STATISTICAL INFERENCE IN TWO-SAMPLE SUMMARY-DATA MENDELIAN RANDOMIZATION USING ROBUST ADJUSTED PROFILE SCORE

By Qingyuan Zhao, Jingshu Wang, Gibran Hemani, Jack Bowden, and Dylan S. Small

University of Pennsylvania and University of Bristol

Mendelian randomization (MR) is a method of exploiting genetic variation to unbiasedly estimate a causal effect in presence of unmeasured confounding. MR is being widely used in epidemiology and other related areas of population science. In this paper, we study statistical inference in the increasingly popular two-sample summary-data MR design. We show a linear model for the observed associations approximately holds in a wide variety of settings when all the genetic variants satisfy the exclusion restriction assumption, or in genetic terms, when there is no pleiotropy. In this scenario, we derive a maximum profile likelihood estimator with provable consistency and asymptotic normality. However, through analyzing real datasets, we find strong evidence of both systematic and idiosyncratic pleiotropy in MR, echoing some recent discoveries in statistical genetics. We model the systematic pleiotropy by a random effects model, where no genetic variant satisfies the exclusion restriction condition exactly. In this case we propose a consistent and asymptotically normal estimator by adjusting the profile score. We then tackle the idiosyncratic pleiotropy by robustifying the adjusted profile score. We demonstrate the robustness and efficiency of the proposed methods using several simulated and real datasets.

1. Introduction. A common goal in epidemiology is to understand the causal mechanisms of disease. If it was known that a risk factor causally influenced an adverse health outcome, effort could be focused to develop an intervention (e.g., a drug or public health intervention) to reduce the risk factor and improve the population's health. In settings where evidence from a randomized controlled trial is lacking, inferences about causality are made using observational data. The most common design of observational study is to control for confounding variables between the exposure and the outcome. However, this strategy can easily lead to biased estimates and false conclusions when one or several important confounding variables are

MSC 2010 subject classifications: primary 65J05; secondary 46N60, 62F35
Keywords and phrases: causal inference, limited information maximum likelihood, weak instruments, errors in variables, path analysis, pleiotropy effects

overlooked.

Mendelian randomization (MR) is an alternative study design that leverages genetic variation to produce an unbiased estimate of the causal effect even when there is unmeasured confounding. MR is both old and new. It is a special case of the instrumental variable (IV) methods [21], which date back to the 1920s [54] and have a long and rich history in econometrics and statistics. The first MR design was proposed by Katan [33] over 3 decades ago and later popularized in genetic epidemiology by Davey Smith and Ebrahim [18]. As a public health study design, MR is rapidly gaining popularity from just 5 publications in 2003 to over 380 publications in the year 2016 [1]. However, due to the inherent complexity of genetics (the understanding of which is rapidly evolving) and the make-up of large international disease databases being utilized in the analysis, MR has many unique challenges compared to classical IV analyses in econometrics and health studies. Therefore, MR does not merely involve plugging genetic instruments in existing IV methods. In fact, the unique problem structure has sparked many recent methodological advancements [8, 9, 24, 32, 34, 50, 51].

Much of the latest developments in Mendelian randomization has been propelled by the increasing availability and scale of genome-wide association studies (GWAS) and other high-throughput genomic data. A particularly attractive proposal is to automate the causal inference by using published GWAS data [15], and a large database and software platform is currently being developed [29]. Many existing IV and MR methods [e.g. 24, 40, 50], though theoretically sound and robust to different kinds of biases, require having individual-level data. Unfortunately, due to privacy concerns, the access to individual-level genetic data is almost always restricted and usually only the GWAS summary statistics are publicly available. This data structure has sparked a number of new statistical methods anchored within the framework of meta-analysis [e.g. 8, 9, 27]. They are intuitively simple and can be conveniently used with GWAS summary data, thus are quickly gaining popularity in practice. However, the existing summary-data MR methods often make unrealistic simplifying assumptions and generally lack theoretical support such as statistical consistency and asymptotic sampling distribution results.

This paper aims to resolve this shortcoming by developing statistical methods that can be used with summary data, have good theoretical properties, and are robust to deviations of the usual IV assumptions. In the rest of the Introduction, we will introduce a statistical model for GWAS summary data and demonstrate the MR problem using a real data example. This example will be repeatedly used in subsequent sections to motivate and illustrate the statistical methods. We will conclude the Introduction by discussing the methodological challenges in MR and outline our solution.

1.1. *Two-sample MR with summary data.* We are interested in estimating the causal effect of an exposure variable X on an outcome variable Y. The causal effect is confounded by unobserved variables, but we have p genetic variants (single nucleotide polymorphisms, SNPs), $Z_1, Z_2, \ldots, Z_p$, that are *valid* instrumental variables (defined in Section 2.1). These IVs can help us to obtain unbiased estimate of the causal effect even when there is unmeasured confounding. The precise problem considered in this paper is two-sample Mendelian randomization with summary data, where we observe, for SNP $j = 1, \ldots, p$, two associational effects: the SNP-exposure effect $\hat{\gamma}_j$ and the SNP-outcome effect $\hat{\Gamma}_j$. These estimated effects are usually computed from two different samples using a simple linear regression or logistic regression and are or are becoming available in public domain.

Throughout the paper we assume

ASSUMPTION 1. *For every $j \in \{1, \ldots, p\} := [p]$, $\hat{\gamma}_j \sim \mathrm{N}(\gamma_j, \sigma_{Xj}^2)$, $\hat{\Gamma}_j \sim \mathrm{N}(\Gamma_j, \sigma_{Yj}^2)$, and the variances $(\sigma_{Xj}^2, \sigma_{Yj}^2)_{j \in [p]}$ are known. Furthermore, the $2p$ random variables $(\hat{\gamma}_j)_{j \in [p]}$ and $(\hat{\Gamma}_j)_{j \in [p]}$ are mutually independent.*

The first assumption is quite reasonable as typically there are hundreds of thousands of samples in modern GWAS, making the normal approximation very accurate. The independence between $(\hat{\gamma}_j)_{j \in [p]}$ and $(\hat{\Gamma}_j)_{j \in [p]}$ is guaranteed because the effects are computed from independent samples. The independence across SNPs is reasonable if we only use uncorrelated SNPs by using a tool called linkage disequilibrium (LD) clumping [29, 43, 44]. See Section 2 for more justifications of the last assumption.

Our key modeling assumption for summary-data MR is

MODEL FOR GWAS SUMMARY DATA. *There exists a real number β_0 such that*

$$\Gamma_j \approx \beta_0 \gamma_j \text{ for almost all } j \in [p]. \tag{1.1}$$

In Section 2 and Appendix A, we will explain why this model likely holds for a variety of situations and why the parameter β_0 may be interpreted as the causal effect of X on Y. However, by investigating a real data example, we will demonstrate in Section 3.4 that it is very likely that the strict equality $\Gamma_j = \beta_0 \gamma_j$ is not true for some if not most j. For now we will proceed with the loose statement in (1.1), but it will be soon made precise in several ways.

Assumption 1 and model (1.1) suggest two different strategies of estimating β_0:

1. Use the Wald ratio $\hat{\beta}_j = \hat{\Gamma}_j/\hat{\gamma}_j$ [53] as each SNP's individual estimate of β_0, then aggregate the estimates using a robust meta-analysis method. Most existing methods for summary-data MR follow this line [8, 9, 27], however the Wald estimator $\hat{\beta}_j$ is heavily biased when γ_j is small, a phenomenon known as "weak instrument bias". See Bound, Jaeger and Baker [7] and Section 1.3 below.
2. Treat equation (1.1) as an errors-in-variables regression problem [16], where we are regressing $\hat{\Gamma}_j$, whose expectation is Γ_j, on $\hat{\gamma}_j$, which can be regarded as a noisy observation of the actual regressor γ_j. Then we directly estimate β_0 in a robust way. This is the novel approach taken in this paper and will be described and tested in detail.

1.2. *A motivating example.* Next we introduce a real data example that will be repeatedly used in the development of this paper. In this example we are interested in estimating the causal effect of a person's Body Mass Index (BMI) on Systolic Blood Pressure (SBP). We obtained publicly available summary data from three GWAS with non-overlapping samples:

`BMI-FEM`: BMI in females by the Genetic Investigation of ANthropometric Traits (GIANT) consortium [35] (sample size: 171977, unit: kg/m^2).

`BMI-MAL`: BMI in males in the same study by the GIANT consortium (sample size: 152893, unit: kg/m^2).

`SBP-UKBB`: SBP using the United Kingdom BioBank (UKBB) data (sample size: 317754, unit: mmHg).

Using the `BMI-FEM` dataset and LD clumping, we selected 25 SNPs that are genome-wide significant (p-value $\leq 5 \times 10^{-8}$) and uncorrelated (10000 kilo base pairs apart and $R^2 \leq 0.001$). We then obtained the 25 SNP-exposure effects $(\hat{\gamma}_j)_{j=1}^{25}$ and the corresponding standard errors from `BMI-MAL` and the SNP-outcome effects $(\hat{\Gamma}_j)_{j=1}^{25}$ and the corresponding standard errors from `SBP-UKBB`.

Figure 1 shows the scatter plot of the 25 pairs of genetic effects. Since they are measured with error, we added error bars of one standard error to every point on both sides. The goal of summary-data MR is to find a straight line through the origin that best fits these points. The statistical method should also be robust to violations of model (1.1) since not all SNPs satisfy the relation $\Gamma_j = \beta_0\gamma_j$ exactly. We will come back to this example in Sections 3.4, 4.4 and 5.3 to illustrate our methods.

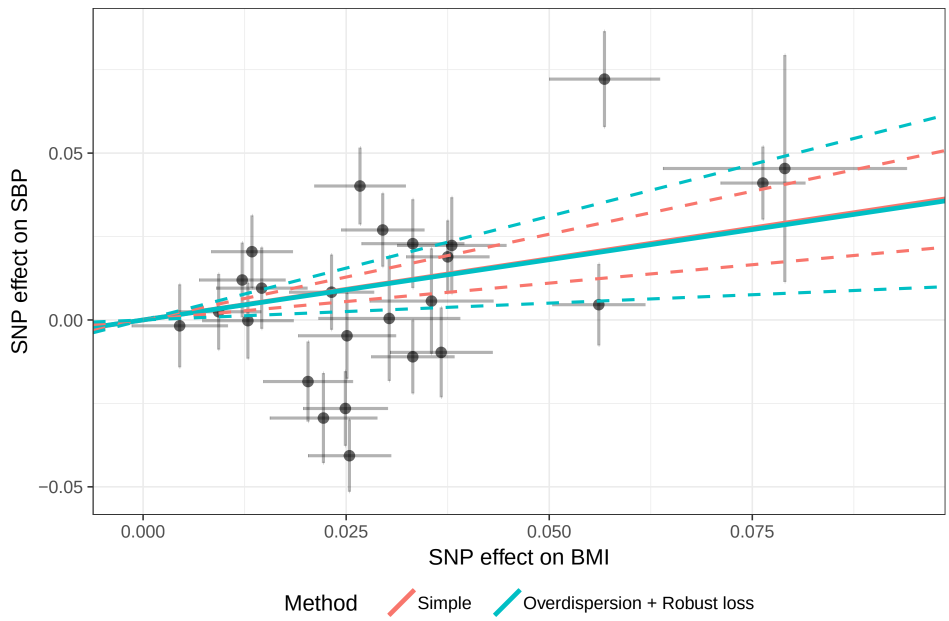


Fig 1: Scatter plot of $\hat{\Gamma}_j$ versus $\hat{\gamma}_j$ in the BMI-SBP example. Each point is augmented by the standard error of $\hat{\Gamma}_j$ and $\hat{\gamma}_j$ on the vertical and horizontal sides. For presentation purposes only, we chose the allele codings so that all $\hat{\gamma}_j$ are positive. Solid lines are the regression slope fitted by two of our methods. Dashed lines are the 95% confidence interval of the slopes. The simple method using unadjusted profile score (PS, described in Section 3) has smaller standard error than the more robust method using robust adjusted profile score (RAPS, described in Section 5), because the simple method does not consider genetic pleiotropy. See also Section 3.4.

1.3. *Statistical Challenges and organization of the paper.* Compared to classical IV analyses in econometrics and health studies, there are many unique challenges in two-sample MR with summary data:

1. Measurement error: Both the SNP-exposure and SNP-outcome effects are clearly measured with error, but most of the existing methods applicable to summary data assume that the sampling error of $\hat{\gamma}_j$ is negligible so a weighted linear regression can be directly used [14].
2. Invalid instruments due to pleiotropy (the phenomenon that one SNP can affect seemingly unrelated traits): A SNP Z_j may causally affect the outcome Y through other pathways not involving the exposure X.

In this case, the approximate linear model $\Gamma_j \approx \beta_0\gamma_j$ might be entirely wrong for some SNPs.

3. Weak instruments: Including a SNP j with very small γ_j can bias the causal effect estimates (especially when the meta-analysis strategy is used). It can also increase the variance of the estimator $\hat{\beta}$. See Section 3.3.2.
4. Selection bias: To avoid the weak instrument bias, the standard practice in MR is to only use the genome-wide significant SNPs as instruments (for example, as implemented in the `TwoSampleMR` R package [29]). However, in many studies the same dataset is used for both selecting SNPs and estimating γ_j, resulting in substantial selection bias even if the selection threshold is very stringent.

Many previous works have considered one or some of these challenges. Bowden et al. [10] proposed a modified Cochran's Q statistic to detect the heterogeneity due to pleiotropy instead of measurement error in $\hat{\gamma}_j$. Addressing the issue of bias due to pleiotropy has attracted lots of attention in the summary-data MR literature [8, 9, 27, 34, 51], but no solid statistical underpinning has yet been given. Other methods with more rigorous statistical theory require individual-level data [24, 40, 50]. The weak instrument problem has been thoroughly studied in the econometrics literature [e.g. 7, 26, 49], but all of this work operates in the individual-level data setting. Finally, the selection bias has largely been overlooked in practice; common widom has been that the selection biases the causal effects towards the null (so it might be less serious) [28] and the bias is perhaps small when a stringent selection criterion is used (in Section 7 we show this is not necessarily the case).

In this paper we develop a novel approach to overcome all the aforementioned challenges by adjusting the profile likelihood of the summary data. The measurement errors of $\hat{\gamma}_j$ and $\hat{\Gamma}_j$ (challenge 1) are naturally incorporated in computing the profile score. To tackle invalid IVs (challenge 2), we will consider three models for the GWAS summary data with increasing complexity:

MODEL 1 (No pleiotropy). *The linear model* $\Gamma_j = \beta_0\gamma_j$ *is true for every* $j \in [p]$.

MODEL 2 (Systematic pleiotropy). *Assume* $\alpha_j = \Gamma_j - \beta_0\gamma_j \overset{i.i.d.}{\sim} \mathrm{N}(0, \tau_0^2)$ *for* $j \in [p]$ *and some small* τ_0^2.

MODEL 3 (Systematic and idiosyncratic pleitropy). *Assume* $\alpha_j,\ j \in [p]$

are from a contaminated normal distribution: most α_j *are distributed as* $\mathrm{N}(0, \tau_0^2)$ *but some* $|\alpha_j|$ *may be much larger.*

The consideration of these three models is motivated by not only the theoretical models in Section 2 but also characteristics observed in real data (Sections 3.4, 4.4 and 5.3) and recent empirical evidence in genetics [13, 46].

The three models are considered in Sections 3 to 5, respectively. We will propose estimators that are provably consistent and asymptotically normal in Models 1 and 2. We will then derive an estimator that is robust to a small proportion of outliers in Model 3. We believe Model 3 best explains the real data and the corresponding Robust Adjusted Profile Score (RAPS) estimator is the clear winner in all the empirical examples.

Although weak IVs may bias the individual Wald's ratio estimator (challenge 3), we will show, both theoretically and empirically, that including additional weak IVs is usually helpful for our new estimators when there are already strong IVs or many weak IVs. Finally, the selection bias (challenge 4) is handled by requiring use of an independent dataset for IV selection as we have done in 1.2.

The rest of the paper is organized as follows. In Section 2 we give theoretical justifications of the model (1.1) for GWAS summary data. Then in Sections 3 to 5 we describe an adjusted profile score approach of statistical inference in Models 1 to 3, respectively. The paper is concluded with a simulation example in Section 6, another real data example in Section 7 and more discussion in Section 8.

2. Statistical model for MR. In this Section we explain why the linear model (1.1) for GWAS summary data may hold in many MR problems.

2.1. *Validity of instrumental variables.* In order to study the origin of the linear model (1.1) for summary data and give a causal interpretation to the parameter β_0, we must specify how the original data $(X, Y, Z_1, \ldots, Z_p)$ are generated and how the summary statistics are computed. Consider the following structural equation model [42] for the random variables:

$$\begin{aligned} X &= g(Z_1, \ldots, Z_p, U, E_X), \text{ and} \\ Y &= f(X, Z_1, \ldots, Z_p, U, E_Y), \end{aligned} \tag{2.1}$$

where U is the unmeasured confounder, E_X and E_Y are independent random noises, $(E_X, E_Y) \perp\!\!\!\perp (Z_1, \cdots, Z_p, U)$ and $E_X \perp\!\!\!\perp E_Y$. In two-sample MR, we observe n_X i.i.d. realizations of $(X, Z_1, \ldots, Z_p)$ and independently n_Y

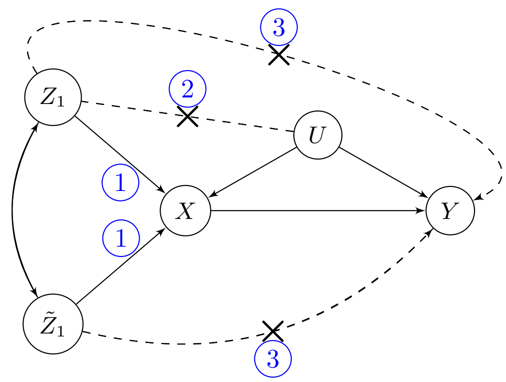


Fig 2: Causal DAG and the three criteria for valid IV. The proposed IV Z_1 can either be a causal variant for X or correlated with a causal variant. Z_1 must be independent of any unmeasured confounder U and cannot have any direct effect on Y or be correlated with another variant that has direct effect on Y.

i.i.d. realizations of $(Y, Z_1, \ldots, Z_p)$. We shall also assume that the SNPs $Z_1, Z_2, \ldots, Z_p$ are discrete random variables supported on $\{0, 1, 2\}$ and are mutually independent. To ensure the independence, in practice we only include SNPs with low pairwise LD score in our model by using standard genetics software like LD clumping [43].

A variable Z_j is called a *valid* IV if it satisfies the following three criteria:

1. Relevance: Z_j is associated with the exposure X. Notice that a SNP that is correlated (in genetics terminology, in LD) with the actual causal variant is also considered relevant and does not affect the statistical analysis below.
2. Effective random assignment: Z_j is independent of the unmeasured confounder U.
3. Exclusion restriction: Z_j only affects the outcome Y through the exposure X. In other words, the function f does not depend on Z_j.

The causal model and the IV conditions are illustrated by a directed acyclic graph (DAG) with a single instrument Z_1 in Figure 2. Readers who are unfamiliar with this language may find the tutorial by Baiocchi, Cheng and Small [3] helpful.

In Mendelian randomization, the first criterion—relevance—is easily satisfied by selecting SNPs that are significantly associated with X. Aside from the effects of population stratification, the second independence to unmeasured confounder assumption is usually easy to justify because most of the common confounders in epidemiology are postnatal, which are independent of genetic variants governed by Mendel's Second Law of independent assortment [18, 20]. Empirically, there is generally a lack of confounding of genetic variants with factors that confound exposures in conventional observational epidemiological studies [19].

The main concern for Mendelian randomization is the possible violation of the third exclusion restriction criterion, due to a genetic phenomenon called pleiotropy [18, 47], a.k.a. the multi-function of genes. The exclusion restriction assumption does not hold if a SNP Z_j affects the outcome Y through multiple causal pathways and some do not involve the exposure X. It is also violated if Z_j is correlated with other variants that affect Y through pathways that does not involve X. Pleiotropy is widely prevalent for complex traits [48]. In fact, a "universal pleiotropy hypothesis" developed by Fisher [22] and Wright [55] theorizes that every genetic mutation is capable of affecting essentially all traits. Recent genetics studies have found strong evidence that there is an extremely large number of causal variants with tiny effect sizes on many complex traits, which in part motivates our random effects Model 2.

Another important concept is the strength of an IV, defined as its association with the exposure X and usually measured by the F-statistic of an instrument-exposure regression. Since we assume all the genetic instruments are independent, the strength of SNP j can be assessed by comparing the statistic $\hat{\gamma}_j^2/\sigma_{Xj}^2$ with the quantiles of χ_1^2 (or equivalently $F_{1,\infty}$). When only a few weak instruments are available (e.g. F-statistic less than 10), the usual asymptotic inference is quite problematic [7]. In this paper, we primarily consider the setting where there is at least one strong IV or many weak IVs.

2.2. *Linear structural model.* We are now ready to derive the linear model (1.1) for GWAS summary data. Assuming all the IVs are valid, we start with the linear structural model where functions f and g in (2.1) are linear in their arguments (see also Bowden et al. [11]):

$$X = \sum_{j=1}^{p} \gamma_j Z_j + \eta_X U + E_X,\ Y = \beta X + \eta_Y U + E_Y. \tag{2.2}$$

In this case, the GWAS summary statistics $(\hat{\gamma}_j)_{j\in[p]}$ and $(\hat{\Gamma}_j)_{j\in[p]}$ are usually computed from simple linear regressions:

$$\hat{\gamma}_j = \frac{\widehat{\mathrm{Cov}}_{n_X}(X, Z_j)}{\widehat{\mathrm{Cov}}_{n_X}(Z_j, Z_j)},\ \hat{\Gamma}_j = \frac{\widehat{\mathrm{Cov}}_{n_Y}(Y, Z_j)}{\widehat{\mathrm{Cov}}_{n_Y}(Z_j, Z_j)}.$$

Here $\widehat{\mathrm{Cov}}_n$ is the sample covariance operator with n i.i.d. samples. Using (2.2), it is easy to show that $\hat{\gamma}_j$ and $\hat{\Gamma}_j$ converge to normal distributions centered at γ_j and $\Gamma_j = \beta\gamma_j$.

