | 0.099 | 92.8 |
| $n = 10000$ | | | | | | |
| | dIVW | 0 | 0.401 | 0.051 | 0.051 | 94.7 |
| | dIVW | 5.45 | select no IV over 25% of runs | | | |
| | dIVW | $\sqrt{2 \log p} = 3.90$ | 0.405 | 0.112 | 0.111 | 95.3 |
| | dIVW | MR-EO $\approx 1.10$ | 0.394 | 0.048 | 0.048 | 94.6 |
| | MR-raps | 0 | 0.400 | 0.049 | 0.049 | 94.5 |
| | MR-raps-shrink | Bayes shrinkage | 0.399 | 0.047 | 0.047 | 94.8 |
| 6 | IVW | 0 | 0.330 | 0.014 | 0.014 | 0.1 |
| $s = 1000$ | IVW | 5.45 | 0.390 | 0.024 | 0.024 | 93.0 |
| $p = 2000$ | IVW | $\sqrt{2 \log p} = 3.90$ | 0.386 | 0.019 | 0.018 | 88.1 |
| $n = 50000$ | | | | | | |
| | dIVW | 0 | 0.400 | 0.017 | 0.017 | 94.8 |
| | dIVW | 5.45 | 0.400 | 0.024 | 0.024 | 94.8 |
| | dIVW | $\sqrt{2 \log p} = 3.90$ | 0.400 | 0.019 | 0.019 | 94.5 |
| | dIVW | MR-EO $\approx 1.31$ | 0.399 | 0.017 | 0.017 | 94.9 |
| | MR-raps | 0 | 0.400 | 0.017 | 0.017 | 94.8 |
| | MR-raps-shrink | Bayes shrinkage | 0.400 | 0.017 | 0.017 | 94.8 |
| 7 | IVW | 0 | 0.196 | 0.023 | 0.023 | 0 |
| $s = 2000$ | IVW | 5.45 | select no IV over 81% of runs | | | |
| $p = 2000$ | IVW | $\sqrt{2 \log p} = 3.90$ | 0.343 | 0.197 | 0.195 | 93.5 |
| $n = 10000$ | | | | | | |
| | dIVW | 0 | 0.400 | 0.050 | 0.049 | 94.5 |
| | dIVW | 5.45 | select no IV over 81% of runs | | | |
| | dIVW | $\sqrt{2 \log p} = 3.90$ | 0.423 | 0.622 | 0.428 | 96.5 |
| | dIVW | MR-EO $\approx 0.44$ | 0.395 | 0.051 | 0.050 | 94.4 |
| | MR-raps | 0 | 0.400 | 0.048 | 0.047 | 94.8 |
| | MR-raps-shrink | Bayes shrinkage | 0.399 | 0.048 | 0.047 | 94.8 |

TABLE 4
*Point estimates of exposure effect and their SEs (in parentheses) from different MR methods in the BMI-CAD example.*

| | | | | MR-EO | MR-EO$_\alpha$ |
|---|---|---|---|---|---|
| $\lambda$ | 0 | 5.45 | $\sqrt{2\log p}=3.75$ | 0.57 | 0.59 |
| # of IVs selected | 1119 | 44 | 165 | 1029 | 1023 |
| $\hat{\kappa}_\lambda\sqrt{\hat{p}_\lambda}/\max(1,\lambda^2)$ | 226.8 | 16.3 | 25.7 | 232.4 | 233.1 |
| IVW | 0.315 (0.050) | 0.282 (0.084) | 0.319 (0.068) | | |
| dIVW | 0.365 (0.058) | 0.287 (0.085) | 0.331 (0.071) | 0.345 (0.058) | |
| dIVW$_\alpha$ | 0.365 (0.067) | 0.287 (0.100) | 0.331 (0.082) | 0.345 (0.067) | |
| MR-Egger | 0.386 (0.077) | 0.513 (0.184) | 0.390 (0.129) | | |
| MR-median | 0.322 (0.097) | 0.278 (0.124) | 0.304 (0.116) | | |
| MR-mode | 0.739 (402.9) | 0.499 (0.402) | 0.488 (4.241) | | |
| MR-raps | 0.382 (0.061) | | | | |
| MR-raps$_\alpha$ | 0.367 (0.067) | | | | |
| MR-raps-shrink (Bayes shrinkage) | | 0.388 (0.060) | | | |
| MR-raps-shrink$_\alpha$(Bayes shrinkage) | | 0.374 (0.067) | | | |

