Instrumental Variables Estimation With Some Invalid Instruments and its Application to Mendelian Randomization

Hyunseung KANG, Anru ZHANG, T. Tony CAI, and Dylan S. SMALL

Instrumental variables have been widely used for estimating the causal effect between exposure and outcome. Conventional estimation methods require complete knowledge about all the instruments' validity; a valid instrument must not have a direct effect on the outcome and not be related to unmeasured confounders. Often, this is impractical as highlighted by Mendelian randomization studies where genetic markers are used as instruments and complete knowledge about instruments' validity is equivalent to complete knowledge about the involved genes' functions. In this article, we propose a method for estimation of causal effects when this complete knowledge is absent. It is shown that causal effects are identified and can be estimated as long as less than 50% of instruments are invalid, without knowing which of the instruments are invalid. We also introduce conditions for identification when the 50% threshold is violated. A fast penalized ℓ_1 estimation method, called sisVIVE, is introduced for estimating the causal effect without knowing which instruments are valid, with theoretical guarantees on its performance. The proposed method is demonstrated on simulated data and a real Mendelian randomization study concerning the effect of body mass index(BMI) on health-related quality of life (HRQL) index. An R package *sisVIVE* is available on CRAN. Supplementary materials for this article are available online.

KEY WORDS: Body mass index; Causal inference; Health-related quality of life; Instrumental variable; ℓ_1 penalization; Pleiotropy.

1. INTRODUCTION

Instrumental variables (IV) is a popular method for estimating the causal effect of an exposure on an outcome when there is unmeasured confounding. Conventional IV estimation methods require that the instruments are valid, or informally speaking, that the instruments are (A1) related to the exposure (A2) have no direct pathway to the outcome and (A3) are not related to unmeasured variables that affect the exposure and the outcome (see Figure 1 and Section 2 for a formal definition of valid IVs). For example, Figure 1 is an illustration of the IV assumptions and one potential violation of the IV assumptions (see, Hernán and Robins 2006 for details on other possible violations). Here, the IV is a genetic marker that is a single nucleotide polymorphism (SNP) whose value is fixed at birth and the unmeasured variables refer to variables that precede the assignment of the genetic marker, such as population stratification (to be discussed later). The challenge in IV estimation is to find valid instruments that satisfy assumptions (A1)–(A3). Unfortunately, this is a difficult task, especially in the case of Mendelian randomization (MR).

In MR, the goal is to estimate the causal effect of an exposure on an outcome by using genetic markers, specifically SNPs, as instruments (Davey Smith and Ebrahim 2003, 2004; Lawlor et al. 2008; Wehby, Ohsfeldt, and Murray 2008). For example, Timpson et al. (2005) studied the causal effect of C-reactive protein (CRP), the exposure, on various metabolic outcomes, such as body mass index (BMI) and cholesterol biomarkers (e.g., tryglycerides), using four haplotypes constructed from three SNPs (rs1800947, rs1130864, rs1205) as instruments. The instruments have been previously associated with plasma CRP levels, thereby agreeing with (A1). However, agreement with (A2) and (A3) is less certain. As the authors of the study noted, it is plausible that one or more of the genes that contain the SNPs, rs1800947, rs1130864, and rs1205, may have multiple functions, known as pleiotropy, where, in addition to changing CRP levels (the exposure), the gene containing one of these SNPs would change triglyceride levels or BMI (the outcome) and (A2) would not hold. Indeed, recent work by Martínez-Calleja et al. (2012) suggested that one of the instruments used, rs1130864, is directly linked to BMI, one of the outcomes, raising doubts about causal estimates when this SNP is assumed to be a valid instrument.

As another example, Katan (1986), in one of the first discussions of MR, proposed to estimate the causal effect of serum cholesterol level on cancer by using the apolipoprotein E polymorphism (APOE)'s effect on serum cholesterol levels. However, as Davey Smith and Ebrahim (2004) argued, the current knowledge about the APOE gene and its multiple pleiotropic effects on longevity, cholesterol biomarkers, and several other variables, would invalidate the APOE gene as a valid instrument, specifically due to its violation of (A2), and make an IV analysis based on it biased.