However, $\hat{\gamma}_j$ and $\hat{\gamma}_k$ are not exactly uncorrelated when $j \neq k$ (same for $\hat{\Gamma}_j$ and $\hat{\Gamma}_k$), even if Z_j and Z_k are independent. After some simple algebra, one can show that

$$\mathrm{Cor}^2(\hat{\gamma}_j, \hat{\gamma}_k) = 4 \cdot \frac{\gamma_j^2 \mathrm{Var}(Z_j)}{\mathrm{Var}(X) - \gamma_j^2 \mathrm{Var}(Z_j)} \frac{\gamma_k^2 \mathrm{Var}(Z_k)}{\mathrm{Var}(X) - \gamma_k^2 \mathrm{Var}(Z_k)}.$$

Notice that $\gamma_j^2 \mathrm{Var}(Z_j)/\mathrm{Var}(X)$ is the proportion of variance of X explained by Z_j. In the genetic context, a single SNP usually has very small predictability of a complex trait [13, 31, 41, 46]. Therefore the correlation between $\hat{\gamma}_j$ and $\hat{\gamma}_k$ (similarly $\hat{\Gamma}_j$ and $\hat{\Gamma}_k$) is almost negligible. In conclusion, the linear model (1.1) is approximately true when the phenotypes are believed to be generated from a linear structural model.

To stick to the main statistical methodology, we postpone additional justifications of (1.1) in nonlinear structural models to Appendix A. In Appendix A.1, we will investigate the case where Y is binary and $\hat{\Gamma}_j$ is obtained via logistic regression, as is very often the case in applied MR investigations. In Appendix A.2, we will show the linearity between X and $\boldsymbol{Z}$ is also not necessary.

2.3. *Violations of exclusion restriction.* Equation (2.2) assumes that all the instruments are valid. In reality, the exclusion restriction assumption is likely violated for many if not most of the SNPs. To investigate its impact in the model for summary data, we consider the following modification of the linear structural model (2.2):

$$X = \sum_{j=1}^{p} \gamma_j Z_j + \eta_X U + E_X,\ Y = \beta X + \sum_{j=1}^{p} \alpha_j Z_j + \eta_Y U + E_Y. \tag{2.3}$$

The difference between (2.2) and (2.3) is that the SNPs are now allowed to directly affect Y and the effect size of SNP Z_j is α_j. In this case, it is not difficult to see that the regression coefficient $\hat{\Gamma}_j$ estimates $\Gamma_j = \alpha_j + \gamma_j\beta$.

This inspires our Models 2 and 3. In Model 2, we assume the direct effects α_j are normally distributed random effects. In Model 3, we further require the statistical procedure to be robust against any extraordinarily large direct effects α_j. See Section 8 for more discussion on the assumptions on the pleiotropy effects.

3. No pleiotropy: A profile likelihood approach.

We now consider Model 1, the case with no pleiotropy effects.

3.1. *Derivation of the profile likelihood.* A good place to start is writing down the likelihood of GWAS summary data. Up to some additive constant, the log-likelihood function is given by

$$l(\beta, \gamma_1, \ldots, \gamma_p) = -\frac{1}{2}\left[\sum_{j=1}^{p} \frac{(\hat{\gamma}_j - \gamma_j)^2}{\sigma_{Xj}^2} + \sum_{j=1}^{p} \frac{(\hat{\Gamma}_j - \gamma_j \beta)^2}{\sigma_{Yj}^2}\right]. \tag{3.1}$$

Since we are only interested in estimating β_0, the other parameters, namely $\boldsymbol{\gamma} := (\gamma_1, \cdots, \gamma_p)$, are considered nuisance parameters. There are two ways to proceed from here. One is to view $\boldsymbol{\gamma}$ as *incidental* parameters [39] and try to eliminate them from the likelihood. The other approach is to assume the sequence $\gamma_1, \gamma_2, \cdots$ is generated from a fixed unknown distribution. When p is large, it is possible to estimate the distribution of $\boldsymbol{\gamma}$ to improve the efficiency using the second approach [38]. In this paper we aim to develop a general method for summary-data MR that can be used regardless of the number of SNPs being used, so we will take the first approach.

The profile log-likelihood of β is given by profiling out $\boldsymbol{\gamma}$ in (3.1):

$$l(\beta) = \max_{\boldsymbol{\gamma}} l(\beta, \boldsymbol{\gamma}) = -\frac{1}{2} \sum_{j=1}^{p} \frac{(\hat{\Gamma}_j - \beta \hat{\gamma}_j)^2}{\sigma_{Xj}^2 \beta^2 + \sigma_{Yj}^2}. \tag{3.2}$$

The maximum likelihood estimator of β is given by $\hat{\beta} = \arg\max_{\beta} l(\beta)$. It is also called a Limited Information Maximum Likelihood (LIML) estimator in the IV literature, a method due to Anderson and Rubin [2] with good consistency and efficiency properties. See also Pacini and Windmeijer [40].

Equation (3.2) can be interpreted as a linear regression of $\hat{\Gamma}$ on $\hat{\gamma}$, with the intercept of the regression fixed to zero and the variance of each observation equaling to $\sigma_{Xj}^2 \beta^2 + \sigma_{Yj}^2$. There is another meta-analysis interpretation. Let $\hat{\beta}_j = \hat{\Gamma}_j / \hat{\gamma}_j$ be the individual Wald's ratio, then (3.2) can be rewritten as

$$l(\beta) = -\frac{1}{2} \sum_{j=1}^{p} \frac{(\hat{\beta}_j - \beta)^2}{\sigma_{Xj}^2 \beta^2 / \hat{\gamma}_j^2 + \sigma_{Yj}^2 / \hat{\gamma}_j^2}. \tag{3.3}$$

This expression is also derived by Bowden et al. [10] by defining a generalized version of Cochran's Q statistic to test for the presence of pleiotropy that takes into account uncertainty in $\hat{\gamma}_j$.

3.2. *Consistency and asymptotic normality.* It is well known that the maximum likelihood estimator can be inconsistent when there are many nuisance parameters in the problem [e.g. 39]. Nevertheless, due to the connection with LIML, we expect and will prove below that $\hat{\beta}$ is consistent and asymptotically normal. However, we will also show that the profile likelihood (3.2) can be information biased [37], meaning the profile likelihood ratio test does not generally have a χ_1^2 limiting distribution under the null.

A major distinction between our asymptotic setting and the classical errors-in-variables regression setting is that our "predictors" $\hat{\gamma}_j$, $j \in [p]$ can be individually weak. This can be seen, for example, from the linear structural model (2.2) that

$$\operatorname{Var}(X) = \sum_{j=1}^{p} \gamma_j^2 \operatorname{Var}(Z_j) + \eta_X^2 \operatorname{Var}(U) + \operatorname{Var}(E_X). \tag{3.4}$$

Note that Z_j takes on the value $0, 1, 2$ with probability p_j^2, $2p_j(1-p_j)$, $(1-p_j)^2$ where p_j is the allele frequency of SNP j. For simplicity, we assume p_j is bounded away from 0 and 1. In other words, only common genetic variants are used as IVs. Together with (3.4), this implies that, if $\operatorname{Var}(X)$ exists, $\|\boldsymbol{\gamma}\|_2$ is bounded.

ASSUMPTION 2 (Collective IV strength is bounded). $\|\boldsymbol{\gamma}\|_2^2 = O(1)$.

As a consequence, the average effect size is decreasing to 0,

$$\frac{1}{p} \sum_{j=1}^{p} |\gamma_j| \le \|\boldsymbol{\gamma}\|_2 / \sqrt{p} \to 0, \text{ when } p \to \infty.$$

This is clearly different from the usual linear regression setting where the "predictors" $\hat{\gamma}_j$ are viewed as random samples from a population. In the one-sample IV literature, this many weak IV setting ($p \to \infty$) has been considered by Bekker [5], Stock and Yogo [49], Hansen, Hausman and Newey [26] among many others in econometrics.

Another difference between our asymptotic setting and the errors-in-variables regression is that our measurement errors also converge to 0 as the sample size converges to infinity. Recall that n_X is the sample size of $(X, Z_1, \ldots, Z_p)$ and n_Y is the sample size of $(Y, Z_1, \ldots, Z_p)$. We assume

ASSUMPTION 3 (Variance of measurement error). *Let* $n = \min(n_X, n_Y)$. *There exist constants* c_σ, c'_σ *such that* $c_\sigma/n \le \sigma^2_{Xj} \le c'_\sigma/n$ *and* $c_\sigma/n \le \sigma^2_{Yj} \le c'_\sigma/n$ *for all* $j \in [p]$.

We write $a = O(b)$ if there exists a constant $c > 0$ such that $|a| \le cb$, and $a = \Theta(b)$ if there exists $c > 0$ such that $c^{-1}b \le |a| \le cb$. In this notation, Assumption 3 assumes the known variances σ^2_{Xj} and σ^2_{Yj} are $\Theta(1/n)$.

In the linear structural model (2.2), $\mathrm{Var}(\hat{\gamma}_j) \le \mathrm{Var}(X)/[\mathrm{Var}(Z_j)/n_X]$. Thus Assumption 3 is satisfied when only common variants are used.

We are ready to to state our first theoretical result.

THEOREM 3.1. *In Model 1 and under Assumptions 1 to 3, if* $p/(n^2\|\boldsymbol{\gamma}\|_2^4) \to 0$*, the maximum likelihood estimator* $\hat{\beta}$ *is statistically consistent, i.e.* $\hat{\beta} \xrightarrow{p} \beta_0$.

A crucial quantity in Theorem 3.1 and the analysis below is the average strength of the IVs, defined as

$$\kappa = \frac{1}{p}\sum_{j=1}^{p} \frac{\gamma_j^2}{\sigma^2_{Xj}} = \Theta(n\|\boldsymbol{\gamma}\|_2^2/p).$$

An unbiased estimator of κ is the average F-statistic minus 1,

$$\hat{\kappa} = \frac{1}{p}\sum_{j=1}^{p} \frac{\hat{\gamma}_j^2}{\sigma^2_{Xj}} - 1.$$

In practice, we require that the average F-statistic to be large (say > 100) when p is small, or not too small (say > 3) when p is large. Thus the condition $p/(n^2\|\boldsymbol{\gamma}\|_2^4) = 1/(p\kappa^2) \to 0$ in Theorem 3.1 is usually quite reasonable.

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

$$\psi(\beta) := -l'(\beta) = \sum_{j=1}^{p} \frac{(\hat{\Gamma}_j - \beta\hat{\gamma}_j)(\hat{\Gamma}_j\sigma^2_{Xj}\beta + \hat{\gamma}_j\sigma^2_{Yj})}{(\sigma^2_{Xj}\beta^2 + \sigma^2_{Yj})^2}. \tag{3.5}$$

The maximum likelihood estimator $\hat{\beta}$ solves the estimating equation $\psi(\hat{\beta}) = 0$, and we consider the Taylor expansion around the truth β_0:

$$0 = \psi(\hat{\beta}) = \psi(\beta_0) + \psi'(\beta_0)(\hat{\beta} - \beta_0) + \frac{1}{2}\psi''(\tilde{\beta})(\hat{\beta} - \beta_0)^2, \tag{3.6}$$

where $\tilde{\beta}$ is between $\hat{\beta}$ and β_0. Since $\hat{\beta}$ is statistically consistent, the last term on the right hand side of (3.6) can be proved to be negligible, and the asymptotic normality of $\hat{\beta}$ can be established by showing, for some appropriate V_1 and V_2, $\psi(\beta_0) \xrightarrow{d} \mathrm{N}(0, V_1)$ and $\psi'(\beta_0) \xrightarrow{p} -V_2$. When $V_1 = V_2$, the profile likelihood/score is called *information unbiased* [37].

THEOREM 3.2. *Under the assumptions in Theorem 3.1 and if at least one of the following two conditions are true: (1) $p \to \infty$ and $\|\boldsymbol{\gamma}\|_3/\|\boldsymbol{\gamma}\|_2 \to 0$; (2) $\kappa \to \infty$; then we have*

$$\frac{V_2}{\sqrt{V_1}}(\hat{\beta} - \beta_0) \xrightarrow{d} \mathrm{N}(0,1), \tag{3.7}$$

where

$$V_1 = \sum_{j=1}^{p} \frac{\gamma_j^2 \sigma_{Yj}^2 + \Gamma_j^2 \sigma_{Xj}^2 + \sigma_{Xj}^2 \sigma_{Yj}^2}{(\sigma_{Xj}^2 \beta_0^2 + \sigma_{Yj}^2)^2}, \; V_2 = \sum_{j=1}^{p} \frac{\gamma_j^2 \sigma_{Yj}^2 + \Gamma_j^2 \sigma_{Xj}^2}{(\sigma_{Xj}^2 \beta_0^2 + \sigma_{Yj}^2)^2}. \tag{3.8}$$

Notice that Theorem 3.2 is very general. It can be applied even in the extreme situation p is fixed and $\kappa \to \infty$ (a few strong IVs) or $p \to \infty$ and $\kappa \to 0$ (many very weak IVs). The assumption $\|\boldsymbol{\gamma}\|_3/\|\boldsymbol{\gamma}\|_2 \to 0$ is used to verify a Lyapunov's condition for a central limit theorem. It essentially says the distribution of IV strengths is not too uneven and this assumption can be further relaxed.

Using our rate assumption for the variances (Assumption 3), $V_2 = \Theta(n\|\boldsymbol{\gamma}\|_2^2) = \Theta(p\kappa)$ and $V_1 = V_2 + \Theta(p)$. This suggests that the profile likelihood is information unbiased if and only if $\kappa \to \infty$. In general, the amount of information bias depends on the instrument strength κ. As an example, suppose $\beta_0 = 0$ and $\sigma_{Yj}^2 \equiv \sigma_{Y1}^2$. Then by (3.7) and (3.8), $\mathrm{Var}(\hat{\beta}) \approx V_1/V_2^2 = (1 + \kappa^{-1})/V_2$. Alternatively, if we make the simplifying assumption that $\sigma_{Yj}^2/\sigma_{Xj}^2$ does not depend on j, it is straightforward to show that

$$\mathrm{Var}(\hat{\beta}) \propto \frac{1 + \kappa^{-1}}{p\kappa}.$$

This approximation can be used as a rule of thumb to select the optimal number of IVs.

In order to obtain standard error of $\hat{\beta}$, we must estimate V_1 and V_2 using the GWAS summary data. We propose to replace γ_j^2 and Γ_j^2 in (3.8) by their unbiased sample estimates, $\hat{\gamma}_j^2 - \sigma_{Xj}^2$ and $\hat{\Gamma}_j^2 - \sigma_{Yj}^2$:

$$\hat{V}_1 = \sum_{j=1}^{p} \frac{(\hat{\gamma}_j^2 - \sigma_{Xj}^2)\sigma_{Yj}^2 + (\hat{\Gamma}_j^2 - \sigma_{Yj}^2)\sigma_{Xj}^2 + \sigma_{Xj}^2\sigma_{Yj}^2}{(\sigma_{Xj}^2\hat{\beta}^2 + \sigma_{Yj}^2)^2},$$
$$\hat{V}_2 = \sum_{j=1}^{p} \frac{(\hat{\gamma}_j^2 - \sigma_{Xj}^2)\sigma_{Yj}^2 + (\hat{\Gamma}_j^2 - \sigma_{Yj}^2)\sigma_{Xj}^2}{(\sigma_{Xj}^2\hat{\beta}^2 + \sigma_{Yj}^2)^2}.$$

THEOREM 3.3. *Under the same assumptions in Theorem 3.2, we have* $\hat{V}_1 = V_1(1 + o_p(1))$, $\hat{V}_2 = V_2(1 + o_p(1))$, *and*

$$\frac{\hat{V}_2}{\sqrt{\hat{V}_1}}(\hat{\beta} - \beta_0) \xrightarrow{d} \mathrm{N}(0,1) \text{ as } n \to \infty. \tag{3.9}$$

3.3. *Practical issues.* Next we discuss several practical implications of the theoretical results above.

3.3.1. *Influence of a single IV.* Under the assumptions in Theorem 3.2, (3.6) and (3.5) lead to the following asymptotically linear form of $\hat{\beta}$:

$$\hat{\beta} = \frac{1 + o_p(1)}{V_2} \sum_{j=1}^{p} \frac{(\hat{\Gamma}_j - \beta_0\hat{\gamma}_j)(\hat{\Gamma}_j\sigma_{Xj}^2\beta_0 + \hat{\gamma}_j\sigma_{Yj}^2)}{(\sigma_{Xj}^2\beta_0^2 + \sigma_{Yj}^2)^2}.$$

The above equation characterizes the influence of a single IV on the estimator $\hat{\beta}$ [25]. Intuitively, the IV Z_j has large influence if it is strong or it has large residual $\hat{\Gamma}_j - \beta_0\hat{\gamma}_j$. Alternatively, we can measure the influence of a single IV by computing the leave-one-out estimator $\hat{\beta}_{-j}$ that maximizes the profile likelihood with all the SNPs except Z_j. In practice, it is desirable to limit the influence of each SNP to make the estimator robust against idiosyncratic pleiotropy (Model 3). This problem will be considered in Section 5.

3.3.2. *Selecting IVs.* The formulas (3.7) and (3.8) suggests that using extremely weak instruments may deteriorate the efficiency. Consider the following example in which we have a new instrument Z_{p+1} that is independent of X, so $\gamma_{p+1} = 0$. When adding Z_{p+1} to the analysis, V_1 increases but V_2 remains the same, thus the variance of $\hat{\beta}$ becomes larger. Generally, this suggests that we should screen out extremely weak IVs to improve efficiency. To avoid selection bias, we recommend to use two independent GWAS datasets in practice, one to screen out weak IVs and perform LD clumping and one to esimate the SNP-exposure effects γ_j unbiasedly.

3.3.3. *Residual quantile-quantile plot.* One way to check the modeling assumptions in Assumption 1 and Model 1 is the residual Quantile-Quantile (Q-Q) plot, which plots the quantiles of standardized residuals

$$\hat{t}_j = \frac{\hat{\Gamma}_j - \hat{\beta}\hat{\gamma}_j}{\sqrt{\hat{\beta}^2\sigma_{Xj}^2 + \sigma_{Yj}^2}}$$

against the quantiles of the standard normal distribution. This is reasonable because when $\hat{\beta} = \beta_0$, $\hat{t}_j \sim \mathrm{N}(0, 1)$ under Assumption 1 and Model 1. The Q-Q plot is helpful at identifying IVs that do not satisfy the linear relation $\Gamma_j = \beta_0\gamma_j$, most likely due to genetic pleiotropy.

Besides the residual Q-Q plot, other diagnostic tools can be found in related works. Bowden et al. [10] considered using each SNP's contribution to the generalized Q statistic to assess whether it is an outlier. Bowden et al. [12] proposed a radial plot $\hat{\beta}_j\sqrt{w_j}$ versus $\sqrt{w_j}$, where w_j is the "weight" of the j-th SNP in (3.3). Since these diagnostic methods are based on the Wald ratio estimates $\hat{\beta}_j$, they can suffer from the weak instrument bias.

3.4. *Example (continued).* We conclude this Section by applying the profile likelihood or Profile Score (PS) estimator in the BMI-SBP example in Section 1.2. Here we used 160 SNPs that have p-values $\leq 10^{-4}$ in the `BMI-FEM` dataset. The PS point estimate is 0.601 with standard error 0.054.

Figure 3 shows the Q-Q plot and the leave-one-out estimates discussed in Section 3.3. The Q-Q plot clearly indicates the linear model Model 1 is not appropriate to describe the summary data. Although the standardized residuals are roughly normally distributed, their standard deviations are apparently larger than 1. This motivates the random pleiotropy effects assumption in Model 2 which will be considered next.

4. Systematic pleiotropy: Adjusted profile score.

4.1. *Failure of the profile likelihood.* Next we consider Model 2, where the deviation from the linear relation $\Gamma_j = \beta_0\gamma_j$ is described by a random effects model $\alpha_j = \Gamma_j - \beta_0\gamma_j \sim \mathrm{N}(0, \tau_0^2)$. In other words, the variance of $\hat{\Gamma}$ is inflated by an unknown additive constant τ_0^2:

$$\hat{\gamma}_j \sim \mathrm{N}(\gamma_j, \sigma_{Xj}^2),\ \hat{\Gamma}_j \sim \mathrm{N}(\gamma_j\beta_0, \sigma_{Yj}^2 + \tau_0^2),\ j \in [p].$$

Similar to Section 3.1, the profile log-likelihood of (β, τ^2) is given by

$$l(\beta, \tau^2) = -\frac{1}{2}\sum_{j=1}^{p} \frac{(\hat{\Gamma}_j - \beta\hat{\gamma}_j)^2}{\sigma_{Xj}^2\beta^2 + \sigma_{Yj}^2 + \tau^2} + \log(\sigma_{Yj}^2 + \tau^2),$$

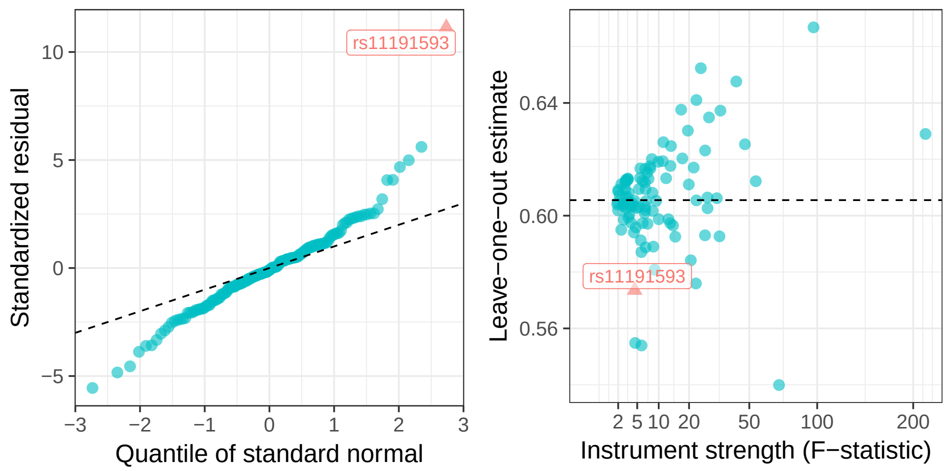


Fig 3: Diagnostic plots of the Profile Score (PS) estimator. Left panel is a Q-Q plot of the standardized residuals against standard normal. Right panel is the leave-one-out estimates against instrument strength.

and the corresponding profile score equations are

$$\frac{\partial}{\partial \beta} l(\beta, \tau^2) = 0, \ \frac{\partial}{\partial \tau^2} l(\beta, \tau^2) = 0.$$

It is straightforward to verify that the first estimating equation is unbiased, i.e. it has expectation 0 at (β_0, τ_0^2). However, the other profile score is

$$\frac{\partial}{\partial \tau^2} l(\beta, \tau^2) = \frac{1}{2} \sum_{j=1}^{p} \frac{(\hat{\Gamma}_j - \beta \hat{\gamma}_j)^2}{(\sigma_{Xj}^2 \beta^2 + \sigma_{Yj}^2 + \tau^2)^2} - \frac{1}{\sigma_{Yj}^2 + \tau^2}. \tag{4.1}$$

It is easy to see that its expectation is not equal to 0 at the true value $(\beta, \tau^2) = (\beta_0, \tau_0^2)$. This means the profile score is biased in Model 2, thus the corresponding maximum likelihood estimator is not statistically consistent.