The subscript $\alpha$ indicates application under balanced horizontal pleiotropy

**7. Summary and Discussion.** In two-sample summary-data MR studies, we show that the IVW estimator requires stringent conditions on the average strength of IVs for consistency and asymptotic normality. The IVW estimator with screening relaxes these conditions somewhat, but requires carefully choosing a threshold $\lambda$ and a third independent dataset. We then propose a simple modification of the IVW estimator, called the debiased IVW (dIVW) estimator. The dIVW estimator, with or without screening, is shown to be consistent and asymptotically normal under conditions that are much weaker than those required by the IVW estimator, with or without screening. Finally, we provide some theoretical and numerical results on assuming the commonly invoked known-variance condition.

While our work primarily focuses on the "standard" IVW estimator commonly used in practice, as suggested by the anonymous referees and the editor, the standard IVW estimator, with or without screening, can be viewed as instances of the generalized IVW estimator

$$\frac{\sum_{j=1}^{p}\hat{\beta}_j f(\hat{w}_j)}{\sum_{j=1}^{p} f(\hat{w}_j)},$$

where $\hat{\beta}_j = \hat{\Gamma}_j/\hat{\gamma}_j$, and $f$ is a general weighting function. This general weighting function can encompass soft thresholding and other IV selection procedures. However, this class of estimators does not include the proposed dIVW estimator since the weights from the dIVW estimator will not sum to 1. Nevertheless, extending the current theory to better understand this broader

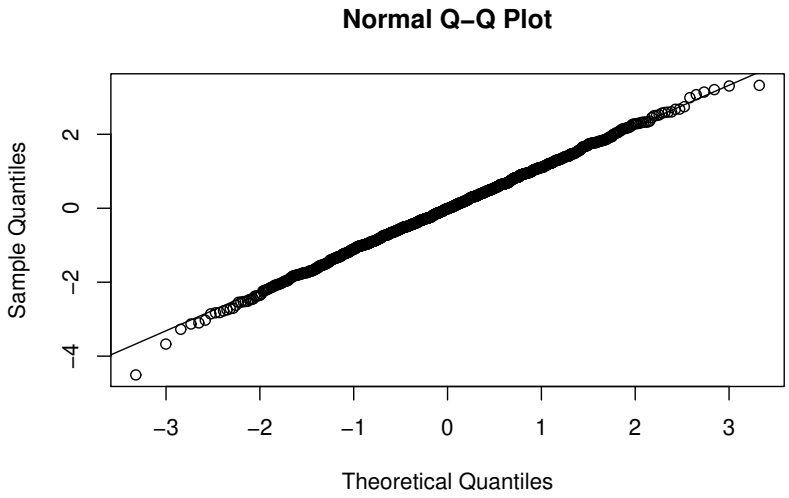


FIG 2. *Quantile-Quantile plot of the standardized residuals against a standard normal.*

class of IVW estimators under many weak IVs is an important direction for future research.

Finally, based on our theoretical and simulation work, we make three recommendations for practice. First, we argue that the dIVW estimator without screening should be the default baseline estimator for two-sample summary-data MR studies instead of the IVW estimator. It is as simple as the IVW estimator, and has provable robustness against many weak instruments and balanced horizontal pleiotropy, lending itself as the baseline estimator for investigating more complex pleiotropy. Second, if there are many irrelevant IVs and summary statistics from a third independent selection dataset are available, we may improve the efficiency of the dIVW estimator by screening with threshold $\lambda$ produced by the MR-EO algorithm; we discourage the use of the genome-wide significance p-value threshold $\lambda \approx 5.45$ as it tends to screen out too many IVs. Third, for the promised theoretical properties of the proposed dIVW estimator to hold, it is important to perform diagnostics by constructing Quantile-Quantile plot of the standardized residuals and also checking that $\hat{\kappa}_\lambda \sqrt{\hat{p}_\lambda} / \max(1, \lambda^2)$ is at least greater than 20.