Both examples highlight a fundamental limitation with MR studies. For one, pleiotropy and its impact on (A2) is a concern in most MR studies (Little and Khoury 2003; Davey Smith and Ebrahim 2003, 2004; Thomas and Conti 2004; Brennan 2004; Lawlor et al. 2008). Lawlor et al. (2008) also list other biological phenomena associated with genetic instruments such

Hyunseung Kang is Ph.D. student (E-mail: *khyuns@wharton.upenn.edu*); Anru Zhang is Ph.D. student (E-mail: *anrzhang@wharton.upenn.edu*), T. Tony Cai is Dorothy Silberberg Professor of Statistics (E-mail: *tcai@wharton.upenn.edu*), and Dylan S. Small is Professor (E-mail: *dsmall@wharton.upenn.edu*), Department of Statistics, Jon M. Huntsman Hall, The Wharton School, University of Pennsylvania, 3730 Walnut Street, Philadelphia, PA 19104. The research of Tony Cai and Anru Zhang was supported in part by NSF FRG Grant DMS-0854973, NSF Grant DMS-1208982 and NIH Grant R01 CA127334-05. The research of Hyunseung Kang and Dylan Small was supported in part by NSF Grant SES-1260782. The authors thank Jack Bowden, two referees, and the associate editor for helpful suggestions.
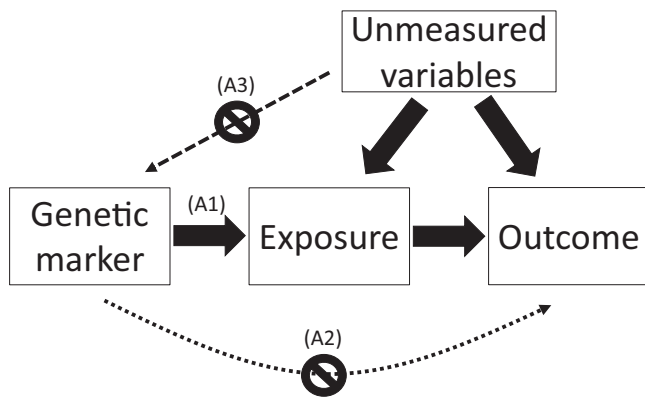
Figure 1. Diagram of one possible violation of instrumental variables assumptions. Arrows represent associations between variables. Absence of arrows indicates no relationship. Numbers (A1), (A2), and (A3) indicate different instrumental variables assumptions. In this example, the unmeasured variable refers to variables that precede the assignment of the genetic marker, such as population stratification, and the genetic marker is a single nucleotide polymorphism (SNP) whose value is fixed at birth. As such, the arrows from the unmeasured variable originate from the unmeasured variable and the arrow from the genetic marker goes from it to the outcome (A2) since the genetic marker is fixed at the time of conception.

as linkage disequilibrium and population stratification that may violate (A2) and (A3). Unfortunately, verifying genetic instruments as valid IVs requires having complete knowledge of the instruments' biological function and pleiotropic effects. As both examples highlight, the biological understanding of many genetic markers and their potential pleiotropic effects are typically incomplete at the time of the study (Solovieff et al. 2013). In the face of incomplete biological knowledge and possible instrument invalidity, can valid causal estimates be derived?

Previous work in IV estimation in the presence of possibly invalid instruments is limited. Traditional instrumental variables literature has stated that to estimate the causal effect of an exposure on an outcome when there are unmeasured confounders, one needs to have at least one instrument that one *knows* is valid (Wooldridge 2010). Andrews (1999) considered the invalid instrument case in the general context of generalized method of moments (GMM) estimation common in econometrics and arrived at an identification result that is similar to our identification result in Theorem 1. The author also proposed an estimation strategy, called the moment selection criteria (MSC), to correctly select the valid instruments, which is similar to (8) in Section 3.2. Unfortunately, as we discuss in Section 3.2, MSC is computationally infeasible when the number of instruments are large. Kolesár et al. (2011) considered the possibility of identifying causal effects when all the instruments are invalid because of direct effects on the outcome. The authors showed that if the direct effects are orthogonal to the instruments' effects on the treatment, then the causal effect can be identified. Kolesár et al. (2011) describes conditions under which this orthogonality is plausible. But, for MR, this stringent structure on the instruments would not hold in most cases as it would mean that the pleiotropic effects of the IVs are orthogonal to the effects of the IVs on the treatment. Gautier and Tsybakov (2011) analyzed instrumental variables regression in the presence of