4.2. *Adjusted profile score.* The failure of maximizing the profile likelihood should not be surprising, because it is well known that maximum likelihood estimator can be biased when there are many nuisance parameters [39]. There are many proposals to modify the profile likelihood, see, for example, Barndorff-Nielsen [4], Cox and Reid [17]. Here we take the approach of McCullagh and Tibshirani [37] that directly modifies the profile score so it has mean 0 at the true value. The *Adjusted Profile Score (APS)* is given by $\boldsymbol{\psi}(\beta,\tau^2) = (\psi_1(\beta,\tau^2), \psi_2(\beta,\tau^2))$, where

$$\psi_1(\beta,\tau^2) = -\frac{\partial}{\partial\beta} l(\beta,\tau^2) = \sum_{j=1}^{p} \frac{(\hat{\Gamma}_j - \beta\hat{\gamma}_j)(\hat{\Gamma}_j\sigma_{Xj}^2\beta + \hat{\gamma}_j(\sigma_{Yj}^2+\tau^2))}{(\sigma_{Xj}^2\beta^2 + \sigma_{Yj}^2 + \tau^2)^2}, \tag{4.2}$$

$$\psi_2(\beta,\tau^2) = \sum_{j=1}^{p} \sigma_{Xj}^2 \frac{(\hat{\Gamma}_j - \beta\hat{\gamma}_j)^2 - (\sigma_{Xj}^2\beta^2 + \sigma_{Yj}^2 + \tau^2)}{(\sigma_{Xj}^2\beta^2 + \sigma_{Yj}^2 + \tau^2)^2}. \tag{4.3}$$

Compared to (4.1), we replaced $(\sigma_{Yj}^2 + \tau^2)^{-1}$ by $(\sigma_{Xj}^2\beta^2 + \sigma_{Yj}^2 + \tau^2)^{-1}$, so each summand in (4.3) has mean 0 at (β_0, τ_0^2). We also weighted the IVs by σ_{Xj}^2 in (4.3), which is useful in the proof of statistical consistency.

Notice that both the denominators and numerators in ψ_1 and ψ_2 are polynomials of β and τ^2. However, the denominators are of higher degrees. This implies that the APS estimating equations always have diverging solutions: $\boldsymbol{\psi}(\beta,\tau^2) \to \mathbf{0}$ if $\beta \to \pm\infty$ or $\tau^2 \to \infty$. We define the APS estimator $(\hat{\beta}, \hat{\tau}^2)$ to be the non-trivial finite solution to $\boldsymbol{\psi}(\beta,\tau^2) = \mathbf{0}$ if it exists.

4.3. *Consistency and asymptotic normality.* Because of the diverging solutions of the APS equations, we need to impose some compactness constraints on the parameter space to study the asymptotic property of $(\hat{\beta}, \hat{\tau}^2)$:

ASSUMPTION 4. *$(\beta_0, p\tau_0^2)$ is in the interior of a bounded set $\mathcal{B} \subset \mathbb{R} \times \mathbb{R}^+$.*

The overdispersion parameter τ_0^2 is scaled up in Assumption 4 by p. This is motivated by the linear structural model (2.3), where $\sum_{j=1}^{2} \tau_0^2 \mathrm{Var}(Z_j) = \Theta(p\tau_0^2)$ is the variance of Y explained by the direct effects of $\boldsymbol{Z}$. Thus it is reasonable to treat $p\tau_0^2$ as a constant.

We also assume, in addition to Assumption 2, that the variance of X explained by the IVs is non-diminishing:

ASSUMPTION 5. $\|\boldsymbol{\gamma}\|_2 = \Theta(1)$.

THEOREM 4.1. *In Model 2 and suppose Assumptions 1 and 3 to 5 hold, $p \to \infty$ and $p/n^2 \to 0$. Then with probability going to 1 there exists a solution of the APS equation such that $(\hat{\beta}, p\hat{\tau}^2)$ is in $\mathcal{B}$. Furthermore, all solutions in $\mathcal{B}$ are statistically consistent, i.e. $\hat{\beta} \xrightarrow{p} \beta_0$ and $p\hat{\tau}^2 - p\tau_0^2 \xrightarrow{p} 0$.*

Next we consider the asymptotic distribution of the APS estimator.

THEOREM 4.2. *In Model 2 and under the assumptions in Theorem 4.1, if additionally $p = \Theta(n)$ and $\|\boldsymbol{\gamma}\|_3/\|\boldsymbol{\gamma}\|_2 \to 0$, then*

$$\left(\tilde{\boldsymbol{V}}_2^{-1}\tilde{\boldsymbol{V}}_1\tilde{\boldsymbol{V}}_2^{-T}\right)^{1/2}\begin{pmatrix}\hat{\beta}-\beta_0\\ \hat{\tau}^2-\tau_0^2\end{pmatrix} \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_2), \tag{4.4}$$

where

$$\tilde{\boldsymbol{V}}_1 = \sum_{j=1}^{p} \frac{1}{(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)^2}\begin{pmatrix}(\gamma_j^2+\sigma_{Xj}^2)(\sigma_{Yj}^2+\tau_0^2)+\Gamma_j^2\sigma_{Xj}^2 & 0\\ 0 & 2(\sigma_{Xj}^2)^2\end{pmatrix},$$

$$\tilde{\boldsymbol{V}}_2 = \sum_{j=1}^{p} \frac{1}{(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)^2}\begin{pmatrix}\gamma_j^2(\sigma_{Yj}^2+\tau_0^2)+\Gamma_j^2\sigma_{Xj}^2 & \sigma_{Xj}^2\beta_0\\ 0 & \sigma_{Xj}^2\end{pmatrix}.$$

Similar to Theorem 3.3, the information matrices $\tilde{\boldsymbol{V}}_1$ and $\tilde{\boldsymbol{V}}_2$ can be estimated by substituting γ_j^2 by $\hat{\gamma}_j^2 - \sigma_{Xj}^2$ and Γ_j^2 by $\hat{\Gamma}_j^2 - \sigma_{Yj}^2 - \hat{\tau}^2$. We omit the details for brevity.

4.4. *Example (continued).* We apply the APS estimator to the BMI-SBP example. Using the same 160 SNPs in Section 3.4, the APS point estimate is $\hat{\beta} = 0.301$ (standard error 0.158) and $\hat{\tau}^2 = 9.2 \times 10^{-4}$ (standard error 1.7×10^{-4}). Notice that the APS point estimate of β is much smaller than the PS point estimate. One possible explanation of this phenomenon is that the PS estimator tends to use a larger β to compensate for the overdispersion in Model 2 (the variance of $\hat{\Gamma}_j - \beta\hat{\gamma}_j$ is $\beta^2\sigma_{Xj}^2 + \sigma_{Yj}^2$ in Model 1 and $\beta^2\sigma_{Xj}^2 + \sigma_{Yj}^2 + \tau_0^2$ in Model 2).

Figure 4 shows the diagnostic plots of the APS estimator. Compared to the PS estimator in Section 3.4, the overdispersion issue is much more benign. However, there is an outlier which corresponds to the SNP `rs11191593`. It heavily biases the APS estimate too: when excluding this SNP, the APS point estimate changes from 0.301 to almost 0.4 in the right panel of Figure 4. The outlier might also inflate $\hat{\tau}^2$ so the Q-Q plot looks a little underdispersed. These observations motivate the consideration of a robust modification of the APS in the next Section.

5. Idiosyncratic pleiotropy: Robustness to outliers. Next we consider Model 3 with idiosyncratic pleiotropy. As mentioned in Section 3.3.1, a single IV can have unbounded influence on the PS (and APS) estimators. When the IV Z_j has other strong causal pathways, its pleiotropy parameter α_j can be much larger than what is predicted by the random effects model

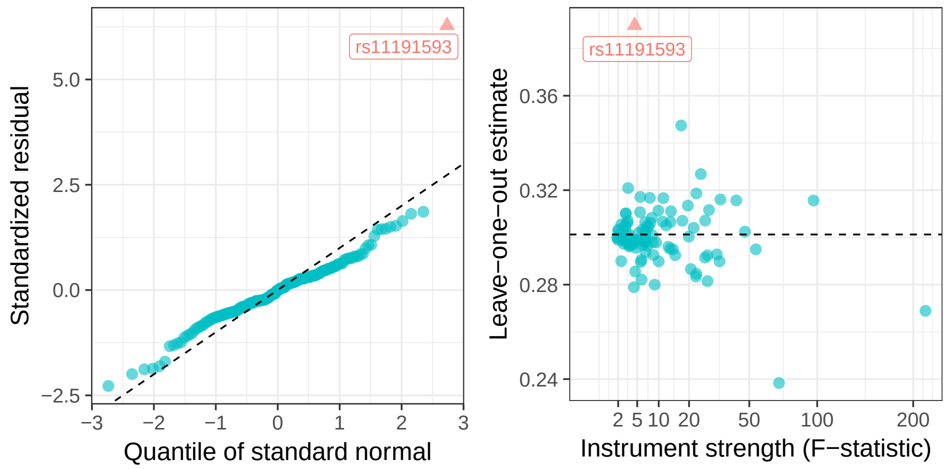


Fig 4: Diagnostic plots of the Adjusted Profile Score (APS) estimator. Left panel is a Q-Q plot of the standardized residuals against standard normal. Right panel is the leave-one-out estimates against instrument strength.

$\alpha_j \sim \mathrm{N}(0, \tau_0^2)$, leading to a biased estimate of the causal effect as illustrated in Section 4.4. In this Section, we propose a general method to robustify the APS to limit the influence of outliers such as SNP `rs11191593` in the example.

5.1. *Robustify the adjusted profile score.* Our approach is an application of the robust regression techniques pioneered by Huber [30]. As mentioned in Section 3.1, the profile likelihood (3.2) can be viewed as a linear regression of $\hat{\Gamma}_j$ on $\hat{\gamma}_j$ using the l_2-loss. To limit the influence of a single IV, we consider changing the l_2-loss to a robust loss function. Two celebrated examples are the Huber loss

$$\rho_{\mathrm{huber}}(r;k) = \begin{cases} r^2/2, & \text{if } |r| \le k, \\ k(|r| - k/2), & \text{otherwise,} \end{cases}$$

and Tukey's biweight loss

$$\rho_{\mathrm{tukey}}(r;k) = \begin{cases} 1 - (1 - (r/k)^2)^3, & \text{if } |r| \le k, \\ 1, & \text{otherwise.} \end{cases}$$

This heuristic motivates the following modification of the profile log-likelihood when $\tau_0^2 = 0$:

$$l_\rho(\beta) := -\sum_{j=1}^{p} \rho\left(\frac{\hat{\Gamma}_j - \beta\hat{\gamma}_j}{\sqrt{\sigma_{Xj}^2\beta^2 + \sigma_{Yj}^2}}\right) \tag{5.1}$$

It is easy to see that $l_\rho(\beta)$ reduces to the regular profile log-likelihood (3.2) if $\rho(r) = r^2/2$.

When $\tau_0^2 > 0$, we cannot directly use the profile score $(\partial/\partial\tau^2)l(\beta,\tau^2)$ as discussed in Section 4.1. This issue can be resolved using the APS approach in Section 4.2 by using ψ_2 in (4.3). However, a single IV can still have unbounded influence in ψ_2. We must further robustify ψ_2, which is analogous to estimating a scale parameter robustly.

Next we briefly review the robust M-estimation of scale parameter. Consider repeated measurements of a scale family with density $f_0(r/\sigma)/\sigma$. Then a general way of robust estimation of σ is to solve the following estimating equation [36, Section 2.5]

$$\hat{\mathbb{E}}[(R/\sigma) \cdot \rho'(R/\sigma)] = \delta,$$

where $\hat{\mathbb{E}}$ stands for the empirical average and $\delta = \mathbb{E}[R \cdot \rho'(R)]$ for $R \sim f_0$.

Based on the above discussion, we propose the following *Robust Adjusted Profile Score (RAPS)* estimator of β. Denote

$$t_j(\beta,\tau^2) = \frac{\hat{\Gamma}_j - \beta\hat{\gamma}_j}{\sqrt{\sigma_{Xj}^2\beta^2 + \sigma_{Yj}^2 + \tau^2}}.$$

Then the RAPS $\boldsymbol{\psi}^{(\rho)} = (\psi_1^{(\rho)}, \psi_2^{(\rho)})$ is given by

$$\psi_1^{(\rho)}(\beta,\tau^2) = \sum_{j=1}^{p} \rho'(t_j(\beta,\tau^2))u_j(\beta,\tau^2), \tag{5.2}$$

$$\psi_2^{(\rho)}(\beta,\tau^2) = \sum_{j=1}^{p} \sigma_{Xj}^2 \frac{t_j(\beta,\tau^2)\cdot\rho'(t_j(\beta,\tau^2)) - \delta}{\sigma_{Xj}^2\beta^2 + \sigma_{Yj}^2 + \tau^2}, \tag{5.3}$$

where $\rho'(\cdot)$ is the derivative of $\rho(\cdot)$, $u_j(\beta,\tau^2) = -(\partial/\partial\beta)t_j(\beta,\tau^2)$ and $\delta = \mathbb{E}[R\cdot\rho'(R)]$ for $R \sim \mathrm{N}(0,1)$. Notice that $\boldsymbol{\psi}^{(\rho)}$ reduces to the non-robust APS $\boldsymbol{\psi}$ in (4.2) and (4.3) when $\rho(r) = r^2/2$ is the squared error loss. Finally, the RAPS estimator $(\hat{\beta},\hat{\tau}^2)$ is given by the non-trivial finite solution of $\boldsymbol{\psi}^{(\rho)}(\beta,\tau^2) = \mathbf{0}$.

5.2. *Asymptotics.* Because the RAPS estimator is the solution of a system of nonlinear equations, its asymptotic behavior is very difficult to analyze. For instance, it is difficult to establish statistical consistency because there could be multiple roots for the RAPS equations in the population level. Thus β might not be globally identified. We can, nevertheless, verify the local identifiability [45]:

THEOREM 5.1 (Local identification of RAPS). *In Model 2,* $\mathbb{E}[\boldsymbol{\psi}^{(\rho)}(\beta_0, \tau_0^2)] = \mathbf{0}$ *and* $\mathrm{E}[\nabla \boldsymbol{\psi}^{(\rho)}]$ *has full rank.*

In practice, we find that the RAPS estimating equation usually only has one finite solution. To study the asymptotic normality of the RAPS estimator, we will assume $(\hat{\beta}, p\hat{\tau}^2)$ is consistent under Model 2. We further impose the following smoothness condition on the robust loss function ρ:

ASSUMPTION 6. *The first three derivatives of* $\rho(\cdot)$ *exist and are bounded.*

THEOREM 5.2. *In Model 2 and under the assumptions in Theorem 4.2, if additionally we assume*

1. *the RAPS estimator is consistent:* $\hat{\beta} - \beta_0 \xrightarrow{p} 0$, $p(\hat{\tau}^2 - \tau_0^2) \xrightarrow{p} 0$,
2. *Assumption 6 holds, and*
3. $\|\boldsymbol{\gamma}\|_3^3 / \|\boldsymbol{\gamma}\|_2^3 = O(p^{-1/2})$,

then

$$\left((\tilde{\boldsymbol{V}}_2^{(\rho)})^{-1} \tilde{\boldsymbol{V}}_1^{(\rho)} (\tilde{\boldsymbol{V}}_2^{(\rho)})^{-T}\right)^{1/2} \begin{pmatrix} \hat{\beta} - \beta_0 \\ \hat{\tau}^2 - \tau_0^2 \end{pmatrix} \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{I}_2), \tag{5.4}$$

where

$$\tilde{\boldsymbol{V}}_1^{(\rho)} = \begin{pmatrix} c_1 (\tilde{\boldsymbol{V}}_1)_{11} & 0 \\ 0 & c_2 (\tilde{\boldsymbol{V}}_1)_{22} \end{pmatrix},$$
$$\tilde{\boldsymbol{V}}_2^{(\rho)} = \begin{pmatrix} \delta (\tilde{\boldsymbol{V}}_2)_{11} & \delta (\tilde{\boldsymbol{V}}_2)_{12} \\ 0 & [(\delta + c_3)/2] (\tilde{\boldsymbol{V}}_2)_{22} \end{pmatrix},$$

and the constants are: for $R \sim \mathrm{N}(0,1)$, $c_1 = \mathbb{E}[\rho'(R)^2]$, $c_2 = \mathrm{Var}(R\rho'(R))/2$, $c_3 = \mathbb{E}[R^2 \rho''(R)]$.

It is easy to verify that when $\rho(r) = r^2/2$, $\delta = c_1 = c_2 = c_3 = 1$, so $\tilde{\boldsymbol{V}}_1^{(\rho)}$ and $\tilde{\boldsymbol{V}}_2^{(\rho)}$ reduce to $\tilde{\boldsymbol{V}}_1$ and $\tilde{\boldsymbol{V}}_2$. In other words, the asymptotic variance formula in Theorem 5.2 is consistent with the one in Theorem 4.2. However, additional technical assumptions are needed in Theorem 5.2 to bound the higher-order terms in the Taylor expansion.

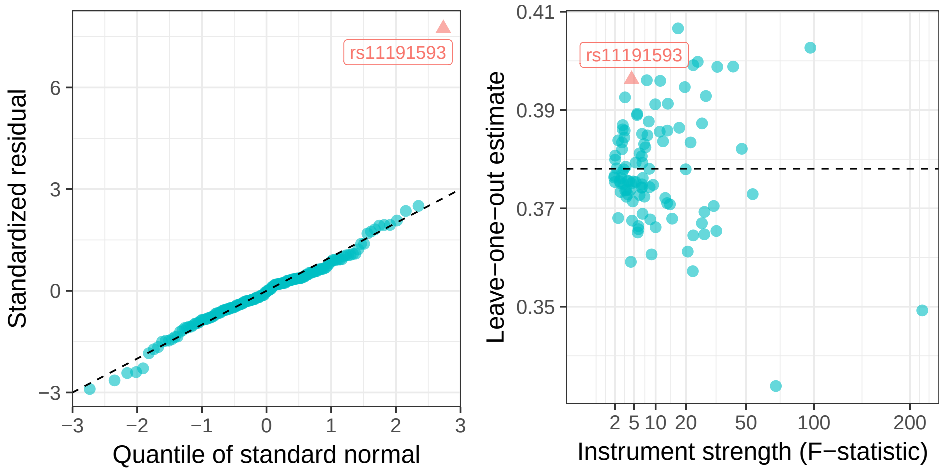
5.3. *Example (continued).* As before, we illustrate the RAPS estimator using the BMI-SBP example. Using the Huber loss with $k = 1.345$ (corresponding to 95% asymptotic efficiency in the simple location problem), the point estimate is $\hat{\beta} = 0.378$ (standard error 0.121), $\hat{\tau}^2 = 4.7 \times 10^{-4}$ (standard error 1.0×10^{-4}). Using the Tukey loss with $k = 4.685$ (also corresponding to 95% asymptotic efficiency in the simple location problem), the point estimate is $\hat{\beta} = 0.402$ (standard error 0.106), $\hat{\tau}^2 = 3.4 \times 10^{-4}$ (standard error 7.8×10^{-5}).

Figure 5 shows the diagnostic plots of the two RAPS estimators. Compared to Figure 4, the robust loss functions limit the influence of the outlier (SNP `rs11191593`), and the resulting $\hat{\beta}$ becomes larger. In Figure 5b, the outlier's influence is essentially zero because the Tukey loss function is redescending. This shows the robustness of our RAPS estimator to the idiosyncratic pleiotropy.