**Acknowledgements.** The authors would like to thank the Associate Editor and two anonymous referees for useful comments that led to a much improved paper.

**Software and Reproducibility.** R code for the methods proposed in this paper can be found in the R package mr.divw, which is posed at https://github.com/tye27/mr.divw. Numerical examples in this article can be reproduced by running examples in the R package.

## SUPPLEMENTARY MATERIAL

**Supplementary Material: Debiased Inverse-Variance Weighted Estimator in Two-Sample Summary-Data Mendelian Randomization**
(doi: COMPLETED BY THE TYPESETTER; .pdf). We provide additional numerical results and theoretical proofs for the theorems in the paper.

## References.


[1] Abbott, L., Bryant, S., Churchhouse, C. and et al. (2018). Round 2 GWAS Results of Thousands of Phenotypes in the UK BioBank. *http://www.nealelab.is/uk-biobank* (14 November 2018, date last accessed).

[2] Akiyama, M., Okada, Y., Kanai, M., Takahashi, A., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., Iwasaki, M., Yamaji, T., Sawada, N., Hachiya, T., Tanno, K., Shimizu, A., Hozawa, A., Minegishi, N., Tsugane, S., Yamamoto, M., Kubo, M. and Kamatani, Y. (2017). Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. *Nature Genetics* **49** 1458-1467.

[3] Andrews, D. W. K. and Stock, J. H. (2005). Inference with Weak Instruments. Working Paper No. 313, National Bureau of Economic Research.

[4] Angrist, J. D. and Krueger, A. B. (2001). Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives* **15** 69-85.

[5] Baiocchi, M., Cheng, J. and Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in medicine* **33** 2297–2340.

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

[7] Bowden, J., Davey Smith, G., Haycock, P. C. and Burgess, S. (2016). Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genetic Epidemiology* **40** 304–314.

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

[9] Bowden, J., Del Greco M, F., Minelli, C., Zhao, Q., Lawlor, D. A., Sheehan, N. A., Thompson, J. and Davey Smith, G. (2019). Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the NOME assumption. *International journal of epidemiology* **48** 728–742.

[10] Burgess, S., Butterworth, A. and Thompson, S. G. (2013). Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data. *Genetic Epidemiology* **37** 658–665.

[11] Burgess, S., Small, D. S. and Thompson, S. G. (2015). A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research* **26** 2333–2355.
[12] Burgess, S. and Thompson, S. G. (2011). Bias in causal estimates from Mendelian randomization studies with weak instruments. *Statistics in Medicine* **30** 1312–1323.
[13] Burgess, S., Thompson, S. G. and Collaboration, C. C. G. (2011). Avoiding bias from weak instruments in Mendelian randomization studies. *International Journal of Epidemiology* **40** 755–764.
[14] Burgess, S. and Thompson, S. G. (2012). Improving bias and coverage in instrumental variable analysis with weak instruments for continuous and binary outcomes. *Statistics in Medicine* **31** 1582–1600.
[15] Chao, J. C. and Swanson, N. R. (2005). Consistent Estimation with a Large Number of Weak Instruments. *Econometrica* **73** 1673–1692.
[16] Cirulli, E. T. and Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics* **11** 415-425.
[17] Corbin, L. J., Richmond, R. C., Wade, K. H., Burgess, S., Bowden, J., Smith, G. D. and Timpson, N. J. (2016). BMI as a Modifiable Risk Factor for Type 2 Diabetes: Refining and Understanding Causal Estimates Using Mendelian Randomization. *Diabetes* **65** 3002.
[18] Davey Smith, G. and Ebrahim, S. (2003). 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* **32** 1–22.
[19] Didelez, V. and Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research* **16** 309–330.
[20] Gibson, G. (2012). Rare and common variants: twenty arguments. *Nature Reviews Genetics* **13** 135 - 145.
[21] Hartwig, F. P., Davey Smith, G. and Bowden, J. (2017). Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *International Journal of Epidemiology* **46** 1985–1998.
[22] Hemani, G., Bowden, J. and Davey Smith, G. (2018). Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Human molecular genetics* **27** R195–R208.
[23] Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., Langdon, R., Tan, V. Y., Yarmolinsky, J., Shihab, H. A., Timpson, N. J., Evans, D. M., Relton, C., Martin, R. M., Davey Smith, G., Gaunt, T. R., Haycock, P. C. and Loos, R. (2018). The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7** e34408.
[24] Hernan, M. A. and Robins, J. M. (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.
[25] Kamstrup, P. R., Tybjærg-Hansen, A., Steffensen, R. and Nordestgaard, B. G. (2009). Genetically Elevated Lipoprotein(a) and Increased Risk of Myocardial Infarction. *JAMA* **301** 2331–2339.
[26] Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N. and Davey Smith, G. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine* **27** 1133–1163.
[27] Pierce, B. L. and Burgess, S. (2013). Efficient Design for Mendelian Randomization Studies: Subsample and 2-Sample Instrumental Variable Estimators. *American*