possibly invalid instruments. However, for their procedure to work, one must have a predefined set of known valid instruments. Finally, Mealli and Pacini (2013) explored how using an auxiliary outcome can tighten bounds or provide identification of the effect of a treatment on a primary outcome when there is only one binary instrument that may violate (A2) by using an using auxiliary outcome. However, their work is different to our problem where we consider multiple candidate instruments.

Our article adds to the prior literature as follows. First, we show that it is indeed possible to identify and estimate the causal effect without a known predefined set of valid instruments. In particular, under a weaker condition where the proportion of invalid instruments is strictly less than 50% of the total instruments, we show that identification and estimation is possible. For example, given four possible haplotypes/instruments in the previous example by Timpson et al. (2005), estimation of the causal effect of CRP on metabolic phenotypes is still possible if no more than one instrument is invalid, without knowing exactly which of the four is invalid. We also show conditions for identification when the 50% threshold may not hold.

Second, we develop a fast ℓ_1 estimation procedure to estimate the causal effect of the exposure on the outcome in the presence of possibly invalid instruments. The procedure has provable theoretical guarantees on estimation performance and is computationally as fast as ordinary least squares (OLS). The procedure is implemented and available on CRAN as an R package *sisVIVE*, which stands for some invalid some valid IV estimator.

Third, we conduct a simulation study that compares our method to two-stage least squares (TSLS), the most popular estimation procedure in IV estimation. We show that our procedure dominates TSLS when the instruments may be invalid. We also conduct a real MR study concerning the effect of BMI on health-related quality of life (HRQL) measure using our new method.

2. CAUSAL MODEL AND INSTRUMENTAL VARIABLES

2.1 Notation

To define valid instruments, the potential outcomes approach (Neyman 1923; Rubin 1974) for instruments laid out in Holland (1988) is used. For each individual $i \in \{1, \ldots, n\}$, let $Y_i^{(d,\mathbf{z})} \in \mathbb{R}$ be the potential outcome if the individual were to have exposure $d \in \mathbb{R}$ and instruments $\mathbf{z} \in \mathbb{R}^L$. Let $D_i^{(\mathbf{z})} \in \mathbb{R}$ be the potential exposure if the individual had instruments $\mathbf{z} \in \mathbb{R}^L$. For each individual, only one possible realization of $Y_i^{(d,\mathbf{z})}$ and $D_i^{(\mathbf{z})}$ is observed, denoted as Y_i and D_i, respectively, based on his observed instrument values $\mathbf{Z}_{i.} \in \mathbb{R}^L$ and exposure D_i. In total, n sets of outcome, exposure, and instruments, denoted as $(Y_i, D_i, \mathbf{Z}_{i.})$, are observed in an iid fashion.

We denote $\mathbf{Y} = (Y_1, \ldots, Y_n)$ to be an n-dimensional vector of observed outcomes, $\mathbf{D} = (D_1, \ldots, D_n)$ to be an n-dimensional vector of observed exposures, and $\mathbf{Z}$ to be a n by L matrix of instruments where row i consists of $\mathbf{Z}_{i.}$.

For any vector $\boldsymbol{\alpha} \in \mathbb{R}^L$, let α_j denote the jth element of $\boldsymbol{\alpha}$. Let $\|\boldsymbol{\alpha}\|_1$, $\|\boldsymbol{\alpha}\|_2$, and $\|\boldsymbol{\alpha}\|_\infty$ be the usual 1, 2 and ∞-norms, respectively. Let $\|\boldsymbol{\alpha}\|_0$ denote the 0-norm, that is the number of nonzero elements in $\boldsymbol{\alpha}$. The support of $\boldsymbol{\alpha}$, denoted as $\text{supp}(\boldsymbol{\alpha}) \subseteq \{1, \ldots, L\}$, is defined as the set containing the nonzero elements

of the vector $\boldsymbol{\alpha}$, that is, $j \in \text{supp}(\boldsymbol{\alpha})$ if and only if $\alpha_j \neq 0$. A vector $\boldsymbol{\alpha}$ is called s-sparse if it has no more than s nonzero entries. Also, for a vector $\boldsymbol{\alpha} \in \mathbb{R}^L$ and set $A \subseteq \{1, \ldots, L\}$, we denote $\boldsymbol{\alpha}_A \in \mathbb{R}^L$ to be the vector where all the elements except whose indices are in A are zero.