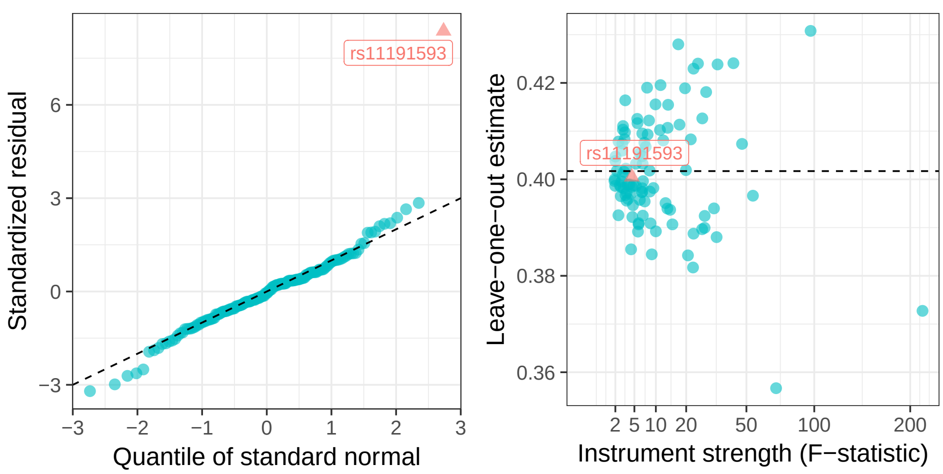
6. Simulation. Throughout the paper all of our theoretical results are asymptotic. We usually require both the sample size n and the number of IVs p to go to infinity (except for Theorem 3.2 where finite p is allowed). We now assess if the asymptotic approximations are reasonably accurate in practical situations, where p may range from tens to hundreds.

To this end, we created simulated datasets that mimic the BMI-SBP example introduced in Section 1.2. In particular, we considered two scenarios: $p = 25$, which corresponds to using the selection threshold 5×10^8 as described in Section 1.2, and $p = 160$, which corresponds to using the threshold 1×10^{-4} as in Sections 3.4, 4.4 and 5.3. The model parameters are chosen as follows: the variances of the measurement error, $\{(\sigma_{Xj}^2, \sigma_{Yj}^2)\}_{j\in[p]}$, are the same as those in the BMI-SBP dataset. The true IV-exposure effects, $\{\gamma_j\}_{j\in[p]}$, are chosen to be the observed effects in the BMI-SBP dataset, and $\hat{\gamma}_j$ is generated according to Assumption 1 by $\hat{\gamma}_j \overset{ind.}{\sim} \mathrm{N}(\gamma_j, \sigma_{Xj}^2)$. The true IV-outcome effects, $\{\Gamma_j\}_{j\in[p]}$, are generated in six different ways with $\beta_0 = 0.4$:

1. $\Gamma_j = \gamma_j\beta_0$;
2. $\Gamma_j = \gamma_j\beta_0 + \alpha_j$, $\alpha_j \overset{i.i.d.}{\sim} \mathrm{N}(0, \tau_0^2)$, where $\tau_0 = 2 \cdot (1/p)\sum_{j=1}^p \sigma_{Yj}$;
3. Γ_j is generated according to setup 2 above, except that α_1 has mean $5 \cdot \tau_0$ (the IVs are sorted so that the first IV has the largest $|\gamma_j|/\sigma_{Xj}$).
4. $\Gamma_j = \gamma_j\beta_0 + \alpha_j$, $\alpha_j \overset{i.i.d.}{\sim} \tau_0 \cdot \mathrm{Lap}(1)$, where $\mathrm{Lap}(1)$ is the Laplace (double exponential) distribution with rate 1.
5. $\Gamma_j = \gamma_j\beta_0 + \alpha_j$, $\alpha_j = |\gamma_j|/(p^{-1}\sum_{j=1}^p |\gamma_j|) \cdot \mathrm{N}(0, \tau_0^2)$.
6. Γ_j is generated according to setup 2 above, except that for 10% ran-



(a) RAPS using the Huber loss.



(b) RAPS using the Tukey loss.

Fig 5: Diagnostic plots of the Robust Adjusted Profile Score (RAPS) estimator. Left panels are Q-Q plots of the standardized residuals against standard normal. Right panels are the leave-one-out estimates against instrument strength.

domly selected IVs, their direct effects α_j have mean $5 \cdot \tau_0$.

The first three setups correspond to Models 1 to 3, respectively, and the last three setups violate our modeling assumptions and are used to assess the robustness of the procedures. Finally, $\hat{\Gamma}_j$ is generated according to Assumption 1 by $\hat{\Gamma}_j \overset{ind.}{\sim} \mathrm{N}(\Gamma_j, \sigma_{Yj}^2)$.

We applied six methods to the simulated data (10,000 replications in each setting). The first three are existing methods to benchmark our performance: the inverse variance weighting (IVW) estimator [14], MR-Egger regression [8], and the weighted median estimator [9]. The next three methods are proposed in this paper: the profile score (PS) estimator in Section 3, the adjusted profile score (APS) estimator in Section 4, and the robust adjusted profile score (RAPS) estimator in Section 5 with Tukey's loss function ($k = 4.685$).

The simulation results are reported in Table 1 for $p = 25$ and Table 2 for $p = 160$. Here is a summary of the results:

1. In setup 1, the PS estimator has the smallest root-median square error (RMSE) and the shortest confidence interval (CI) with the desired coverage rate. The IVW estimator performs very well when $p = 25$ but has considerable bias and less than nominal coverage when $p = 160$. The APS and RAPS estimators have slightly longer CI than PS. The MR-Egger and weighted median estimators are less accurate than the other methods.
2. In setup 2, the PS estimator, as well as the weighted median, have substantial bias and perform poorly. The APS estimator is overall the best with very small bias and desired coverage, followed very closely by RAPS. The IVW and MR-Egger estimators also perform quite well, though their relative biases are more than 10% when $p = 160$.
3. In setup 3, all estimators besides RAPS have very large bias and poor CI coverage. The RMSE of the RAPS estimator is slightly larger than the RMSE in Model 2, and the coverage of RAPS is slightly below the nominal rate.
4. In setup 4, the direct effects α_j are distributed as Laplace instead of normal. The RAPS estimator has the smallest bias and RMSE, though the coverage is slightly below the nominal level.
5. In setup 5, the variance of α_j is proportional to $|\gamma_j|$. In this case APS and RAPS are approximately unbiased but the coverage is significantly lower than 95%.
6. In setup 6, 10% of the IVs have very large but roughly balanced

TABLE 1

Simulation results for $p = 25$. The summary statistics reported are: bias divided by β_0, root-median-square error (RMSE) divided by β_0, length of the confidence interval (CI) divided by β_0, and the coverage rate of the CI (nominal rate is 95%), all in %.

Setup	Method	Bias %	RMSE %	CI Len. %	Cover. %
1	IVW	− 2.9	12.7	73.8	95.4
	Egger	− 7.4	24.4	142.3	95.3
	W. Median	− 5.2	17.0	105.5	96.5
	PS	− 0.1	12.7	74.9	95.1
	APS	− 0.4	12.7	76.8	96.0
	RAPS	− 0.4	13.0	79.0	96.1
2	IVW	− 3.0	29.3	167.9	93.3
	Egger	− 8.2	59.7	319.2	92.1
	W. Median	− 12.8	39.9	121.4	70.6
	PS	14.7	36.1	71.4	49.2
	APS	− 0.2	28.8	165.4	93.4
	RAPS	− 0.1	30.1	170.2	93.1
3	IVW	−115.5	115.2	225.6	48.1
	Egger	−264.2	262.8	409.1	25.5
	W. Median	− 80.7	79.5	151.4	47.3
	PS	−122.3	121.3	66.1	6.9
	APS	− 86.2	85.6	207.0	65.0
	RAPS	− 11.6	40.6	168.7	84.3
4	IVW	− 5.1	25.1	159.5	96.0
	Egger	− 54.5	58.8	300.9	90.0
	W. Median	− 22.5	26.0	113.2	83.8
	PS	13.4	31.2	71.7	55.9
	APS	4.0	25.6	158.4	96.1
	RAPS	2.6	20.3	117.5	93.3
5	IVW	− 2.4	48.2	169.7	76.3
	Egger	− 8.2	98.0	321.0	72.9
	W. Median	− 24.4	60.4	136.7	56.0
	PS	15.8	57.2	71.6	33.0
	APS	0.9	46.8	183.0	81.1
	RAPS	1.5	44.9	169.0	78.3
6	IVW	− 8.1	64.2	382.8	94.8
	Egger	−102.2	134.8	723.7	90.7
	W. Median	− 30.8	50.3	130.6	63.1
	PS	200.2	309.6	82.1	4.1
	APS	13.7	62.1	327.1	92.8
	RAPS	12.3	50.3	298.2	85.4

Table 2

Simulation results for $p = 160$. The summary statistics reported are: bias divided by β_0, root-median-square error (RMSE) divided by β_0, length of the confidence interval (CI) divided by β_0, and the coverage rate of the CI (nominal rate is 95%), all in %.

Setup	Method	Bias %	RMSE %	CI Len. %	Cover. %
1	IVW	− 11.1	12.2	51.0	87.0
	Egger	− 10.1	15.2	79.9	92.6
	W. Median	− 12.6	15.6	84.3	93.9
	PS	0.1	9.6	57.0	95.2
	APS	− 0.4	9.5	58.3	95.8
	RAPS	− 0.5	9.8	59.9	95.8
2	IVW	− 11.6	23.2	122.5	92.6
	Egger	− 10.8	34.9	191.5	93.6
	W. Median	− 25.7	34.3	105.5	68.9
	PS	119.2	119.8	51.0	6.2
	APS	− 0.4	23.0	134.8	95.1
	RAPS	− 0.4	23.8	138.7	95.1
3	IVW	− 70.1	69.9	131.3	44.7
	Egger	− 125.5	125.6	203.8	32.3
	W. Median	− 65.0	65.0	111.5	41.5
	PS	4.1	77.9	44.6	15.5
	APS	− 47.9	48.3	139.3	73.2
	RAPS	− 3.9	27.4	137.9	90.6
4	IVW	− 11.9	20.5	121.5	94.7
	Egger	− 13.6	31.5	189.5	94.7
	W. Median	− 24.1	24.8	93.9	80.2
	PS	134.7	114.3	51.4	7.1
	APS	4.8	20.8	133.6	96.5
	RAPS	4.3	16.1	91.3	93.6
5	IVW	− 11.0	53.9	139.7	62.2
	Egger	− 9.8	92.5	217.7	56.9
	W. Median	− 56.0	63.7	125.2	49.3
	PS	− 819.8	244.0	57.8	4.7
	APS	− 0.3	55.3	170.7	71.6
	RAPS	1.5	48.6	120.4	59.8
6	IVW	− 12.7	47.2	278.8	95.0
	Egger	− 16.4	74.2	435.3	94.9
	W. Median	− 34.9	43.6	115.2	63.1
	PS	> 999.9	> 999.9	> 999.9	12.8
	APS	13.6	50.2	291.2	95.2
	RAPS	10.8	42.7	258.4	91.2

pleiotropy effects α_j. All estimators are biased in this case. The RAPS estimator has the smallest RMSE but the CI coverage is slightly below 95%. The IVW and APS estimators have slightly larger RMSE and the CI has the desired coverage rate.

Overall, the RAPS estimator is the clear winner in this simulation: when there is no idiosyncratic outlier (setups 1 and 2), it behaves almost as well as the best performer; when there is an idiosyncratic outlier (setup 3), it still has very small bias and close-to-nominal coverage; when our modeling assumptions are not satisfied (setups 4, 5, 6), it still has the smallest bias and RMSE, though the CI may fail to cover β_0 at the nominal rate.

7. A real-data comparison. Finally, we use another real dataset to compare the estimators in this paper and some existing methods. We shall refer to it as the BMI-BMI example, because both the "exposure" and the "outcome" are BMI. Although there is no "causal" effect of BMI on itself, the Model 1 of GWAS summary data should technically hold with $\beta_0 = 1$. Therefore, this is a rare scenario where we know the truth in real data. Since there are many SNPs that are only weakly associated with BMI, this example also offers a good opportunity to probe the issue of weak instrument bias and the efficiency gain by including many weak IVs.

We obtained three datasets for this example:

`BMI-GIANT`: full dataset from the GIANT consortium [35] (i.e. combining `BMI-FEM` and `BMI-MAL`), used to select SNPs.
`BMI-UKBB-1`: half of the UKBB data, used as the "exposure".
`BMI-UKBB-2`: another half of UKBB data, used as the "outcome".

We applied in total six methods. Four have been previously developed: besides the three estimators considered in Section 6, we also included the weighted mode estimator of Hartwig, Davey Smith and Bowden [27]. We use the implementation in the `TwoSampleMR` software package [29] for the existing methods. The last two methods were the PS and RAPS estimators developed in this paper (APS performs similarly to PS and RAPS and is omitted).

The results are reported in Table 3. Overall, the PS and RAPS estimators provided very accurate estimate of $\beta_0 = 1$. PS has the smallest standard error because there is no pleiotropy at all in this example. When there is pleiotropy (as expected in most real studies), PS can perform poorly as demonstrated in Section 6. All the existing methods are biased especially when there are many weak IVs.

TABLE 3

Results of the BMI-BMI example. The true β_0 should be 1. We considered 8 selection thresholds p_{sel} from 1×10^{-9} to 1×10^{-2}. The mean and median of the F-statistics $\hat{\gamma}_j^2/\sigma_{Xj}^2$ are reported. In each setting, we report the point estimate and the standard error of all the methods.

p_{sel}	# SNPs	Mean F	IVW	W. Median	W. Mode
1e-9	48	78.6	0.983 (0.026)	0.945 (0.039)	0.941 (0.042)
1e-8	58	69.2	0.983 (0.024)	0.945 (0.039)	0.939 (0.044)
1e-7	84	55.0	0.988 (0.024)	0.945 (0.036)	0.933 (0.041)
1e-6	126	44.1	0.986 (0.022)	0.944 (0.034)	0.931 (0.038)
1e-5	186	34.3	0.986 (0.019)	0.943 (0.033)	0.928 (0.039)
1e-4	287	26.1	0.981 (0.017)	0.941 (0.031)	0.929 (0.035)
1e-3	474	18.8	0.955 (0.015)	0.903 (0.027)	0.917 (0.231)
1e-2	812	12.7	0.928 (0.014)	0.879 (0.023)	0.739 (7.130)

p_{sel}	# SNPs	Median F	Egger	PS	RAPS
1e-9	48	51.8	0.926 (0.055)	0.999 (0.023)	0.998 (0.026)
1e-8	58	42.0	0.928 (0.050)	0.999 (0.023)	0.998 (0.025)
1e-7	84	32.1	0.905 (0.048)	1.012 (0.021)	1.004 (0.025)
1e-6	126	27.4	0.881 (0.043)	1.017 (0.019)	1.009 (0.023)
1e-5	186	21.0	0.874 (0.036)	1.020 (0.018)	1.013 (0.020)
1e-4	287	15.8	0.921 (0.031)	1.023 (0.017)	1.018 (0.018)
1e-3	474	10.8	0.913 (0.027)	1.010 (0.016)	1.006 (0.016)
1e-2	812	5.6	0.909 (0.022)	1.010 (0.015)	1.005 (0.015)

In Table 4 we illustrate the danger of selection bias. In this example we discard the `BMI-GIANT` dataset and use `BMI-UKBB-1` for both selection and inference (estimating γ_j). The estimators are biased towards 0 in almost all cases, even if we only use the genome-wide significant p-value threshold 10^{-9} or 10^{-8}. This is because the assumption $\hat{\gamma}_j \sim \text{N}(\gamma_j, \sigma_{Xj}^2)$ is violated. In fact, due to selection bias, the selected $\hat{\gamma}_j$ are stochastically larger than their mean γ_j (if $\gamma_j > 0$). Compared with other methods, the MR-Egger regression seems to be less affected by the selection bias.

8. Discussion. In this paper we have proposed a systematic approach for two-sample summary-data Mendelian randomization based on modifying the profile score function. By considering increasingly more complex models, we arrived at the Robust Adjusted Profile Score (RAPS) estimator which is robust to both systematic and idiosyncratic pleiotropy and performed excellently in all the numerical examples. Thus we recommend to routinely use the RAPS estimator in practice, especially if the exposure and the outcome are both complex traits.

Our theoretical and empirical results advocate for a new design of two-sample MR. Instead of using just a few strong SNPs (those with large $|\hat{\gamma}_j|/\sigma_{Xj}$), we find that adding many (potentially hundreds of) weak SNPs

TABLE 4

Illustration of selection bias. The same `BMI-UKBB-1` dataset is used for both selecting SNPs and estimating the SNP-exposure effects γ_j. All estimators are biased (true $\beta_0 = 1$) due to not accounting for selection bias.

p_{sel}	# SNPs	Mean F	IVW	W. Median	W. Mode
1e-9	110	68.63	0.851 (0.02)	0.83 (0.025)	0.896 (0.046)
1e-8	168	57.00	0.823 (0.017)	0.8 (0.022)	0.885 (0.053)
1e-7	228	50.08	0.799 (0.016)	0.768 (0.019)	0.886 (0.058)
1e-6	305	43.92	0.761 (0.015)	0.736 (0.019)	0.865 (0.079)
1e-5	443	36.98	0.721 (0.013)	0.667 (0.016)	0.824 (0.12)
1e-4	652	30.68	0.678 (0.012)	0.616 (0.015)	0.593 (0.122)
1e-3	929	25.36	0.629 (0.011)	0.57 (0.014)	0.576 (0.096)
1e-2	1289	20.70	0.592 (0.01)	0.528 (0.013)	0.554 (0.093)

p_{sel}	# SNPs	Median F	Egger	PS	RAPS
1e-9	110	49.20	1.071 (0.051)	0.871 (0.015)	0.862 (0.021)
1e-8	168	41.12	1.018 (0.046)	0.848 (0.014)	0.831 (0.018)
1e-7	228	37.12	1.016 (0.043)	0.824 (0.012)	0.803 (0.016)
1e-6	305	33.68	1.006 (0.041)	0.793 (0.011)	0.763 (0.016)
1e-5	443	28.74	0.957 (0.037)	0.762 (0.01)	0.716 (0.015)
1e-4	652	23.23	0.89 (0.033)	0.724 (0.009)	0.66 (0.014)
1e-3	929	19.12	0.823 (0.03)	0.687 (0.008)	0.594 (0.013)
1e-2	1289	15.26	0.749 (0.025)	0.657 (0.008)	0.541 (0.012)

usually substantially decreases the variance of the estimator. This is not feasible with existing methods for MR because they usually require the instruments to be strong. An additional advantage of using many weak instruments is that outliers in the sense of Model 3 are more easily detected, so the results are generally more robust to pleiotropy. There is one caveat: selection bias is more significant for weaker instruments, so a sample-splitting design (such as the one in Section 1.2) should be used.

In Models 2 and 3, we have assumed that the pleiotropy effects are completely independent and normally or nearly normally distributed. We view this assumption as an approximate modeling assumption rather than the precise data generating mechanism. It is motivated by the real data (Section 3.4) and seems to fit the data very well (Section 5.3). It is a special instance of the INstrument Strength Independent of Direct Effect (INSIDE) assumption [11] that is common in the summary-data MR literature. Apart from normality, two other implicit but key assumptions we made are:

1. The pleiotropy effects α_j are additive rather than multiplicative (the variance of α_j is proportional to σ_{Yj}) [8]. Multiplicative random effects model are easier to fit especially if the measurement error in $\hat{\gamma}_j$ is ignored, however it is quite unrealistic because α_j is a population quantity and thus is unlikely to be dependent on a sample quantity (for

example, σ_{Yj} may vary due to missing data). In contrast, the additive model is well motivated by the linear structural model in 2.3.

2. The pleiotropy effects α_j have mean 0. In comparison, the MR-Egger regression [8] assumes α_j has an unknown mean μ and refers to the case $\mu \neq 0$ as "directional pleiotropy". We have not seen strong evidence of "directional pleiotropy" in real datasets, and, more importantly, assuming $\mu \neq 0$ implies that there is a "special" allele coding so that $\alpha_j \sim \mathrm{N}(\mu, \tau^2)$. It is thus impossible to obtain estimators of β that are invariant to allele recoding without completely reformulating the MR-Egger model. For further details see Bowden et al. [12].

There are many technical challenges in the development of this paper. Due to the nature of the many weak IV problem, the asymptotics we considered are quite different from the classical measurement error literature. In Section 3 we showed the profile likelihood is information biased when there are many weak IVs, and in Section 4.1 we showed the profile likelihood is biased when there is overdispersion caused by systematic pleiotropy. This issue is solved by adjusting the profile score, but the proof of the consistency of the APS estimator is nontrivial. Consistency of the the RAPS estimator is even more challenging and still open because the estimating equations may have multiple roots, although we found its practical performance is usually quite benign. A possible solution is to initialize by another robust and consistent estimator (similar to the MM-estimation in robust regression, see Yohai [56]). However, we are not aware of any other provably robust and consistent estimator in our setting, and deriving such estimator is beyond the scope of this paper.

Software and reproducibility. `R` code for the methods proposed in this paper can be found in the package `mr.raps` that is publicly available at `https://github.com/qingyuanzhao/mr.raps` and can be directly called from `TwoSampleMR`. Numerical examples can be reproduced by running examples in the `R` package.

References.

[1] Clarivate Analytics (2017). Web of Science Topic: Mendelian Randomization. Data retrieved from http://www.webofknowledge.com.

[2] Anderson, T. W. and Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics* **20** 46–63.

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

[4] Barndorff-Nielsen, O. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70** 343–365.