*Journal of Epidemiology* **178** 1177–1184.
[28] Pingault, J.-B., O'Reilly, P. F., Schoeler, T., Ploubidis, G. B., Rijsdijk, F. and Dudbridge, F. (2018). Using genetic data to strengthen causal inference in observational research. *Nature Reviews Genetics* **19** 566–580.
[29] Qi, G. and Chatterjee, N. (2019). Mendelian randomization analysis using mixture models for robust and efficient estimation of causal effects. *Nature Communications* **10** 1941.
[30] Sawa, T. (1969). The Exact Sampling Distribution of Ordinary Least Squares and Two-Stage Least Squares Estimators. *Journal of the American Statistical Association* **64** 923–937.
[31] Smith, G. D. and Ebrahim, S. (2004). Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology* **33** 30–42.
[32] Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. and Smoller, J. W. (2013). Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics* **14** 483 - 495.
[33] Staiger, D. and Stock, J. H. (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica* **65** 557–586.
[34] Stock, J. H., Wright, J. H. and Yogo, M. (2002). A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments. *Journal of Business & Economic Statistics* **20** 518–529.
[35] the CARDIoGRAMplusC4D Consortium, Nikpay, M., Goel, A., Won, H.-H., Hall, L. M., Willenborg, C. and et al. (2015). A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics* **47** 1121 -1130.
[36] Verbanck, M., Chen, C.-Y., Neale, B. and Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature Genetics* **50** 693–698.
[37] Visscher, P. M., Hill, W. G. and Wray, N. R. (2008). Heritability in the genomics era —concepts and misconceptions. *Nature Reviews Genetics* **9** 255–266.
[38] Wang, S. and Kang, H. (2019). Weak-Instrument Robust Tests in Two-Sample Summary-Data Mendelian Randomization. *arXiv 1909.06950*.
[39] Yavorska, O. O. and Burgess, S. (2017). MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *International Journal of Epidemiology* **46** 1734–1739.
[40] Zhao, Q., Wang, J., Hemani, G., Bowden, J. and Small, D. S. (2019a). Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *To appear in Annals of Statistics*.
[41] Zhao, Q., Chen, Y., Wang, J. and Small, D. S. (2019b). Powerful three-sample genome-wide design and robust statistical inference in summary-data Mendelian randomization. *International Journal of Epidemiology*.
[42] Zheng, J., Baird, D., Borges, M.-C., Bowden, J., Hemani, G., Haycock, P., Evans, D. M. and Smith, G. D. (2017). Recent Developments in Mendelian Randomization Studies. *Current Epidemiology Reports* **4** 330–345.

Department of Statistics
University of Pennsylvania
Philadelphia, PA 19104
E-mail: tingye@wharton.upenn.edu

Department of Statistics
University of Wisconsin-Madison
Madison, WI 53706
E-mail: shao@stat.wisc.edu
hyunseung@stat.wisc.edu