For any n by L matrix $\mathbf{M} \in \mathbb{R}^{n \times L}$, we denote the (i, j) element of matrix $\mathbf{M}$ as M_{ij}, the ith row as $\mathbf{M}_{i.}$, and the jth column as $\mathbf{M}_{.j}$. Let $\mathbf{M}^T$ be the transpose of $\mathbf{M}$. Let $\mathbf{P_M}$ be the n by n orthogonal projection matrix onto the column space of $\mathbf{M}$, specifically $\mathbf{P_M} = \mathbf{M}(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T$; it is assumed that $\mathbf{M}^T\mathbf{M}$ has a proper inverse, unless otherwise noted. Let $\mathbf{P_{M^\perp}}$ be the residual projection matrix, specifically $\mathbf{P_{M^\perp}} = \mathbf{I} - \mathbf{P_M}$ where $\mathbf{I}$ is an n by n identity matrix.

For any sets $A \subseteq \{1, \ldots, L\}$, we denote A^C to be the complement of set A. Also, we denote $|A|$ to be the cardinality of set A.

2.2 Model

We consider the Additive LInear, Constant Effects (ALICE) model of Holland (1988) and extend it to allow for multiple valid and possibly invalid instruments as in Small (2007). Let $d', d \in \mathbb{R}$ be possible values of the exposure and $\mathbf{z}', \mathbf{z} \in \mathbb{R}^L$ be possible values of the instruments. Let $\epsilon_i = Y_i^{(0,\mathbf{0})} - E[Y_i^{(0,\mathbf{0})}|\mathbf{Z}_{i.}]$ and the collection of ϵ_i be denoted as $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$. Suppose we have the following potential outcomes model for the outcome

$$Y_i^{(d',\mathbf{z}')} - Y_i^{(d,\mathbf{z})} = (\mathbf{z}' - \mathbf{z})^T\boldsymbol{\phi}^* + (d' - d)\beta^* \qquad (1)$$

$$E(Y_i^{(0,\mathbf{0})}|\mathbf{Z}_{i.}) = \mathbf{Z}_{i.}^T\boldsymbol{\psi}^*, \qquad (2)$$

where $\boldsymbol{\phi}^*, \boldsymbol{\psi}^* \in \mathbb{R}^L$, and $\beta^* \in \mathbb{R}$ are unknown parameters. In Equation (1), the parameter β^* represents the causal parameter of interest, the causal effect on the outcome of changing the exposure by one unit. Also in Equation (1), the parameter $\boldsymbol{\phi}^*$ represents the direct effect of the instruments on the outcome; changing instruments from $\mathbf{z}'$ to $\mathbf{z}$ results in a direct effect on the outcome of $(\mathbf{z} - \mathbf{z})^T\boldsymbol{\phi}^*$. In Equation (2), the parameter $\boldsymbol{\psi}^*$ represents the confounders that affect the instrument and the outcome. In particular, without any confounders, there should not be any relationship between the instruments $\mathbf{Z}_{i.}$ and the potential outcome $Y_i^{(0,\mathbf{0})}$. Instead, in Equation (2), they are related via $\boldsymbol{\psi}^*$.