[5] Bekker, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica* 657–681.
[6] Bentkus, V. (2005). A Lyapunov-type bound in $\mathbb{R}^d$. *Theory of Probability and Its Applications* **49** 311–323.
[7] Bound, J., Jaeger, D. A. and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of American Statistical Association* **90** 443–450.
[8] 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.
[9] 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.
[10] Bowden, J., Fabiola Del Greco, M., Minelli, C., Lawlor, D., Sheehan, N., Thompson, J. and Smith, G. D. (2017a). Improving the accuracy of two-sample summary data Mendelian randomization: moving beyond the NOME assumption. *bioRxiv:159442*.
[11] Bowden, J., Del Greco, M., Minelli, C., Davey Smith, G., Sheehan, N. and Thompson, J. (2017b). A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Statistics in Medicine* **36** 1783–1802.
[12] Bowden, J., Spiller, W., Del-Greco, F., Sheehan, N., Thompson, J., Minelli, C. and Smith, G. D. (2017c). Improving the visualisation, interpretation and analysis of two-sample summary data Mendelian randomization via the radial plot and radial regression. *BioRxiv* 200378.
[13] Boyle, E. A., Li, Y. I. and Pritchard, J. K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169** 1177–1186.
[14] Burgess, S., Butterworth, A. and Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology* **37** 658–665.
[15] Burgess, S., Scott, R. A., Timpson, N. J., Smith, G. D., Thompson, S. G. and Consortium, E.-I. (2015). Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *European Journal of Epidemiology* **30** 543–552.
[16] Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. CRC press.
[17] Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B (Methodological)* **49** 1–39.
[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] Davey Smith, G. and Hemani, G. (2014). Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics* **23** R89–R98.
[20] Davey Smith, G., Lawlor, D. A., Harbord, R., Timpson, N., Day, I. and Ebrahim, S. (2007). Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. *PLoS Medicine* **4** e352.
[21] Didelez, V. and Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research* **16**

309–330.
[22] Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Oxford University Press.
[23] Greenland, S., Robins, J. M. and Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science* **14** 29–46.
[24] Guo, Z., Kang, H., Cai, T. T. and Small, D. S. (2016). Confidence Intervals for Causal Effects with Invalid Instruments using Two-Stage Hard Thresholding with Voting. *arXiv:1603.05224*.
[25] Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of American Statistical Association* **69** 383–393.
[26] Hansen, C., Hausman, J. and Newey, W. (2008). Estimation with many instrumental variables. *Journal of Business and Economic Statistics* **26** 398–422.
[27] Hartwig, F. P., Davey Smith, G. and Bowden, J. (2017). Robust inference in two-sample Mendelian randomisation via the zero modal pleiotropy assumption. *bioRxiv:126102*.
[28] Haycock, P. C., Burgess, S., Wade, K. H., Bowden, J., Relton, C. and Smith, G. D. (2016). Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. *The American journal of clinical nutrition* **103** 965–978.
[29] Hemani, G., Zheng, J., Wade, K. H., Laurin, C., Elsworth, B., Burgess, S., Bowden, J., Langdon, R., Tan, V., Yarmolinsky, J., Shihab, H. A., Timpson, N., Evans, D. M., Relton, C., Martin, R. M., Smith, G. D., Gaunt, T. R., Haycock, P. C. and The MR-Base Collaboration (2016). MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations. *bioRxiv:078972*.
[30] Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics* **35** 73–101.
[31] Ioannidis, J. P., Trikalinos, T. A. and Khoury, M. J. (2006). Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases. *American journal of epidemiology* **164** 609–614.
[32] 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 American Statistical Association* **111** 132–144.
[33] Katan, M. (1986). Apoupoprotein E isoforms, serum cholesterol, and cancer. *Lancet* **327** 507–508.
[34] Li, S. (2017). Mendelian randomization when many instruments are invalid: hierarchical empirical Bayes estimation. *arXiv:1706.01389*.
[35] Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., Powell, C., Vedantam, S., Buchkovich, M. L., Yang, J. et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518** 197–206.
[36] Maronna, R., Martin, R. D. and Yohai, V. (2006). *Robust Statistics: Theory and Methods*. John Wiley & Sons.
[37] McCullagh, P. and Tibshirani, R. (1990). A simple method for the adjustment of profile likelihoods. *Journal of the Royal Statistical Society. Series B (Methodological)* **52** 325–344.
[38] Murphy, S. A. and Van der Vaart, A. W. (1996). Likelihood inference in the errors-in-variables model. *Journal of Multivariate Analysis* **59** 81–108.
[39] Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially con-

sistent observations. *Econometrica* **16** 1–32.
[40] Pacini, D. and Windmeijer, F. (2016). Robust inference for the Two-Sample 2SLS estimator. *Economics Letters* **146** 50–54.
[41] Park, J.-H., Wacholder, S., Gail, M. H., Peters, U., Jacobs, K. B., Chanock, S. J. and Chatterjee, N. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature Genetics* **42** 570–575.
[42] Pearl, J. (2009). *Causality*. Cambridge University Press.
[43] Purcell, S. PLINK (software V1.07). http://pngu.mgh.harvard.edu/purcell/plink/.
[44] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J. and Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81** 559–575.
[45] Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica* **39** 577–591.
[46] Shi, H., Kichaev, G. and Pasaniuc, B. (2016). Contrasting the genetic architecture of 30 complex traits from summary association data. *The American Journal of Human Genetics* **99** 139–153.
[47] 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.
[48] Stearns, F. W. (2010). One hundred years of pleiotropy: a retrospective. *Genetics* **186** 767–773.
[49] Stock, J. H. and Yogo, M. (2005). Asymptotic distributions of instrumental variables statistics with many instruments. In *Essays in Honor of Thomas Rothenberg* 6, 109–120. Cambridge University Press, Cambridge.
[50] Tchetgen Tchetgen, E. J., Sun, B. and Walter, S. (2017). The GENIUS Approach to Robust Mendelian Randomization Inference. *arXiv:1709.07779*.
[51] van Kippersluis, H. and Rietveld, C. A. (2017). Pleiotropy-robust Mendelian randomization. *International Journal of Epidemiology* **in press**.
[52] Vansteelandt, S., Bowden, J., Babanezhad, M., Goetghebeur, E. et al. (2011). On instrumental variables estimation of causal odds ratios. *Statistical Science* **26** 403–422.
[53] Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *Annals of Mathematical Statistics* **11** 284–300.
[54] Wright, P. G. (1928). *Tariff on Animal and Vegetable Oils*. MacMillan, New York.
[55] Wright, S. (1968). *Evolution and the Genetics of Populations, Volume 1: Genetic and Biometric Foundations* **1**. University of Chicago Press.
[56] Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Annals of Statistics* **15** 642–656.

APPENDIX A: LINEAR MODEL FOR GWAS SUMMARY DATA

In this Appendix we give additional justifications of the linear model (1.1) for GWAS summary data. We will show (1.1) is very likely to hold in very general situations, much beyond the linear structural model considered in Section 2.

A.1. Binary outcome and logistic model. When the outcome Y is binary, the linear structural model (2.2) is no longer appropriate. Instead, we consider the following logistic model of Y (let $H(t) = 1/(1+e^{-t})$ be the logistic link function):

$$X = \sum_{j=1}^{p} \gamma_j Z_j + \eta_X U + E_X,\ Y \sim \text{Bernoulli}\big(H(\beta X + \eta_Y U)\big). \tag{A.1}$$

Next we derive an approximation of the coefficient Γ_j when we run a logistic regression of Y on Z_j. By (A.1), we have

$$\mathbb{P}(Y = 1 | Z_j = z_j) = \mathbb{E}\big[H\big(\beta\gamma_j z_j + E'\big)\big],$$

where $E' = \beta \sum_{k \neq j} \gamma_k Z_k + \beta\eta_X U + \beta E_X + \eta_Y U$. If we assume $E' \sim \mathrm{N}(\mu, \sigma_j^2)$, then

$$\mathbb{P}(Y = 1 | Z_j = z_j) = \int_{-\infty}^{\infty} H\big(\mu + \beta\gamma_j z_j + \sigma_j e\big)\phi(e)\,\mathrm{d}e.$$

Note that $\sigma_j^2 \approx \sigma^2 = \mathrm{Var}(\beta X + \eta_Y U)$ when γ_j is small.

To proceed further we introduce a well-known probit approximation of logistic function [16]:

$$H(t) \approx \Phi(t/1.7).$$

By using the following Gaussian integral identity,

$$\int_{-\infty}^{\infty} \Phi(a + bx)\phi(x)\,\mathrm{d}x = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right),$$

we obtain

$$\begin{aligned}\mathbb{P}(Y = 1 | Z_j = z_j) &\approx \int_{-\infty}^{\infty} \Phi\left(\frac{\mu + \beta\gamma_j z_j + \sigma e}{1.7}\right)\phi(e)\,\mathrm{d}e \\ &= \Phi\left(\frac{\mu + \beta\gamma_j z_j}{1.7\sqrt{1 + (\sigma/1.7)^2}}\right) \\ &\approx H\left(\frac{\mu + \beta\gamma_j z_j}{\sqrt{1 + (\sigma/1.7)^2}}\right).\end{aligned}$$

Therefore $\Gamma_j \approx \beta\gamma_j/\sqrt{1 + (\sigma/1.7)^2}$. In other words, model (1.1) is approximately correct with $\beta_0 = \beta/\sqrt{1 + (\sigma/1.7)^2}$. The attenuation bias is due to the non-collapsibility of odds ratio [23]. Notice that although we assumed E' is normally distributed in our calculation, this approximation is quite accurate for many other distributions [16, Section 4.8.2]. A similar result can be found in Vansteelandt et al. [52] who also discussed the general intepretation of causal odds ratios.

A.2. General situation: a local argument. The linear model (1.1) may actually hold in much broader situations than the linear and logistic models considered above. The main reason is that for most SNPs, the influence on a complex trait X is usually minuscule [13, 31, 41, 46]. Let's consider a continuous exposure X and the quantity $\mathbb{E}[h(Y)|Z_j = 1] - \mathbb{E}[h(Y)|Z_j = 0]$ for some function h of interest. Assuming appropriate differentiability and using the shorthand notation $X(z_1) = g(z_1, Z_2, \ldots, Z_p, U, E_X)$, we have,

$$\begin{aligned} & \mathbb{E}[h(Y)|Z_1 = 1] - \mathbb{E}[h(Y)|Z_1 = 0] \\ =& \mathbb{E}\big[h\big(f(X(1), U, E_Y)\big) - h\big(f(X(0), U, E_Y)\big)\big] \\ \approx& \mathbb{E}[h'(f^{(1)}(X, U, E_Y)) \cdot (X(1) - X(0))], \end{aligned}$$

where h' is th derivative of h and $f^{(1)}$ is the partial derivative of f with respect to its first argument. In this approximation we have used the assumption that $X(1) - X(0)$ is small, i.e. the exposure X is not changed by a single instrument Z_1 by much.

In many epidemiological problems, the causal effect of X on the outcome Y is also very small compared to the variance of Y. Therefore, when it is reasonable to assume that the variability of the term $f^{(1)}(X, U, E_Y)$ is mostly driven by the noise variable E_Y which is independent of $X(0)$ and $X(1)$, we have

$$\mathbb{E}[h(Y)|Z_1 = 1] - \mathbb{E}[h(Y)|Z_1 = 0] \approx \mathbb{E}[h'(f^{(1)}(X, U, E_Y))] \cdot \mathbb{E}[X(1) - X(0)].$$

The left hand side of the above equation may be regarded as a general version of Γ_1 and $\mathrm{E}[X(1) - X(0)]$ a general version of γ_1. Thus we arrive at the approximation $\Gamma_1 \approx \beta_0 \gamma_1$ for $\beta_0 = \mathbb{E}[h'(f^{(1)}(X, U, E_Y))]$. This may be interpreted as the average of "local" causal effect: let h be the identity function and write the potential outcome $Y(x) = f(x, U, E_Y)$, then

$$\beta_0 \approx \lim_{\Delta x \to 0} \frac{\mathbb{E}[Y(X + \Delta x) - Y(X)]}{\Delta x},$$

where the expectation is taken jointly over X, U, and E_Y.

The above local argument reflects a meta-analysis interpretation of MR [8]: each SNP can be viewed as randomized experiment that changes the exposure X by just a little. Because all the changes are relatively small compared to the variability of X and Y, the relationship between γ_j and Γ_j is almost linear. This is why we expect the approximate linear relation (1.1) may hold in many problems beyond those discussed in Section 2.2 and Appendix A.1.

APPENDIX B: PROOFS

B.1. Proof of Theorem 3.1. Notice that by Assumption 2, $p/(n^2\|\boldsymbol{\gamma}\|_2^2) \to 0$ implies that $n \to \infty$. Let $e_j = \hat{\Gamma}_j - \Gamma_j$ and $\epsilon_j = \hat{\gamma}_j - \gamma_j$. After some algebra, we have

$$l(\beta) = -\frac{1}{2}\sum_{j=1}^{p} \frac{\gamma_j^2(\beta_0-\beta)^2 + (e_j - \beta\epsilon_j)^2 + 2\gamma_j(\beta_0-\beta)(e_j-\beta\epsilon_j)}{\sigma_{Xj}^2\beta^2 + \sigma_{Yj}^2}.$$

Notice that $e_j - \beta\epsilon_j \sim \mathrm{N}(0, \sigma_{Xj}^2\beta^2 + \sigma_{Yj}^2)$. By Assumption 3, suppose $\sigma_{Xj}^2 \geq c_\sigma/n$ and $\sigma_{Yj}^2 \geq c_\sigma/n$ for all $j \in [p]$. Using the elementary inequality $2/(a+b) \geq \min(1/a, 1/b)$ for $a, b > 0$, we obtain

$$\begin{aligned}
&-2l(\beta) \\
=&(\beta_0-\beta)^2\Bigg[\sum_{j=1}^{p}\frac{\gamma_j^2}{\sigma_{Xj}^2\beta^2+\sigma_{Yj}^2}\Bigg] + p + O_p\big(\sqrt{p} + \sqrt{n}\|\boldsymbol{\gamma}\|\cdot|\beta_0-\beta|\big) \\
\geq&\frac{1}{2}(\beta_0-\beta)^2\min\left(\sum_{j=1}^{p}\frac{\gamma_j^2}{\sigma_{Yj}^2}, \sum_{j=1}^{p}\frac{\gamma_j^2}{\sigma_{Xj}^2\beta^2}\right) + p + O_p\big(\sqrt{p} + \sqrt{n}\|\boldsymbol{\gamma}\|\cdot|\beta_0-\beta|\big) \\
\geq&\frac{n\|\boldsymbol{\gamma}\|_2^2}{2c_\sigma}\min\left((\beta_0-\beta)^2, \frac{(\beta_0-\beta)^2}{\beta^2}\right) + p + O_p\big(\sqrt{p} + \sqrt{n}\|\boldsymbol{\gamma}\|\cdot|\beta_0-\beta|\big)
\end{aligned}$$

Consider the case $\beta_0 > 0$. By taking derivative, it is easy to verify that $f(\beta) = (\beta_0-\beta)^2/\beta^2$ is decreasing in β when $0 < \beta < \beta_0$ and increasing in β when $\beta < 0$ or $\beta > \beta_0$. Since $f(\beta) \to 1$ as $|\beta| \to \infty$, for any $\epsilon > 0$ there exists constant $C(\beta_0, \epsilon) > 0$ such that $\inf_{|\beta-\beta_0|\geq\epsilon}(\beta_0-\beta)^2/\beta^2 \geq C(\beta_0,\epsilon)$. Similarly, we can show this is also true for $\beta_0 < 0$ and $\beta_0 = 0$. Also, notice that the last term $O_p(\sqrt{n}|\beta_0-\beta|)$ is negligible compared to the first term when $|\beta-\beta_0| \geq \epsilon$. Let $C'(\beta_0,\epsilon) = \min(\epsilon^2, C(\beta_0,\epsilon)) > 0$. We have

$$\inf_{|\beta-\beta_0|\geq\epsilon} -2l(\beta) \geq (1+o_p(1))C'(\beta_0,\epsilon)\frac{n\|\boldsymbol{\gamma}\|_2^2}{2c_\sigma} + p + O_p\big(\sqrt{p}\big).$$

Finally, by comparing this to $-2l(\beta_0) = p + O_p(\sqrt{p})$, we have

$$\mathbb{P}\Big(l(\beta_0) > \sup_{|\beta-\beta_0|\geq\epsilon} l(\beta)\Big) = \mathbb{P}\Big(O_p(\sqrt{p}) \leq (1+o_p(1))C'(\beta_0,\epsilon)\frac{n\|\boldsymbol{\gamma}\|_2^2}{2c_\sigma} + O_p\big(\sqrt{p}\big)\Big).$$

When $p \ll n^2\|\boldsymbol{\gamma}\|^4$, it is easy to see that this probability converges to 1.

B.2. Proof of Theorem 3.2. By (3.6) and the consistency of $\hat{\beta}$, we have

$$\begin{aligned}\hat{\beta} - \beta_0 &= \frac{-\psi(\beta_0)}{\psi'(\beta_0) + (1/2)\psi''(\tilde{\beta})(\hat{\beta} - \beta_0)} \\ &= \frac{V_2}{\sqrt{V_1}} \cdot \frac{-\psi(\beta_0)/\sqrt{V_1}}{[\psi'(\beta_0) + o_p(\psi''(\tilde{\beta}))]/V_2}\end{aligned}$$

The central limit theorem Equation (3.7) is immediately proven using Slutsky's lemma after showing the following three lemmas:

LEMMA B.1. $(1/\sqrt{V_1})\psi(\beta_0) \xrightarrow{d} \mathrm{N}(0,1)$.

LEMMA B.2. $(-1/V_2)\psi'(\beta_0) \xrightarrow{p} 1$.

LEMMA B.3. *For a neighborhood $\mathcal{N}$ of β_0, $\sup_{\beta\in\mathcal{N}}(1/V_2)\psi''(\beta) = O_p(1)$.*

Next we prove the three lemmas. For the first lemma, let $\psi_j(\beta)$ be the j-th summand in (3.5), so $\psi(\beta) = \sum_{j=1}^{p}\psi_j(\beta)$. It is easy to show that

$$\psi_j(\beta_0) = \frac{(e_j - \beta_0\epsilon_j)[\gamma_j(\sigma_{Yj}^2 + \sigma_{Xj}^2\beta_0^2) + e_j\sigma_{Xj}^2\beta_0 + \epsilon_j\sigma_{Yj}^2]}{(\sigma_{Yj}^2 + \sigma_{Xj}^2\beta_0^2)^2}. \tag{B.1}$$

The expectation of $\psi_j(\beta_0)$ is

$$\mathbb{E}[\psi_j(\beta_0)] = \frac{\mathbb{E}[e_j^2\sigma_{Xj}^2\beta_0 - \epsilon_j^2\sigma_{Yj}^2\beta_0]}{(\sigma_{Yj}^2 + \sigma_{Xj}^2\beta_0^2)^2} = 0.$$

This shows that $\mathbb{E}[\psi(\beta_0)] = 0$. The second moment of $\psi_j(\beta_0)$ is given by

$$\mathbb{E}[\psi_j(\beta_0)^2] = A_j + B_j,$$

where

$$A_j = \mathbb{E}\Big[\frac{(e_j - \beta_0\epsilon_j)^2\gamma_j^2(\sigma_{Yj}^2 + \sigma_{Xj}^2\beta_0^2)^2}{(\sigma_{Yj}^2 + \sigma_{Xj}^2\beta_0^2)^4}\Big] = \frac{\gamma_j^2}{\sigma_{Yj}^2 + \sigma_{Xj}^2\beta_0^2},$$

and

$$\begin{aligned}
B_j &= \mathbb{E}\Big[\frac{(e_j - \beta_0\epsilon_j)^2(e_j\sigma_{Xj}^2\beta_0 + \epsilon_j\sigma_{Yj}^2)^2}{(\sigma_{Yj}^2 + \sigma_{Xj}^2\beta_0^2)^4}\Big] \\
&= \mathbb{E}\Big[\frac{e_j^4\sigma_{Xj}^4\beta_0^2 + \epsilon_j^4\sigma_{Yj}^4\beta_0^2 + e_j^2\epsilon_j^2(\sigma_{Yj}^4 - 4\beta_0\sigma_{Xj}^2\beta_0\sigma_{Yj}^2 + \beta_0^2\sigma_{Xj}^4\beta_0^2)}{(\sigma_{Yj}^2 + \sigma_{Xj}^2\beta_0^2)^4}\Big] \\
&= \frac{3\sigma_{Yj}^4\sigma_{Xj}^4\beta_0^2 + 3\sigma_{Xj}^4\sigma_{Yj}^4\beta_0^2 + \sigma_{Yj}^2\sigma_{Xj}^2(\sigma_{Yj}^4 - 4\beta_0\sigma_{Xj}^2\beta_0\sigma_{Yj}^2 + \beta_0^2\sigma_{Xj}^4\beta_0^2)}{(\sigma_{Yj}^2 + \sigma_{Xj}^2\beta_0^2)^4} \\
&= \frac{\sigma_{Yj}^2\sigma_{Xj}^2(\sigma_{Yj}^4 + 2\beta_0^2\sigma_{Xj}^2\sigma_{Yj}^2 + \beta_0^4\sigma_{Xj}^4)}{(\sigma_{Yj}^2 + \sigma_{Xj}^2\beta_0^2)^4} \\
&= \frac{\sigma_{Xj}^2\sigma_{Yj}^2}{(\sigma_{Yj}^2 + \sigma_{Xj}^2\beta_0^2)^2}.
\end{aligned}$$

In summary,

$$\mathbb{E}[\psi_j(\beta_0)^2] = \frac{\gamma_j^2\sigma_{Yj}^2 + \Gamma_j^2\sigma_{Xj}^2 + \sigma_{Xj}^2\sigma_{Yj}^2}{(\sigma_{Yj}^2 + \sigma_{Xj}^2\beta_0^2)^2}.$$

Notice that by Assumption 3, $\mathbb{E}[\psi_j(\beta_0)^2] = \Theta(n\gamma_j^2 + 1)$.

To prove Lemma B.1, we consider two scenarios:

Scenario 1: $p \to \infty$. In this case, we hope to use central limit theorem to show

$$\frac{1}{\sqrt{V_1}}\psi(\beta_0) = \frac{1}{\sqrt{V_1}}\sum_{j=1}^{p}\psi_j(\beta_0) \to \mathrm{N}(0,1).$$

Next we check Lyapunov's condition by computing the third moment of $\psi_j(\beta_0)$. Notice that

$$\begin{aligned}
\mathbb{E}[|\psi_j(\beta_0)|^3] &= \frac{\mathbb{E}\big\{\big|(e_j - \beta_0\epsilon_j)^3[\gamma_j(\sigma_{Yj}^2 + \sigma_{Xj}^2\beta_0^2) + e_j\sigma_{Xj}^2\beta_0 + \epsilon_j\sigma_{Yj}^2]^3\big|\big\}}{(\sigma_{Yj}^2 + \sigma_{Xj}^2\beta_0^2)^6} \\
&= \frac{\mathbb{E}[|C_0 + C_1\gamma_j + C_2\gamma_j^2 + C_3\gamma_j^3|]}{(\sigma_{Yj}^2 + \sigma_{Xj}^2\beta_0^2)^6} \\
&\le \frac{\mathbb{E}[|C_0|] + \gamma_j\mathbb{E}[|C_1|] + \gamma_j^2\mathbb{E}[|C_2|] + \gamma_j^3\mathbb{E}[|C_3|]}{(\sigma_{Yj}^2 + \sigma_{Xj}^2\beta_0^2)^6}
\end{aligned}$$

We omit the detailed expressions for C_0, C_1, C_2 and C_3 but note that there exists constant $C(\beta_0) > 0$ such that $\mathbb{E}[|C_i|] \le C(\beta_0)(1/n)^{6-i/2}$ for $i =$

$0, 1, 2, 3$. Therefore

$$\sum_{j=1}^{p} \mathbb{E}[|\psi_j(\beta_0)|^3] = O(p + \sqrt{n}\|\boldsymbol{\gamma}\|_1 + n\|\boldsymbol{\gamma}\|_2^2 + n^{3/2}\|\boldsymbol{\gamma}\|_3^3).$$

By the Cauchy-Schwarz inequality, $\sqrt{n}\|\boldsymbol{\gamma}\|_1 \le \sqrt{n}\sqrt{p}\|\boldsymbol{\gamma}\|_2 \le (p + n\|\boldsymbol{\gamma}\|_2^2)/2$. Using the assumption $\|\boldsymbol{\gamma}\|_3/\|\boldsymbol{\gamma}\|_2 \to 0$, it is easy to show the Lyapunov condition

$$\frac{\sum_{j=1}^{p} \mathbb{E}[|\psi_j(\beta_0)|^3]}{\big\{\sum_{j=1}^{p} \mathbb{E}[\psi_j(\beta_0)^2]\big\}^{3/2}} = O\bigg(\frac{p + n\|\boldsymbol{\gamma}\|_2^2 + n^{3/2}\|\boldsymbol{\gamma}\|_3^3}{(p + n\|\boldsymbol{\gamma}\|_2^2)^{3/2}}\bigg) \to 0.$$

Scenario 2: p is finite. By the assumption in the Theorem statement, $\kappa = n\|\boldsymbol{\gamma}\|_2^2/p \to \infty$. We can rewrite (B.1) to obtain

$$\psi(\beta_0) = \sum_{j=1}^{p} \frac{(e_j - \beta_0\epsilon_j)\gamma_j}{\sigma_{Yj}^2 + \sigma_{Xj}^2\beta_0^2} + \sum_{j=1}^{p} \frac{(e_j - \beta_0\epsilon_j)[e_j\sigma_{Xj}^2\beta_0 + \epsilon_j\sigma_{Yj}^2]}{(\sigma_{Yj}^2 + \sigma_{Xj}^2\beta_0^2)^2}.$$

The first term on the right hand side is distributed as $\mathrm{N}(0, V_2)$ and $V_2 = \Theta(n\|\boldsymbol{\gamma}\|_2^2)$. The second term has variance $O(p)$ and is thus ignorable compared to the first term. Therefore, $(1/\sqrt{V_2})\psi(\beta_0) \xrightarrow{d} \mathrm{N}(0,1)$. Since $p/n \to 0$, it is easy to show that $V_1/V_2 \to 1$. By Slutsky's lemma, $(1/\sqrt{V_1})\psi(\beta_0) \xrightarrow{d} \mathrm{N}(0,1)$.

Now we turn to Lemma B.2. It suffices to prove $\mathrm{E}[\psi'(\beta_0)] = V_2$ and $\mathrm{Var}(\psi'(\beta_0)/V_2) \to 0$. Next we compute the first two moments of $\psi'(\beta_0)$. By differentiating (3.5), we get

$$\psi'(\beta_0) = \sum_{j=1}^{p} \frac{A_j + B_j}{(\sigma_{Yj}^2 + \sigma_{Xj}^2\beta_0^2)^4},$$

where

$$\begin{aligned} A_j &= (\hat{\Gamma}_j^2\sigma_{Xj}^2 - \hat{\gamma}_j^2\sigma_{Yj}^2 - 2\beta_0\hat{\gamma}_j\hat{\Gamma}_j\sigma_{Xj}^2)(\sigma_{Yj}^2 + \sigma_{Xj}^2\beta_0^2)^2, \\ B_j &= -(\hat{\Gamma}_j - \beta_0\hat{\gamma}_j)(\hat{\gamma}_j\sigma_{Yj}^2 + \hat{\Gamma}_j\sigma_{Xj}^2\beta_0) \cdot 2(\sigma_{Yj}^2 + \sigma_{Xj}^2\beta_0^2) \cdot 2\sigma_{Xj}^2\beta_0. \end{aligned}$$

The expected values of these two terms are

$$\begin{aligned} \mathbb{E}[A_j] &= \big[(\Gamma_j^2 + \sigma_{Yj}^2)\sigma_{Xj}^2 - (\gamma_j^2 + \sigma_{Xj}^2)\sigma_{Yj}^2 - 2\beta_0\gamma_j\Gamma_j\sigma_{Xj}^2\big](\sigma_{Yj}^2 + \sigma_{Xj}^2\beta_0^2)^2 \\ &= -(\gamma_j^2\sigma_{Yj}^2 + \Gamma_j^2\sigma_{Xj}^2)(\sigma_{Yj}^2 + \sigma_{Xj}^2\beta_0^2)^2, \end{aligned}$$

$$\begin{aligned}\mathbb{E}[B_j] &= -\big[(\Gamma_j^2+\sigma_{Yj}^2)\sigma_{Xj}^2\beta_0 - (\gamma_j^2+\sigma_{Xj}^2)\sigma_{Yj}^2\beta_0 + \Gamma_j\gamma_j(\sigma_{Yj}^2-\sigma_{Xj}^2\beta_0^2)\big]\cdot \\ &\quad \cdot 2(\sigma_{Yj}^2+\sigma_{Xj}^2\beta_0^2)\cdot 2\sigma_{Xj}^2\beta_0 \\ &= -\big[(\Gamma_j^2\sigma_{Xj}^2-\gamma_j^2\sigma_{Yj}^2)\beta_0 + \Gamma_j\gamma_j(\sigma_{Yj}^2-\sigma_{Xj}^2\beta_0^2)\big]\cdot 4(\sigma_{Yj}^2+\sigma_{Xj}^2\beta_0^2)\sigma_{Xj}^2\beta_0 \\ &= 0.\end{aligned}$$

Therefore $\mathrm{E}[\psi'(\beta_0)] = -V_2$.

For the variance of $\psi'(\beta_0)$, consider any β in a neighborhood $\mathcal{N}$ of β. Our argument is based on the key observation that $\psi_j(\beta)$ is a homogeneous quadratic polynomial of $(\tilde{\gamma}_j,\tilde{e}_j,\tilde{\epsilon}_j) = (\sqrt{n}\gamma_j,\sqrt{n}e_j,\sqrt{n}\epsilon_j)$:

$$\psi_j(\beta) = \frac{((\beta_0-\beta)\tilde{\gamma}_j+\tilde{e}_j-\beta\tilde{\epsilon}_j)[(n\sigma_{Xj}^2\beta\beta_0+n\sigma_{Yj}^2)\tilde{\gamma}_j+(n\sigma_{Xj}^2\beta)\tilde{e}_j+(n\sigma_{Yj}^2)\tilde{\epsilon}_j]}{(n\sigma_{Xj}^2\beta^2+n\sigma_{Yj}^2)^2}.$$

Therefore its derivative, $\psi_j'(\beta)$, remains to be a homogeneous quadratic polynomial of $(\tilde{\gamma}_j,\tilde{e}_j,\tilde{\epsilon}_j)$. This observation suggests that $\mathbb{E}[\psi_j'(\beta_0)]$ is a quadratic function of $\tilde{\gamma}_j$. Notice that the any term in $\psi_j'(\beta)$ that has odd degree of $\tilde{\gamma}_j$ must have expectation equal to 0, because it must have odd degree in either $\tilde{e}_j$ or $\tilde{\epsilon}_j$. A simple calculation then yields $\mathbb{E}[\psi_j'(\beta)]] = \Theta(\tilde{\gamma}_j^2)$, so for $\beta\in\mathcal{N}$,

$$\mathbb{E}[\psi'(\beta)] = \Theta(n\|\boldsymbol{\gamma}\|_2^2).$$

Similarly, the variance $\mathrm{Var}[\psi_j'(\beta)^2]$ is also a quadratic polynomial of $\tilde{\gamma}_j$ (because the $\tilde{\gamma}_j^2$ term in $\psi_j'(\beta)$ is non-random). Thus for $\beta\in\mathcal{N}$,

$$\mathrm{Var}(\psi'(\beta)) = O(n\|\boldsymbol{\gamma}\|_2^2+p).$$

Using the assumption $p/(n^2\|\boldsymbol{\gamma}\|^4)\to 0$, it is then easy to see that

$$\mathrm{Var}(\psi'(\beta_0)) \ll \big(\mathrm{E}[\psi'(\beta_0)]\big)^2.$$

This concludes the proof of Lemma B.2.

The above argument for $\psi'(\beta_0)$ can also be applied to $\psi''(\beta)$ for any β in a neighborhood $\mathcal{N}$ of β_0, so $\mathrm{Var}(\psi''(\beta)) = O(n\|\boldsymbol{\gamma}\|_2^2+p) = o(V_2^2)$. Since $\psi(\beta)$ is smooth in β, this proves the Lemma B.3.

B.3. Proof of Theorem 3.3. By Theorem 3.2, $\hat{\beta}-\beta_0 = O_p(1/\sqrt{n})$. Thus

$$\sigma_{Yj}^2+\sigma_{Xj}^2\hat{\beta}^2 = (\sigma_{Yj}^2+\sigma_{Xj}^2\beta_0^2)(1+o_p(1/n)).$$

This implies that

$$\hat{V}_1 = (1+o_p(1))\cdot\sum_{j=1}^p \frac{(\hat{\gamma}_j^2-\sigma_{Xj}^2)\sigma_{Yj}^2+(\hat{\Gamma}_j^2-\sigma_{Yj}^2)\sigma_{Xj}^2+\sigma_{Xj}^2\sigma_{Yj}^2}{(\sigma_{Yj}^2+\sigma_{Xj}^2\beta_0^2)^2}.$$

It is easy to show that the summation on the right hand side has mean V_1. Similar to the proof of Theorem 3.2, the variance of this term is $O(p)$. Note that $V_1 = \Theta(n\|\boldsymbol{\gamma}\|_2^2 + p) = \Theta(p\cdot(1+\kappa))$, using the assumption in Theorem 3.2 that $p \to \infty$ or $\kappa \to \infty$, $\hat{V}_1 = (1 + o_p(1))V_1$. Similarly, $\hat{V}_2 = (1 + o_p(1))V_2$. Equation (3.9) follows immediately from Slutsky's lemma.

B.4. Proof of Theorem 4.1. To prove consistency of $(\hat{\beta}, \hat{\tau}_0^2)$, we need to study the asymptotic behavior of the adjusted profile score. Let $\psi_{1j}(\beta, \tau^2)$ and $\psi_{2j}(\beta, \tau^2)$ be the j-th term in the summation in (4.2) and (4.3), so $\psi_i(\beta, \tau^2) = \sum_{j=1}^p \psi_{ij}(\beta, \tau^2),\ i = 1, 2$. We first consider the expectation of ψ_{1j} and ψ_{2j}:

$$\begin{aligned}
&\mathbb{E}[\psi_{1j}(\beta,\tau^2)] \\
=&\mathbb{E}\left\{\frac{[(e_j - \beta\epsilon_j) + (\beta_0 - \beta)\gamma_j][\gamma_j(\sigma_{Xj}^2\beta\beta_0 + \sigma_{Yj}^2 + \tau^2) + \epsilon_j(\sigma_{Yj}^2 + \tau^2) + e_j\sigma_{Xj}^2\beta]}{(\sigma_{Xj}^2\beta^2 + \sigma_{Yj}^2 + \tau^2)^2}\right\} \\
=&\frac{(\tau^2 + \sigma_{Yj}^2 + \sigma_{Xj}^2\beta\beta_0)(\beta_0 - \beta)\gamma_j^2 + \sigma_{Xj}^2\beta(\tau_0^2 - \tau^2)}{(\tau^2 + \sigma_{Yj}^2 + \sigma_{Xj}^2\beta^2)^2},
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[\psi_{2j}(\beta,\tau^2)] &= \mathbb{E}\left\{\sigma_{Xj}^2\frac{[(e_j - \beta\epsilon_j) + (\beta_0 - \beta)\gamma_j]^2 - (\tau^2 + \sigma_{Yj}^2 + \sigma_{Xj}^2\beta^2)}{(\tau^2 + \sigma_{Yj}^2 + \sigma_{Xj}^2\beta^2)^2}\right\} \\
&= \frac{\sigma_{Xj}^2(\beta_0 - \beta)^2\gamma_j^2 + \sigma_{Xj}^2(\tau_0^2 - \tau^2)}{(\tau^2 + \sigma_{Yj}^2 + \sigma_{Xj}^2\beta^2)^2}.
\end{aligned} \tag{B.2}$$

Now consider the following contrast of the two estimating equations:

$$\tilde{\psi}(\beta, \tau^2) = \psi_1(\beta, \tau^2) - \beta\psi_2(\beta, \tau^2).$$

It is straightforward to verify that

$$\begin{aligned}
\mathbb{E}[\tilde{\psi}(\beta,\tau^2)] &= \sum_{j=1}^p \frac{(\beta_0 - \beta)\gamma_j^2(\tau^2 + \sigma_{Yj}^2 + \sigma_{Xj}^2\beta\beta_0 - \beta\sigma_{Xj}^2(\beta_0 - \beta))}{(\tau^2 + \sigma_{Yj}^2 + \sigma_{Xj}^2\beta^2)^2} \\
&= \sum_{j=1}^p \frac{(\beta_0 - \beta)\gamma_j^2}{\tau^2 + \sigma_{Yj}^2 + \sigma_{Xj}^2\beta^2}.
\end{aligned}$$

In other words, $\mathbb{E}[\tilde{\psi}(\beta, \tau^2)] = 0$ if and only if $\beta = \beta_0$.

Next we bound the variance of $\tilde{\psi}(\beta, \tau^2)$ over $\mathcal{B}$. Because of Assumptions 3 and 4, $\mathrm{Var}(\epsilon_j) = \Theta(1/n)$ and $\mathrm{Var}(e_j) = \Theta(1/n + 1/p)$. Using the inequality

$\mathrm{Var}(X+Y) \le 2[\mathrm{Var}(X) + \mathrm{Var}(Y)]$ repeatedly, we have

$$\mathrm{Var}(\tilde{\psi}(\beta,\tau^2)) = O\big(\mathrm{Var}(\psi_1(\beta,\tau^2))\big) + O\big(\mathrm{Var}(\psi_2(\beta,\tau^2))\big),$$

and, after some algebra,

$$\mathrm{Var}(\psi_1(\beta,\tau^2)) = O((n+p)\|\boldsymbol{\gamma}\|_2^2 + p) = o(n^2), \text{ and}$$
$$\mathrm{Var}(\psi_2(\beta,\tau^2)) = O((n+p)\|\boldsymbol{\gamma}\|_2^2 + p) = o(n^2).$$

To summarize, we have shown that

$$\tilde{\psi}(\beta,\tau^2) = (\beta_0 - \beta)\cdot\Theta(n) + o_p(n), \tag{B.3}$$
$$\psi_2(\beta,\tau^2) = (\beta_0-\beta)^2\Theta(n) + (p\tau_0^2 - p\tau)\Theta(n) + o_p(n). \tag{B.4}$$

Consider a box $\mathcal{B}' = [-C_1, C_1] \times [0, C_2]$ that contains $\mathcal{B}$. Using (B.3) and (B.4), if C_1 and C_2/C_1 are sufficiently large, all the following events have probabilities going to 1:

$$\sup_{|p\tau^2|\le C_2} \tilde{\psi}(C_1,\tau^2) \le 0, \quad \inf_{|p\tau^2|\le C_2} \tilde{\psi}(-C_1,\tau^2) \ge 0,$$
$$\sup_{|\beta|\le C_1} \psi_2(\beta, C_2/p) \le 0, \quad \inf_{|\beta|\le C_1} \psi_2(\beta, 0) \ge 0.$$

If all the events are true, by continuity of $\tilde{\psi}$ and ψ_2 and the Poincaré-Miranda theorem, there exists $(\hat{\beta}, p\hat{\tau}^2) \in \mathcal{B}'$ such that

$$\tilde{\psi}(\hat{\beta},\hat{\tau}^2) = \psi_2(\hat{\beta},\hat{\tau}^2) = 0.$$

Using (B.3) and (B.4), it is then straightforward to show $\hat{\beta} \xrightarrow{p} \beta_0$ and $(p\hat{\tau}^2 - p\tau_0^2) \xrightarrow{p} 0$. As a consequence, $(\hat{\beta}, p\hat{\tau}^2) \in \mathcal{B}$ with probability going to 1, thus concluding our proof.

B.5. Proof of Theorem 4.2. We begin with proving $\tilde{\boldsymbol{V}}_1$ and $\tilde{\boldsymbol{V}}_2$ are the corresponding moments of $\boldsymbol{\psi}(\beta_0,\tau_0^2)$ and $\nabla\boldsymbol{\psi}(\beta_0,\tau_0^2)$.

LEMMA B.4. $\mathrm{Var}(\boldsymbol{\psi}(\beta_0,\tau_0^2)) = \tilde{\boldsymbol{V}}_1$, $\mathbb{E}[\nabla\boldsymbol{\psi}(\beta_0,\tau_0^2)] = \tilde{\boldsymbol{V}}_2$.

PROOF OF LEMMA B.4. In Section 4.2 we have already shown that $\mathbb{E}[\boldsymbol{\psi}(\beta_0,\tau_0^2)] = \mathbf{0}$. The variance of $\psi_1(\beta_0,\tau_0^2)$ and the expectation of $(\partial/\partial\beta)\psi_1(\beta_0,\tau_0^2)$ can be obtained from the proof of Theorem 3.2 by replacing σ_{Yj}^2 with $\sigma_{Yj}^2 + \tau_0^2$.

Next we compute the other moments. Let $\psi_{1j}(\beta,\tau^2)$ and $\psi_{2j}(\beta,\tau^2)$ be the j-th summand in (4.2) and (4.3), so $\psi_i(\beta,\tau^2)=\sum_{j=1}^p \psi_{ij}(\beta,\tau^2)$ for $i=1,2$. Because $\hat\Gamma_j-\beta_0\hat\gamma_j \sim \mathrm{N}(0,\beta_0^2\sigma_{Xj}^2+\Sigma_{Yj}^2+\tau_0^2)$, it is easy to see that

$$\mathrm{Var}(\psi_{2j}(\beta_0,\tau_0^2)) = \frac{2(\sigma_{Xj}^2)^2}{(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)^2}$$

The covariance of ψ_{1j} and ψ_{2j} is given by

$$\begin{aligned}
&\mathrm{Cov}(\psi_{1j}(\beta_0,\tau_0^2),\psi_{2j}(\beta_0,\tau_0^2))\\
=&\mathbb{E}\Big[\frac{(\hat\Gamma_j-\beta_0\hat\gamma_j)(\hat\gamma_j(\tau_0^2+\sigma_{Yj}^2)+\hat\Gamma_j\sigma_{Xj}^2\beta_0)\sigma_{Xj}^2[(\hat\Gamma_j-\beta_0\hat\gamma_j)^2-(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)]}{(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)^4}\Big]\\
=&\sigma_{Xj}^2\mathbb{E}\Big[\frac{(\hat\Gamma_j-\beta_0\hat\gamma_j)^3(\hat\gamma_j(\sigma_{Yj}^2+\tau_0^2)+\hat\Gamma_j\sigma_{Xj}^2\beta_0)}{(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)^4}\Big]\\
=&\sigma_{Xj}^2\mathbb{E}\Big[\frac{(e_j-\beta_0\epsilon_j)^3[\gamma_j(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)+\epsilon_j(\sigma_{Yj}^2+\tau_0^2)+e_j\sigma_{Xj}^2\beta_0]}{(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)^4}\Big]\\
=&\sigma_{Xj}^2\mathbb{E}\Big[\frac{(e_j-\beta_0\epsilon_j)^3[\epsilon_j(\sigma_{Yj}^2+\tau_0^2)+e_j\sigma_{Xj}^2\beta_0]}{(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)^4}\Big]\\
=&\sigma_{Xj}^2\mathbb{E}\Big[\frac{e_j^4\sigma_{Xj}^2\beta_0+e_j^2\epsilon_j^2[-3\beta_0(\sigma_{Yj}^2+\tau_0^2)+3\beta_0^2\sigma_{Xj}^2\beta_0]-\epsilon_j^4\beta_0^3(\sigma_{Yj}^2+\tau_0^2)}{(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)^4}\Big]\\
=&0.
\end{aligned}$$