Let $\boldsymbol{\alpha}^* = \boldsymbol{\phi}^* + \boldsymbol{\psi}^*$. When we combine Equations (1) and (2) along with the definition of ϵ_i, we have the observed data model

$$Y_i = \mathbf{Z}_{i.}^T\boldsymbol{\alpha}^* + D_i\beta^* + \epsilon_i, \qquad E(\epsilon_i|\mathbf{Z}_{i.}) = 0. \qquad (3)$$

We make the following remarks regarding the model (3). First, the model can include exogenous measured covariates, say $\mathbf{X}_{i.} \in \mathbb{R}^p$ which may include the intercept term, and we can replace the variables Y_i, D_i, and $\mathbf{Z}_{i.}$ with the residuals after regressing them on $\mathbf{X}$ (e.g., replace $\mathbf{Y}$ by $(\mathbf{I} - \mathbf{P_X})\mathbf{Y}$) where $\mathbf{X}$ is the n by p matrix of covariates (Wang and Zivot 1998). The results in this article will hold generally when working with such data that is transformed by regressing out the effect of $\mathbf{X}$. In the same spirit, the model can be extended to nonlinear models by including appropriate basis transformations of $\mathbf{Z}_{i.}$. However, for simplicity of exposition, we will focus on a model without any measured covariates or nonlinear terms. We will also assume that $\mathbf{Y}$, $\mathbf{D}$, and the columns of $\mathbf{Z}$ are centered, which can also

result from a residual transformation with $\mathbf{X}$ containing only the intercept term.

Second, following Heckman and Robb (1985), Björklund and Moffitt (1987), and Small (2007), we can incorporate heterogeneous effects as follows. Suppose, instead of Equation (1), the potential outcomes model for the outcome is

$$Y_i^{(d',\mathbf{z}')} - Y_i^{(d,\mathbf{z})} = (\mathbf{z}' - \mathbf{z})^T\boldsymbol{\phi}^* + (d' - d)\beta_i^*, \qquad (4)$$

where $\beta^* = E(\beta_i^*)$ is the average effect of the exposure for everyone in the population. Then, the observed data model can be derived from (4) as follows

$$Y_i = \mathbf{Z}_{i.}^T\boldsymbol{\alpha}^* + D_i\beta^* + (\beta_i^* - \beta^*)D_i + \epsilon_i, \qquad E(\epsilon_i|\mathbf{Z}_{i.}) = 0. \qquad (5)$$

If $(\beta_i^* - \beta^*)$ is independent of D_i given $\mathbf{Z}_{i.}$, the heterogeneous model in (5) is identical to model (3) and our result for Theorem 1 in Section 3.1 hold. Also, as Small (2007) notes in page 1055, the assumption that $(\beta_i^* - \beta^*)$ is independent of D_i given $\mathbf{Z}_{i.}$ is equivalent to that "units do not select their treatment levels D_i given $\mathbf{Z}_{i.}$ based on the gains they would experience from treatment D_i given $\mathbf{Z}_{i.}$." If this assumption is violated, different groups of people will have different treatment effects, which in turn would lead to possibly nonzero $\boldsymbol{\alpha}^*$ (see, Angrist and Imbens 1995 and Small 2007 for details). For simplicity of exposition, we'll focus on a model with constant linear effect β^*.

2.3 Definition of Valid Instruments

Based on the observed model in (3), the parameter $\boldsymbol{\alpha}^*$ combines both the direct effect, represented by $\boldsymbol{\phi}^*$, and the effect of confounders on the $\mathbf{Z}_{i.}$ and $Y_i^{(0,0)}$ relationship, represented by $\boldsymbol{\psi}^*$. If there is no direct effect and no effect of the confounders, then $\boldsymbol{\alpha}^* = 0$. Hence, the value of $\boldsymbol{\alpha}^*$ captures the notion of valid and invalid instruments. The definition below formalizes this idea:

Definition 1. Suppose we have the models in (1)–(3) with L instruments. We say instrument $j \in \{1, \ldots, L\}$ is valid if $\alpha_j^* = 0$ and invalid if $\alpha_j^* \neq 0$.

Definition 1 distinguishes valid and invalid instruments based on $\text{supp}(\boldsymbol{\alpha}^*)$, the support of $\boldsymbol{\alpha}^*$. If instrument $j = 1, \ldots, L$ is not in the support, it is valid. If the instrument is in the support of $\boldsymbol{\alpha}^*$, it is invalid. Consequently, not knowing which instruments are valid and invalid directly translates to not knowing the support of $\boldsymbol{\alpha}^*$ in model (3).

In the case of only one instrument (i.e., $L = 1$), Definition 1 of a valid instrument matches with the informal definition (A2) and (A3) in the Introduction and the formal definition in Holland (1988). Specifically, the notion of exclusion restriction (A2), $Y_i^{(d,z)} = Y_i^{(d,z')}$ for all $z, z' \in \mathbb{R}$ is equivalent to the parameter ϕ^* in Equation (1) being zero. Also, the assumption of no unmeasured confounding of the IV-outcome relationship (A3) where $Y_i^{(d,z)}$ and $D_i^{(z)}$ are independent of Z_i for all $d, z \in \mathbb{R}$, is encoded by ψ^* in (2) being zero. Hence, $\phi^* = \psi^* = 0$, which implies $\alpha^* = 0$ and a valid IV in Holland (1988) is also a valid IV in our definition. Also, for one instrument, our model and definition is a special case of the definition of valid instrument discussed in Angrist, Imbens, and Rubin (1996) where

our model assumes an additive, linear, and constant treatment effect β^*.

For more than one instruments (i.e., $L > 1$), our model (1)–(3) and definition of valid IVs can be viewed as a generalization of Holland (1988). It is important to note that in this generalization, Definition 1 defines the validity of an instrument j in the context of the set of instruments $\{1, \ldots, L\}$ being considered. Specifically, an instrument j could be valid in the context of the set $\{1, \ldots, L\}$ (i.e., $\alpha_j^* = 0$), but invalid if considered alone because $\mathbf{Z}_{.j}$ may be associated with or causally affect another IV $\mathbf{Z}_{.j'}$, $j \neq j'$ where $\alpha_{j'}^* \neq 0$.

3. ESTIMATION OF CAUSAL EFFECT

3.1 Identifiability of Model

We first address whether the model in Equation (3) is identifiable, that is whether we can estimate the unknown parameters if we were given infinite data, even without any knowledge about which instruments are valid and invalid. We begin by making the assumptions.

(a) $E(\mathbf{Z}^T\mathbf{Z})$ is full rank.
(b) For $E(\mathbf{Z}^T\mathbf{D}) = E(\mathbf{Z}^T\mathbf{Z})\boldsymbol{\gamma}^*$, the components of $\boldsymbol{\gamma}^*$ are all not equal to zero, that is, $\gamma_j^* \neq 0$ for $j = 1, \ldots, L$.

Assumption (a) states that the matrix of instruments $\mathbf{Z}$ is full rank, a common assumption in the instrumental variables literature (Wooldridge 2010). Assumption (b) states that the instruments are associated with the exposure, akin to assumption (A1), that the instruments are relevant to the exposure; note that there does not need to be a causal relationship between the instrument $\mathbf{Z}$ and the exposure $\mathbf{D}$, just an association (Hernán and Robins 2006; Didelez and Sheehan 2007; Glymour, Tchetgen, and Robins 2012). As one reviewer remarked, assumption (b) requires that all L instruments are related to the exposure, $\gamma_j^* \neq 0$ for all j. If we have instruments that are not relevant to the exposure, $\gamma_j^* = 0$, we can exclude them from further analysis and concentrate only on those instruments that affect the exposure.

Now, the model in (3) implies the following moment condition

$$E(\mathbf{Z}^T(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha}^* - \mathbf{D}\beta^*)) = 0. \tag{6}$$

Suppose the assumptions (a) and (b) hold. Then, the moment equation in Equation (6) simplifies to

$$\boldsymbol{\Gamma}^* = \boldsymbol{\alpha}^* + \boldsymbol{\gamma}^*\beta^* \tag{7}$$

where $\boldsymbol{\Gamma}^* = E(\mathbf{Z}^T\mathbf{Z})^{-1}E(\mathbf{Z}^T\mathbf{Y})$. Since both $\boldsymbol{\Gamma}^*$ and $\boldsymbol{\gamma}^*$, defined by (b), can be identified by their moments based on observed data $E(\mathbf{Z}^T\mathbf{Z})^{-1}E(\mathbf{Z}^T\mathbf{Y})$ and $E(\mathbf{Z}^T\mathbf{Z})^{-1}E(\mathbf{Z}^T\mathbf{D})$, respectively, $\boldsymbol{\alpha}^*$ and β^* are identified if we can find a bijective mapping between $\boldsymbol{\alpha}^*$, β^*, and $\boldsymbol{\Gamma}^*$, $\boldsymbol{\gamma}^*$, that is, a unique solution $\boldsymbol{\alpha}^*$, β^* given $\boldsymbol{\Gamma}^*$, $\boldsymbol{\gamma}^*$.