Thus $\mathrm{Var}(\boldsymbol{\psi}(\beta_0,\tau_0^2))=\tilde{\boldsymbol{V}}_1$. Next we consider the expectation of $\nabla\boldsymbol{\psi}(\beta_0,\tau_0^2)$:

$$\begin{aligned}
&\mathbb{E}\Big[\frac{\partial}{\partial\tau^2}\psi_{1j}(\beta_0^2,\tau_0^2)\Big]\\
=&\mathbb{E}\Big[\frac{(\hat\Gamma_j-\beta_0\hat\gamma_j)\hat\gamma_j(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)^2}{(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)^4}\Big]-\\
&-\mathbb{E}\Big[\frac{(\hat\Gamma_j-\beta_0\hat\gamma_j)(\hat\gamma_j(\sigma_{Yj}^2+\tau_0^2)+\hat\Gamma_j\sigma_{Xj}^2\beta_0)\cdot 2(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)}{(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)^4}\Big]\\
=&\mathbb{E}\Big[\frac{(\hat\Gamma_j-\beta_0\hat\gamma_j)[\hat\gamma_j(\sigma_{Xj}^2\beta_0^2-\sigma_{Yj}^2-\tau_0^2)-2\hat\Gamma_j\sigma_{Xj}^2\beta_0]}{(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)^3}\Big]\\
=&\mathbb{E}\Big[\frac{(e_j-\beta_0\epsilon_j)[\epsilon_j(\sigma_{Xj}^2\beta_0^2-\sigma_{Yj}^2-\tau_0^2)-2e_j\sigma_{Xj}^2\beta_0]}{(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)^3}\Big]\\
=&\frac{-2(\sigma_{Yj}^2+\tau_0^2)\sigma_{Xj}^2\beta_0-\beta_0\sigma_{Xj}^2(\sigma_{Xj}^2\beta_0^2-\sigma_{Yj}^2-\tau_0^2)}{(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)^3}\\
=&\frac{-\sigma_{Xj}^2\beta_0}{(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)^2}.
\end{aligned}$$

Furthermore,

$$\frac{\partial}{\partial \beta}\psi_{2j}(\beta_0,\tau_0) = \sigma_{Xj}^2 \frac{D_j - E_j}{(\sigma_{Xj}^2\beta_0^2 + \sigma_{Yj}^2 + \tau_0^2)^4}$$

where

$$D_j = [2(\hat{\Gamma}_j - \beta_0\hat{\gamma}_j)(-\hat{\gamma}_j) - 2\sigma_{Xj}^2\beta_0](\tau_0^2 + \sigma_{Yj}^2 + \sigma_{Xj}^2\beta_0^2)^2,$$

$$E_j = [(\hat{\Gamma}_j - \beta_0\hat{\gamma}_j)^2 - (\tau_0^2 + \sigma_{Yj}^2 + \sigma_{Xj}^2\beta_0^2)] \cdot 2(\tau_0^2 + \sigma_{Yj}^2 + \sigma_{Xj}^2\beta_0^2)\sigma_{Xj}^2 \cdot 2\beta_0.$$

It is not hard to see that both D_j and E_j have mean 0. Finally,

$$\begin{aligned}&\frac{\partial}{\partial \tau_0^2}\psi_{2j}(\beta_0,\tau_0)\\ &= -\frac{\sigma_{Xj}^2}{(\sigma_{Xj}^2\beta_0^2 + \sigma_{Yj}^2 + \tau_0^2)^2} - \frac{2\sigma_{Xj}^2[(\hat{\Gamma}_j - \beta_0\hat{\gamma}_j)^2 - (\sigma_{Xj}^2\beta_0^2 + \sigma_{Yj}^2 + \tau_0^2)]}{(\sigma_{Xj}^2\beta_0^2 + \sigma_{Yj}^2 + \tau_0^2)^3}\end{aligned}$$

It is easy to see that

$$\mathbb{E}\Big[\frac{\partial}{\partial \tau^2}\psi_{2j}(\beta_0,\tau_0)\Big] = -\frac{\sigma_{Xj}^2}{(\sigma_{Xj}^2\beta_0^2 + \sigma_{Yj}^2 + \tau_0^2)^2}.$$

In summary, we have proved that $\mathbb{E}[\nabla\boldsymbol{\psi}(\beta_0,\tau_0^2)] = -\tilde{\boldsymbol{V}}_2$. □

It is useful to write down the order of $\tilde{\boldsymbol{V}}_1$ and $\tilde{\boldsymbol{V}}_2$:

(B.5) $$\tilde{\boldsymbol{V}}_1 = \begin{pmatrix}\Theta(n) & 0\\ 0 & \Theta(n)\end{pmatrix},\ \tilde{\boldsymbol{V}}_2 = \begin{pmatrix}\Theta(n) & \Theta(n^2)\\ 0 & \Theta(n^2)\end{pmatrix}.$$

Similar to the proof of Theorem 3.2, consider the Taylor expansion (let $\boldsymbol{\theta} = (\beta,\tau^2)$)

(B.6)
$$\mathbf{0} = \boldsymbol{\psi}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\psi}(\boldsymbol{\theta}_0) + \nabla\psi(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \frac{1}{2}\begin{pmatrix}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T\partial^2\psi_1(\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\\ (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T\partial^2\psi_2(\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).\end{pmatrix}$$

By the consistency of $(\hat{\beta}, p\hat{\tau}^2)$ and three Lemmas listed after this paragraph, the third term on the right hand side is negligible compared to the second term. The central limit theorem (4.4) can then be proven by the same arguments (normalizing by $\tilde{\boldsymbol{V}}_1$ and $\tilde{\boldsymbol{V}}_2$ and using Slutsky's lemma) as in the beginning of the proof of Theorem 3.2.

LEMMA B.5. $(\tilde{\boldsymbol{V}}_1)^{-1/2}\boldsymbol{\psi}(\beta_0,\tau_0^2) \overset{d}{\to} \mathrm{N}(\mathbf{0}, \boldsymbol{I}_2)$.

LEMMA B.6. $(\tilde{\boldsymbol{V}}_2)^{-1}\nabla\boldsymbol{\psi}(\beta_0,\tau_0^2) \xrightarrow{p} -\boldsymbol{I}_2$.

LEMMA B.7. *Denote $\partial^2\psi_i(\beta,\tau^2)$ to be all the second-order partial derivatives of $\psi_i(\beta,\tau^2)$, $i=1,2$. For a neighborhood $\mathcal{N}$ of $(\beta_0, p\tau_0^2)$ and $l=0,1,2$,*

$$\sup_{(\beta,\tau^2)\in\mathcal{N}} \Big|\frac{\partial^2}{\partial\beta^{2-l}(\partial\tau^2)^l}\psi_1(\beta,\tau^2)\Big| = O_p(n^{l+1}), \text{ and}$$
$$\sup_{(\beta,\tau^2)\in\mathcal{N}} \Big|\frac{\partial^2}{\partial\beta^{2-l}(\partial\tau^2)^l}\psi_2(\beta,\tau^2)\Big| = O_p(n^{l+1}).$$

Next we prove Lemmas B.5 to B.7. Let $\boldsymbol{\psi}_j(\beta,\tau^2) = (\psi_{1j}(\beta,\tau^2), \psi_{2j}(\beta,\tau^2))$ for $j\in[p]$. Since $\boldsymbol{\psi}_j, j\in[p]$ are mutually independent and $p\to\infty$, it suffices to verify the following Lyapunov condition [6]

$$\sum_{j=1}^{p}\mathbb{E}\big[\|\tilde{\boldsymbol{V}}_1^{-1/2}\boldsymbol{\psi}_j(\beta_0,\tau_0^2)\|^3\big] \to 0$$

to prove Lemma B.5. Because $\tilde{\boldsymbol{V}}_1$ is diagonal, it suffices to verify this for each coordinate of $\boldsymbol{\psi}_j$. Similar to the proof of Lemma B.1, we can show that

$$\sum_{j=1}^{p}\mathbb{E}[|\psi_{1j}(\beta_0,\tau_0^2)|^3] = O(p+\sqrt{n}\|\boldsymbol{\gamma}\|_1 + n\|\boldsymbol{\gamma}\|_2^2 + n^{3/2}\|\boldsymbol{\gamma}\|_3^3). \tag{B.7}$$

Therefore, using $\|\boldsymbol{\gamma}\|_3/\|\boldsymbol{\gamma}\|_2 \to 0$, we obtain

$$\frac{\sum_{j=1}^{p}\mathbb{E}[|\psi_{1j}(\beta_0)|^3]}{(\tilde{\boldsymbol{V}}_1)_{11}^{3/2}} = O\Big(\frac{p+\sqrt{n}\|\boldsymbol{\gamma}\|_1 + n\|\boldsymbol{\gamma}\|_2^2 + n^{3/2}\|\boldsymbol{\gamma}\|_3^3}{n^{3/2}}\Big) \to 0.$$

For ψ_2, since the third moment of a χ_1^2 distribution exists,

$$\mathbb{E}[|\psi_{2j}(\beta_0,\tau_0^2)|^3] = O\Big(\frac{\sigma_{Xj}^6}{(\sigma_{Xj}^2\beta_0^2 + \sigma_{Yj}^2 + \tau_0^2)^3}\Big) = O(1).$$

Thus
$$\frac{\sum_{j=1}^{p}\mathbb{E}[|\psi_{2j}(\beta_0)|^3]}{(\tilde{\boldsymbol{V}}_1)_{22}^{3/2}} = O\Big(\frac{p}{n^{3/2}}\Big) = O(n^{-1/2}) \to 0.$$

This completes our proof of Lemma B.5.

For Lemmas B.6 and B.7, it remains to bound the variance of $\nabla\psi$ and $\partial^2\psi$. Notice that, similar to the proof of Lemmas B.2 and B.3, $\psi_{1j}(\beta,\tau^2)$ is a homogeneous quadratic polynomial of $(\tilde{\gamma}_j,\tilde{e}_j,\tilde{\epsilon}_j)=(\gamma_j/\sqrt{n},e_j/\sqrt{n},\epsilon_j/\sqrt{n})$:

$$\begin{aligned}\psi_{1j}(\beta)=&\frac{\big[(\beta_0-\beta)\tilde{\gamma}_j+\tilde{e}_j-\beta\tilde{\epsilon}_j\big]}{[n\sigma_{Xj}^2\beta^2+n(\sigma_{Yj}^2+\tau^2)]^2}\cdot\\&\cdot\big[(n\sigma_{Xj}^2\beta\beta_0+n(\sigma_{Yj}^2+\tau^2))\tilde{\gamma}_j+(n\sigma_{Xj}^2\beta)\tilde{e}_j+(n\sigma_{Yj}^2)\tilde{\epsilon}_j\big].\end{aligned}$$

Therefore, its derivatives with respect to β and τ^2 remain to be homogeneous quadratic polynomials. As in the proof of Lemma B.2, this suggests that, for $(\beta,p\tau^2)\in\mathcal{B}$, (recall that $\|\boldsymbol{\gamma}\|_4\le\|\boldsymbol{\gamma}\|_3\ll\|\boldsymbol{\gamma}\|_2$)

$$\begin{aligned}\operatorname{Var}\Big(\frac{\partial}{\partial\beta}\psi_1(\beta,\tau^2)\Big)&\le\mathbb{E}\Big[\Big(\frac{\partial}{\partial\beta}\psi_1(\beta,\tau^2)\Big)^2\Big]\\&=O\big(n^2\|\boldsymbol{\gamma}\|_4^4+n\|\boldsymbol{\gamma}\|_2^2+p\big)=o((\tilde{\boldsymbol{V}}_2)_{11}^2).\end{aligned}\tag{B.8}$$

Therefore $(\tilde{\boldsymbol{V}}_2)_{11}^{-1}(\partial/\partial\beta)\psi_1(\beta_0,\tau_0^2)\xrightarrow{p}-1$, where $(\boldsymbol{V})_{ij}^{-1}$ means the reciprocal of the (i,j)-th entry of $\boldsymbol{V}$. Similarly,

$$\operatorname{Var}\Big((\partial/\partial\tau^2)\psi_1(\beta,\tau^2)\Big)=O\big(n[n^2\|\boldsymbol{\gamma}\|_4^4+n\|\boldsymbol{\gamma}\|_2^2+p]\big)=o((\tilde{\boldsymbol{V}}_2)_{12}^2).\tag{B.9}$$

The extra n comes from differentiating with respect to $\tau^2=O(1/p)=O(1/n)$. So $((\tilde{\boldsymbol{V}}_2)_{12})^{-1}(\partial/\partial\tau^2)\psi_1(\beta,\tau^2)\xrightarrow{p}-1$.

For ψ_2, we have

$$\psi_{2j}(\beta,\tau^2)=\frac{\sigma_{Xj}^2}{\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2}\Bigg[\frac{\big[(\beta_0-\beta)\tilde{\gamma}_j+\tilde{e}_j-\beta\tilde{\epsilon}_j\big]^2}{n\sigma_{Xj}^2\beta^2+n(\sigma_{Yj}^2+\tau^2)}-1\Bigg].$$

Using the same argument,

$$\operatorname{Var}\Big((\partial/\partial\beta)\psi_2(\beta,\tau^2)\Big)=O\big(n^2\|\boldsymbol{\gamma}\|_4^4+n\|\boldsymbol{\gamma}\|_2^2+p\big)=o((\tilde{\boldsymbol{V}}_2)_{11}),\tag{B.10}$$

$$\operatorname{Var}\Big((\partial/\partial\tau^2)\psi_2(\beta,\tau^2)\Big)\le O\big(n[n^2\|\boldsymbol{\gamma}\|_4^4+n\|\boldsymbol{\gamma}\|_2^2+p]\big)=o((\tilde{\boldsymbol{V}}_2)_{22}^2).\tag{B.11}$$

Therefore $((\tilde{\boldsymbol{V}}_2)_{22})^{-1}(\partial/\partial\tau^2)\psi_2(\beta,\tau^2)\xrightarrow{p}-1$. We cannot claim the same conclusion for $(\partial/\partial\tau^2)\psi_2(\beta,\tau^2)$ because $(\tilde{\boldsymbol{V}}_2)_{21}=0$. Nevertheless, the above results are already enough to verify Lemma B.6, because

$$\begin{aligned}(\tilde{\boldsymbol{V}}_2)^{-1}\nabla\boldsymbol{\psi}(\beta_0,\tau_0^2)&=\begin{pmatrix}(\tilde{\boldsymbol{V}}_2)_{11}^{-1}&-(\tilde{\boldsymbol{V}}_2)_{11}^{-1}(\tilde{\boldsymbol{V}}_2)_{22}^{-1}(\tilde{\boldsymbol{V}}_2)_{12}\\0&(\tilde{\boldsymbol{V}}_2)_{22}^{-1}\end{pmatrix}\begin{pmatrix}\frac{\partial}{\partial\beta}\psi_1&\frac{\partial}{\partial\tau^2}\psi_1\\\frac{\partial}{\partial\beta}\psi_2&\frac{\partial}{\partial\tau^2}\psi_2\end{pmatrix}\\&=\begin{pmatrix}(\tilde{\boldsymbol{V}}_2)_{11}^{-1}\frac{\partial}{\partial\beta}\psi_1-L\frac{\partial}{\partial\beta}\psi_2&(\tilde{\boldsymbol{V}}_2)_{11}^{-1}\frac{\partial}{\partial\tau^2}\psi_1-L\frac{\partial}{\partial\tau^2}\psi_2\\(\tilde{\boldsymbol{V}}_2)_{22}^{-1}\frac{\partial}{\partial\beta}\psi_2&(\tilde{\boldsymbol{V}}_2)_{22}^{-1}\frac{\partial}{\partial\tau^2}\psi_2\end{pmatrix},\end{aligned}$$

where $L = (\tilde{\boldsymbol{V}}_2)_{11}^{-1}(\tilde{\boldsymbol{V}}_2)_{22}^{-1}(\tilde{\boldsymbol{V}}_2)_{12} = \Theta(n^{-1})$. Using equations (B.8) to (B.11), it is straightforward to verify that the right hand side converges to $\boldsymbol{I}_2$ in probability.

Finally, Lemma B.7 can be proven similarly to Lemma B.3 using the rate of the variances established above as they also extend to the second-order derivative of ψ_2.

B.6. Proof of Theorem 5.1. It is easy to show $\mathbb{E}[\psi_2^{(\rho)}(\beta_0, \tau_0^2)] = 0$ by using $t_j(\beta_0, \tau_0^2) \sim \mathrm{N}(0, 1)$. For $\psi_1^{(\rho)}$, since

$$u_j(\beta, \tau^2) = -\frac{\partial}{\partial \beta} t_j(\beta, \tau^2) = \frac{\hat{\Gamma}_j \sigma_{Xj}^2 \beta + \hat{\gamma}_j(\sigma_{Yj}^2 + \tau^2)}{(\sigma_{Xj}^2 \beta^2 + \sigma_{Yj}^2 + \tau^2)^{3/2}}, \tag{B.12}$$

it is straightforward to verify that

$$\mathbb{E}\big[t_j(\beta_0, \tau_0^2) \cdot u_j(\beta_0, \tau_0^2)\big] = \mathbb{E}\Bigg[\frac{(\hat{\Gamma}_j - \beta_0 \hat{\gamma}_j)\big[\hat{\Gamma}_j \sigma_{Xj}^2 \beta_0 + \hat{\gamma}_j(\sigma_{Yj}^2 + \tau_0^2)\big]}{(\sigma_{Xj}^2 \beta_0^2 + \sigma_{Yj}^2 + \tau_0^2)^2}\Bigg] = 0.$$

Since $t_j(\beta_0, \tau_0^2)$ and $u_j(\beta_0, \tau_0^2)$ are linear transformations of jointly normal random variables, this implies that $t_j(\beta_0, \tau_0^2) \perp\!\!\!\perp u_j(\beta_0, \tau_0^2)$. Therefore

$$\mathbb{E}[\psi_{1j}^{(\rho)}(\beta_0, \tau_0^2)] \propto \mathbb{E}[\rho'(t_j(\beta_0, \tau_0^2)) \cdot u_j(\beta_0, \tau_0^2)] = 0.$$

By Lemma B.8 below, $\mathrm{E}[\nabla \boldsymbol{\psi}^{(\rho)}] = -\tilde{\boldsymbol{V}}_2^{(\rho)}$ has full rank because $\delta, c_3 > 0$.

B.7. Proof of Theorem 5.2. Similar to the proof of Theorem 4.2, we first show $\tilde{\boldsymbol{V}}_1^{(\rho)}$ is the variance of $\boldsymbol{\psi}^{(\rho)}$ and $\tilde{\boldsymbol{V}}_2^{(\rho)}$ is the expectation of $-\nabla \boldsymbol{\psi}^{(\rho)}$ at the true parameter $(\beta, \tau^2) = (\beta_0, \tau_0^2)$.

Lemma B.8. $\mathrm{Var}\big(\boldsymbol{\psi}^{(\rho)}(\beta_0, \tau_0^2)\big) = \tilde{\boldsymbol{V}}_1^{(\rho)}$, $\mathbb{E}\big[\nabla \boldsymbol{\psi}^{(\rho)}(\beta_0, \tau_0^2)\big] = -\tilde{\boldsymbol{V}}_2^{(\rho)}$.

Proof of Lemma B.8. Let $\psi_{1j}^{(\rho)}$ and $\psi_{2j}^{(\rho)}$ be the j-th summand in (5.2) and (5.3). We will use the shorthand notation $t_{j0} = t_j(\beta_0, \tau_0^2)$ and $u_{j0} = u_j(\beta_0, \tau_0^2)$.

Because ρ' is an odd function and $t_{j0} \sim \mathrm{N}(0, 1)$, we have $\mathbb{E}[\rho'(t_{j0})] = 0$. Using $t_{j0} \perp\!\!\!\perp u_{j0}$ (see Appendix B.6), we obtain

$$\mathrm{Var}\big(\psi_{1j}^{(\rho)}(\beta_0, \tau_0^2)\big) = \mathrm{Var}\big(\rho'(t_{j0}) u_{j0}\big) = c_1 \mathbb{E}\big[u_{j0}^2\big].$$

If we let $\rho(r) = r^2/2$ be the l_2-loss (so $c_1 = 1$), we recover the APS so $\mathbb{E}[u_{j0}] = (\tilde{\boldsymbol{V}}_1)_{11}$. Thus $\mathrm{Var}\big(\psi_{1j}^{(\rho)}(\beta_0, \tau_0^2)\big) = c_1 (\tilde{\boldsymbol{V}}_1)_{11}$.

The covariance of $\psi_{1j}^{(\rho)}$ and $\psi_{2j}^{(\rho)}$ is

$$\begin{aligned}&\operatorname{Cov}\big(\psi_{1j}^{(\rho)}(\beta_0,\tau_0^2),\psi_{2j}^{(\rho)}(\beta_0,\tau_0^2)\big)\\ \propto&\mathbb{E}\Big\{\rho'(t_{j0})u_{j0}\cdot\big[t_{j0}\rho'(t_{j0})-\delta\big]\Big\}\\ =&\mathbb{E}[u_{j0}]\cdot\mathbb{E}\Big\{\rho'(t_{j0})\cdot\big[t_{j0}\rho'(t_{j0})-\delta\big]\Big\}\\ =&0.\end{aligned}$$

The last expectation is 0 because ρ' is an odd function.

The variance of $\psi_{2j}^{(\rho)}$ is

$$\begin{aligned}\operatorname{Var}\big(\psi_{2j}^{(\rho)}(\beta_0,\tau_0^2)\big)&=\frac{(\sigma_{Xj}^2)^2}{(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)^2}\operatorname{Var}\big(t_{j0}\rho'(t_{j0})\big)\\ &=c_2(\tilde{\boldsymbol{V}}_1)_{22}.\end{aligned}$$

Next we turn to the derivatives of $\boldsymbol{\psi}^{(\rho)}$. First, by the chain rule,

$$\frac{\partial}{\partial\beta}\psi_{1j}^{(\rho)}(\beta,\tau^2)=-\rho''(t_j(\beta,\tau^2))u_j(\beta,\tau^2)^2+\rho'(t_j(\beta,\tau^2))\frac{\partial}{\partial\beta}u_j(\beta,\tau^2).$$

The expectation of the first term at (β_0,τ_0^2) is

$$\mathbb{E}[-\rho''(t_{j0})u_{j0}^2]=-\mathbb{E}[\rho''(t_{j0})]\mathbb{E}[u_{j0}^2]=\delta(\tilde{\boldsymbol{V}}_2)_{11},$$

where we have used the identity $\mathbb{E}[\rho''(R)]=\mathbb{E}[R\rho'(R)]$ for $R\sim\mathrm{N}(0,1)$, which can be proved by integration by parts and the fact that $\phi'(x)=-x\phi(x)$.

The second term requires more calculations:

$$
\begin{aligned}
&\mathbb{E}\Big[[\rho'(t_{j0})\cdot\Big(\frac{\partial}{\partial\beta}u_j(\beta,\tau^2)\Big)\Big|_{(\beta,\tau^2)=(\beta_0,\tau_0^2)}\Big]\\
=&\mathbb{E}\Bigg[\rho'(t_{j0})\cdot\frac{\partial}{\partial\beta}\Big(\hat{\gamma}_j(\tau_0^2+\sigma_{j2}^2)+\hat{\Gamma}_j\sigma_{j1}^2\beta\Big)\Big|_{\beta=\beta_0}\cdot\frac{1}{(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)^{3/2}}\Bigg]\\
&+\mathbb{E}\Bigg[\rho'(t_{j0})\cdot(\hat{\gamma}_j(\tau_0^2+\sigma_{j2}^2)+\hat{\Gamma}_j\sigma_{j1}^2\beta_0)\cdot\frac{\partial}{\partial\beta}\Big(\frac{1}{(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)^{3/2}}\Big)\Big|_{\beta=\beta_0}\Bigg]\\
=&\mathbb{E}\Bigg[\rho'(t_{j0})\cdot\frac{\partial}{\partial\beta}\Big(\hat{\gamma}_j(\tau_0^2+\sigma_{j2}^2)+\hat{\Gamma}_j\sigma_{j1}^2\beta\Big)\Big|_{\beta=\beta_0}\cdot\frac{1}{(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)^{3/2}}\Bigg]\\
=&\mathbb{E}\Bigg[\rho'(t_{j0})\cdot\hat{\Gamma}_j\sigma_{j1}^2\cdot\frac{1}{(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)^{3/2}}\Bigg]\\
=&\frac{\sigma_{Xj}^2}{(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)^{3/2}}\mathbb{E}[\rho'(t_{j0})(\Gamma_j+e_j)]\\
=&\frac{\sigma_{Xj}^2\sqrt{\sigma_{Yj}^2+\tau_0^2}}{(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)^{3/2}}\mathbb{E}\Bigg[\rho'(t_{j0})\cdot\frac{e_j}{\sqrt{\tau_0^2+\sigma_{j2}^2}}\Bigg].
\end{aligned}
$$

The second equality above is because $t_{j0} \perp\!\!\!\perp u_{j0}$ and $\mathbb{E}[\rho'(t_{j0})] = 0$. Notice that t_{j0} and $e_j/\sqrt{\tau_0^2+\sigma_{j2}^2}$ are marginally distributed as the standard normal and

$$
\mathrm{Cov}\Bigg(t_{j0},\frac{e_j}{\sqrt{\sigma_{Yj}^2+\tau_0^2}}\Bigg)=\Big(\frac{\sigma_{Yj}^2+\tau_0^2}{\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2}\Big)^{1/2}.
$$

It is not difficult to verify that if R_1, R_2 are $\mathrm{N}(0,1)$ marginally and $\mathrm{Cov}(R_1,R_2) = \lambda$, then $\mathbb{E}[\rho'(R_1)R_2] = \lambda\delta$. Thus

$$
\text{(B.13)}\quad \mathbb{E}\Big[[\rho'(t_{j0})\cdot\frac{\partial}{\partial\beta}u_{j0}\Big]=\delta\cdot\frac{\sigma_{Xj}^2(\sigma_{Yj}^2+\tau_0^2)}{(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)^2}=\delta[(\tilde{\boldsymbol{V}}_1)_{11}-(\tilde{\boldsymbol{V}}_2)_{11}].
$$

To summarize,

$$
\mathbb{E}\Big[\frac{\partial}{\partial\beta}\psi_{1j}^{(\rho)}(\beta_0,\tau_0^2)\Big]=-\delta(\tilde{\boldsymbol{V}}_1)_{11}+\delta[(\tilde{\boldsymbol{V}}_1)_{11}-(\tilde{\boldsymbol{V}}_2)_{11}]=-\delta(\tilde{\boldsymbol{V}}_2)_{11}.
$$

The other first-order derivative of $\psi_{1j}^{(\rho)}$ is

$$
\mathbb{E}\Big[\frac{\partial}{\partial\tau^2}\psi_{1j}^{(\rho)}(\beta_0,\tau_0^2)\Big]=\mathbb{E}\Big[\rho''(t_{j0})\Big(\frac{\partial}{\partial\tau^2}t_{j0}\Big)u_{j0}+\rho'(t_{j0})\Big(\frac{\partial}{\partial\tau^2}u_{j0}\Big)\Big].
$$

Using

$$\frac{\partial}{\partial \tau^2} t_j(\beta, \tau^2) = -\frac{t_j(\beta, \tau^2)}{2(\sigma_{Xj}^2 \beta_0^2 + \sigma_{Yj}^2 + \tau_0^2)} \tag{B.14}$$

and the independent of t_{j0}, it is straightforward to show the first term has mean 0. For the second term,

$$\mathbb{E}\Big[\rho'(t_{j0})\Big(\frac{\partial}{\partial \tau^2} u_{j0}\Big)\Big] = \mathbb{E}\Big[\rho'(t_{j0}) \cdot \hat{\gamma}_j \frac{1}{(\sigma_{Xj}^2 \beta_0^2 + \sigma_{Yj}^2 + \tau_0^2)^{3/2}}\Big]$$

Similar to the derivation of (B.13), one can show that

$$\mathbb{E}\Big[\rho'(t_{j0})\Big(\frac{\partial}{\partial \tau^2} u_{j0}\Big)\Big] = \frac{-\delta \beta_0 \sigma_{Xj}^2}{\sigma_{Xj}^2 \beta_0^2 + \sigma_{Yj}^2 + \tau_0^2} = -\delta \cdot (\tilde{\boldsymbol{V}}_2)_{12}.$$

Finally we consider the derivatives of $\psi_{2j}^{(\rho)}$. Using (B.14), we have

$$\begin{aligned}\mathbb{E}\Big[\frac{\partial}{\partial \tau^2} \psi_{2j}^{(\rho)}(\beta_0, \tau_0^2)\Big] &= \mathbb{E}\Big[\sigma_{Xj}^2 \frac{-t_{j0}\rho'(t_{j0})/2 - t_{j0}^2 \rho''(t_{j0})/2 - (t_{j0}\rho'(t_{j0}) - \delta)}{(\sigma_{Xj}^2 \beta_0^2 + \sigma_{Yj}^2 + \tau_0^2)^2}\Big] \\ &= -\frac{\delta + c_3}{2} \cdot \frac{\sigma_{Xj}^2}{(\sigma_{Xj}^2 \beta_0^2 + \sigma_{Yj}^2 + \tau_0^2)^2}\end{aligned}$$

Hence

$$\mathbb{E}\Big[\frac{\partial}{\partial \tau^2} \psi_2^{(\rho)}(\beta_0, \tau_0^2)\Big] = -[(\delta + c_3)/2](\tilde{\boldsymbol{V}}_2)_{22}.$$

The last partial derivative is $(\partial/\partial\beta)\psi_2^{(\rho)}$. Its expectation at (β_0, τ_0^2) is

$$\begin{aligned}&\mathbb{E}\Big[\frac{\partial}{\partial \beta} \psi_{2j}^{(\rho)}(\beta_0, \tau_0^2)\Big] \\ =&\mathbb{E}\Big[\frac{\frac{\partial}{\partial \beta}\Big[t_{j0}\rho'(t_{j0}) - \delta\Big]}{(\sigma_{Xj}^2 \beta_0^2 + \sigma_{Yj}^2 + \tau_0^2)^2}\Big] \\ &+ \mathbb{E}\Big[\frac{\partial}{\partial \beta}\Big[\frac{1}{(\sigma_{Xj}^2 \beta^2 + \sigma_{Yj}^2 + \tau_0^2)^2}\Big]\Big|_{\beta = \beta_0} \cdot \Big[t_{j0}\rho'(t_{j0}) - \delta\Big]\Big] \\ =&\mathbb{E}\Big[\frac{\frac{\partial}{\partial \beta}\Big[t_{j0}\rho'(t_{j0}) - \delta\Big]}{(\sigma_{Xj}^2 \beta_0^2 + \sigma_{Yj}^2 + \tau_0^2)^2}\Big] \\ =&\mathbb{E}\Big[\frac{-u_{j0}\rho'(t_{j0}) - t_{j0}\rho''(t_{j0}) u_{j0}}{(\sigma_{Xj}^2 \beta_0^2 + \sigma_{Yj}^2 + \tau_0^2)^2}\Big] \\ =&0.\end{aligned}$$

The last equation is due to the independence of t_{j0} and u_{j0} and the fact that $\rho'(r)$ and $r\rho''(r)$ are odd functions of r. □

To prove asymptotic normality of the RAPS estimator, we just need to verify Lemmas B.5 to B.7 with $\boldsymbol{\psi}$ replaced by $\boldsymbol{\psi}^{(\rho)}$ and $\tilde{\boldsymbol{V}}_1$, $\tilde{\boldsymbol{V}}_2$ replaced by $\tilde{\boldsymbol{V}}_1^{(\rho)}$, $\tilde{\boldsymbol{V}}_2^{(\rho)}$. This requires verifying the Lyapunov condition for the central limit theorem in Lemma B.5 and bounding the derivatives of $\boldsymbol{\psi}^{(\rho)}$ to prove Lemmas B.6 and B.7.

It is useful to notice that the rates of $\tilde{\boldsymbol{V}}_1$ and $\tilde{\boldsymbol{V}}_2$ in (B.5) still apply to $\tilde{\boldsymbol{V}}_1^{(\rho)}$ and $\tilde{\boldsymbol{V}}_2^{(\rho)}$. First, using the boundedness of ρ', we have

$$\sum_{j=1}^{p} \mathbb{E}\big[|\psi_{1j}^{(\rho)}(\beta_0,\tau_0^2)|^3\big] = \sum_{j=1}^{p} \mathbb{E}[|\rho'(t_{j0})|^3]\cdot \mathbb{E}[|u_{j0}|^3] = O\Big(\sum_{j=1}^{p} \mathbb{E}[|u_{j0}|^3]\Big).$$

We can rewrite (B.12) as

$$u_{j0} = \frac{(\gamma_j\sqrt{n})(\sigma_{Xj}^2\beta_0+\sigma_{Yj}^2+\tau_0^2)+(e_j\sqrt{n})(\sigma_{Xj}^2\beta_0)+(\epsilon_j\sqrt{n})(\sigma_{Yj}^2+\tau_0^2)}{(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)^{3/2}\sqrt{n}}.$$

So u_{j0} is a linear combination of $\gamma_j\sqrt{n}$, $e_j\sqrt{n}$, $\epsilon_j\sqrt{n}$. Using the same argument in Appendix B.5, for any positive integer k,

$$\mathbb{E}[|u_{j0}|^k] = O\Big(\sum_{l=0}^{k}\big(\sqrt{n}|\gamma_j|+1\big)^l\Big) = O\Big(\sum_{l=0}^{k}(\sqrt{n})^l|\gamma_j|^l\Big). \tag{B.15}$$

From this it is easy to verify equation (B.7) still holds for $\psi_{1j}^{(\rho)}(\beta_0,\tau_0^2)$, $j=1,2,\dots$, which implies that

$$\frac{\sum_{j=1}^{p}\mathbb{E}\big[\big|\psi_{1j}^{(\rho)}(\beta_0,\tau_0^2)\big|^3\big]}{(\tilde{\boldsymbol{V}}_1^{(\rho)})_{11}^{3/2}} \to 0.$$

Furthermore,

$$\sum_{j=1}^{p}\mathbb{E}\big[|\psi_{2j}^{(\rho)}(\beta_0,\tau_0^2)|^3\big] = \sum_{j=1}^{p}\frac{(\sigma_{Xj}^2)^3}{(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)^3}\cdot c_5$$

where $c_5 = \mathrm{E}[|R\rho'(R)-\delta|^3]$ for $R\sim \mathrm{N}(0,1)$. Thus

$$\sum_{j=1}^{p}\mathbb{E}\big[|\psi_{2j}^{(\rho)}(\beta_0,\tau_0^2)|^3\big] = O(p) \ll (\tilde{\boldsymbol{V}}_1^{(\rho)})_{22}^{3/2}.$$

To summarize, we have verified the Lyapunov condition for $\boldsymbol{\psi}_j^{(\rho)}$. Consequently, the central limit theorem $(\tilde{\boldsymbol{V}}_1^{(\rho)})^{-1/2}\boldsymbol{\psi}^{(\rho)}(\beta_0,\tau_0^2) \xrightarrow{d} \mathrm{N}(\mathbf{0},\boldsymbol{I}_2)$ holds.

Next we restablish the variance bounds, namely equations (B.8) to (B.11), for $\boldsymbol{\psi}^{(\rho)}$. Similar to Appendix B.5, we extend (B.15), the bound on the moments of u_j, to the derivatives of u_j: for $(\beta, p\tau^2) \in \mathcal{B}$,

$$\mathbb{E}\Big[\Big|\frac{\partial^{l_1+l_2}}{\partial\beta^{l_1}(\partial\tau^2)^{l_2}}u_j(\beta,\tau^2)\Big|^k\Big] = O\Big(n^{l_2}\sum_{l=0}^{k}(\sqrt{n})^l|\gamma_j|^l\Big). \tag{B.16}$$

Similarly,

$$\mathbb{E}\big[\big|t_j(\beta,\tau^2)\big|^k\big] = O\Big(\sum_{l=0}^{k}(\sqrt{n})^l|\gamma_j|^l\Big). \tag{B.17}$$

Consider a partial derivative of $\psi_{1j}^{(\rho)}$:

$$\frac{\partial^{l_1+l_2}}{\partial\beta^{l_1}(\partial\tau^2)^{l_2}}\psi_{1j}^{(\rho)}(\beta,\tau^2) = \frac{\partial^{l_1+l_2}}{\partial\beta^{l_1}(\partial\tau^2)^{l_2}}\rho'(t_j(\beta,\tau^2))u_j(\beta,\tau^2).$$

By equations (B.12) and (B.14), It is a polynomial of derivatives (up to (l_1+l_2+1)-th order) of $\rho(t_j(\beta,\tau^2))$, $t_j(\beta,\tau^2)$, $u_j(\beta,\tau^2)$, and derivatives of $u_j(\beta,\tau^2)$, for which we all have moment bounds. We will use the shorthand notation $t_j = t_j(\beta,\tau^2)$ and $u_j = u_j(\beta,\tau^2)$ below. In particular,

$$\frac{\partial}{\partial\beta}\psi_{1j}^{(\rho)}(\beta,\tau^2) = -\rho''(t_j)u_j^2 + \rho'(t_j)\frac{\partial}{\partial\beta}u_j.$$

Using the boundedness of ρ' and ρ'' and equation (B.16), we have

$$\begin{aligned}
\mathrm{Var}\Big(\frac{\partial}{\partial\beta}\psi_1^{(\rho)}\Big) &\le \mathrm{E}\Big[\Big(\frac{\partial}{\partial\beta}\psi_1^{(\rho)}\Big)^2\Big] \\
&= O\Big(\sum_{j=1}^{p}u_j^4 + \Big(\frac{\partial}{\partial\beta}u_j\Big)^2\Big) \\
&= O\Big(\sum_{l=0}^{4}(\sqrt{n})^l\|\boldsymbol{\gamma}\|_l^l\Big) = o\big((\tilde{\boldsymbol{V}}_2^{(\rho)})_{11}^2\big)
\end{aligned}$$

Similarly,

$$\frac{\partial}{\partial\tau^2}\psi_{1j}^{(\rho)} = -\rho''(t_j)\frac{t_ju_j}{2(\sigma_{Xj}^2\beta_0^2+\sigma_{Yj}^2+\tau_0^2)} + \rho'(t_j)\frac{\partial}{\partial\tau^2}u_j.$$

Using the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}
\mathrm{Var}\Big(\frac{\partial}{\partial\tau^2}\psi_1^{(\rho)}\Big) &\le \mathrm{E}\Big[\Big(\frac{\partial}{\partial\tau^2}\psi_1^{(\rho)}\Big)^2\Big] \\
&= O\bigg(\mathbb{E}\Big[\sum_{j=1}^p \frac{t_j^4+u_j^4}{(1/n)^2} + \Big(\frac{\partial}{\partial\tau^2}u_j\Big)^2\Big]\bigg) \\
&= O\Big(n^2\sum_{l=0}^4(\sqrt{n})^l\|\boldsymbol{\gamma}\|_l^l + n\sum_{l=0}^2(\sqrt{n})^l\|\boldsymbol{\gamma}\|_l^l\Big) = o\big((\tilde{\boldsymbol{V}}_2^{(\rho)})_{12}^2\big).
\end{aligned}$$

Next we consider the derivatives of $\psi_2^{(\rho)}$:

$$\begin{aligned}
\frac{\partial}{\partial\beta}\psi_{2j}^{(\rho)} =& \frac{\sigma_{Xj}^2}{\sigma_{Xj}^2\beta^2+\sigma_{Yj}^2+\tau^2}\cdot\Big[-u_j\rho'(t_j) - t_j\rho''(t_j)u_j\Big] - \\
&- \frac{\sigma_{Xj}^4\beta}{(\sigma_{Xj}^2\beta^2+\sigma_{Yj}^2+\tau^2)^2}\cdot[t_j\rho'(t_j)-\delta],
\end{aligned}$$

thus, again using the Cauchy-Schwarz inequality,

$$\begin{aligned}
\mathrm{E}\Big[\Big(\frac{\partial}{\partial\beta}\psi_2^{(\rho)}\Big)^2\Big] &= O\bigg(\mathbb{E}\Big[\sum_{j=1}^p u_j^2 + (t_j^4+u_j^4) + t_j^2\Big]\bigg) \\
&= O\Big(\sum_{l=0}^4(\sqrt{n})^l\|\boldsymbol{\gamma}\|_l^l\Big) = o\big((\tilde{\boldsymbol{V}}_2^{(\rho)})_{22}\big).
\end{aligned}$$

Finally,

$$\begin{aligned}
\frac{\partial}{\partial\tau^2}\psi_{2j}^{(\rho)} =& \frac{\sigma_{Xj}^2}{\sigma_{Xj}^2\beta^2+\sigma_{Yj}^2+\tau^2}\cdot\Big[-\frac{t_j}{2(\sigma_{Xj}^2\beta^2+\sigma_{Yj}^2+\tau^2)}(\rho'(t_j)-t_j\rho''(t_j))\Big] - \\
&- \frac{\sigma_{Xj}^2}{(\sigma_{Xj}^2\beta^2+\sigma_{Yj}^2+\tau^2)^2}\cdot[t_j\rho'(t_j)-\delta].
\end{aligned}$$

Thus

$$\begin{aligned}
\mathrm{E}\Big[\Big(\frac{\partial}{\partial\tau^2}\psi_2^{(\rho)}\Big)^2\Big] &= O\bigg(\frac{1}{1/n}\mathbb{E}\Big[\sum_{j=1}^p t_j^2 + t_j^4\Big]\bigg) \\
&= O\Big(n\sum_{l=0}^4(\sqrt{n})^l\|\boldsymbol{\gamma}\|_l^l\Big) \\
&= o\big((\tilde{\boldsymbol{V}}_2^{(\rho)})_{22}^2\big).
\end{aligned}$$

In summary, we have re-established equations (B.8) to (B.11) for $\boldsymbol{\psi}^{(\rho)}$. Therefore Lemma B.6 still holds for $\boldsymbol{\psi}^{(\rho)}$ with $\tilde{\boldsymbol{V}}_2$ replaced by $\tilde{\boldsymbol{V}}_2^{(\rho)}$.

Finally we prove Lemma B.7 for the RAPS $\boldsymbol{\psi}^{(\rho)}$. Notice that, for $(\beta, p\tau^2) \in \mathcal{B}$, $t_j(\beta,\tau^2) = O_p(\sqrt{n}|\gamma_j| + 1)$ and $u_j(\beta,\tau^2) = O_p(\sqrt{n}|\gamma_j| + 1)$. These rates also hold for the partial derivatives of t_j and u_j with respect to β. Therefore, by the boundedness of ρ', ρ'' and ρ''',

$$\frac{\partial^2}{\partial\beta^2}\psi_{1j}^{(\rho)}(\beta,\tau^2) = \rho'''(t_j)u_j^3 - 3\rho''(t_j)u_j\frac{\partial}{\partial\beta}u_j + \rho'(t_j)\frac{\partial}{\partial\beta}u_j = O_p\big((\sqrt{n}|\gamma_j| + 1)^3\big).$$

Hence

$$\frac{\partial^2}{\partial\beta^2}\psi_1^{(\rho)}(\beta,\tau^2) = \sum_{j=1}^p \frac{\partial^2}{\partial\beta^2}\psi_{1j}^{(\rho)}(\beta,\tau^2) = O_p(p + \sqrt{n}\|\boldsymbol{\gamma}\|_1 + n\|\boldsymbol{\gamma}\|_2^2 + n^{3/2}\|\boldsymbol{\gamma}\|_3^3).$$

Using the assumption that $\|\boldsymbol{\gamma}\|_3^3 = O(1/\sqrt{p})$, it is easy to show that the right hand side is $O_p(n) = O_p\big((\tilde{\boldsymbol{V}}_2^{(\rho)})_{11}\big)$. Rates of the other partial derivatives can be proved analogously and we omit further detail.

Department of Statistics,
The Wharton School,
University of Pennsylvania,
USA
E-mail: qyzhao@wharton.upenn.edu; jingshuw@wharton.upenn.edu; dsmall@wharton.upenn.edu.
MRC Integrative Epidemiology Unit,
University of Bristol,
UK
E-mail: g.hemani@bristol.ac.uk; jack.bowden@bristol.ac.uk.