If we know exactly which instruments are invalid $A^* = \text{supp}(\boldsymbol{\alpha}^*) = \{j : \alpha_j^* \neq 0\}$ and hence, know the set of valid instruments $(A^*)^C = \{j : \alpha_j^* = 0\}$, Equation (7) becomes

$$\boldsymbol{\alpha}_{(A^*)^C} + \boldsymbol{\gamma}_{(A^*)^C}^*\beta^* = \boldsymbol{\gamma}_{(A^*)^C}^*\beta^* = \Gamma_{(A^*)^C}^*.$$

There is a unique β^* so long as $|(A^*)^C| > 0$, or there is at least one known valid instrument. This is a special case of the classic

identification result for linear simultaneous equation models (Koopmans, Rubin, and Leipnik 1950).

If we know that there is a valid instrument, but are not sure of the identity of the valid instrument(s), then a unique solution to (7) and hence, identification, is not guaranteed. For example, let there be four instruments, $L = 4$ with $\boldsymbol{\gamma}^* = (1, 2, 3, 4)$ and $\boldsymbol{\Gamma}^* = (1, 2, 3, 8)$. Then, depending on the set of valid instruments $(A^*)^C$, which is unknown, we have two different β^* that satisfy Equation (7). If the set of valid instruments $(A^*)^C$ is $(A^*)^C = \{1, 2, 3\}$, we have $\boldsymbol{\gamma}_{(A^*)^C}^*\beta^* = \Gamma_{(A^*)^C}^*$ and $\beta^* = 1$. However, if the set of valid instruments is $(A^*)^C = \{4\}$, $\beta^* = 2$. Without knowing exactly which $(A^*)^C$ is the true set of valid instruments, we cannot choose between the two β^*'s and hence, there is not a unique solution to (7).

But, suppose we impose constraints on A^*. Specifically, suppose the number of invalid instruments, $s = |A^*|$, has to be less than some number U, $s < U$, without knowing which instruments are invalid or knowing exactly the number of invalid instruments. For example, geneticists may have a rough idea on the maximum number of invalid instruments, U, but not know exactly the number of invalid instruments nor do they know exactly which instruments are invalid. Note that this condition of knowing the maximum number of invalid instruments is a much weaker requirement than what is traditionally required in IV and MR literature where one must know exactly which instruments are invalid, that is, know exactly the set A^*; here, we only need an upper bound on the cardinality of A^*. Under the weaker condition $s < U$, a unique solution to (7) can exist and this is stated in Theorem 1.

Theorem 1 (Uniqueness of solution). Suppose we assume assumptions (a) and (b) and the modeling assumption (3). Let $s \in \{0, 1, \ldots, L\}$ with $s < U$ where $U = 1, \ldots, L$. Consider all sets $C_m \subseteq \{1, \ldots, L\}$, $m = 1, \ldots, M$ of size $|C_m| = L - U + 1$ with the property

$$\gamma_j^* q_m = \Gamma_j^* \quad j \in C_m$$

where q_m is a constant. There is a unique solution $\boldsymbol{\alpha}^*$ and β^* to (7) if and only if $q_m = q_{m'}$ for all $m, m' \in \{1, \ldots, M\}$.

To understand Theorem 1, note that if the valid instruments are those in the set C_m, then the causal effect $\beta^* = q_m$. Theorem 1 says that β^* is identified as long as there are not two subsets of the instruments of cardinality $L - U + 1$ that give internally consistent estimates of β^* (i.e., all instruments in each subset give the same estimate of β^*), but are externally inconsistent (i.e., the estimates of β^* from the two subsets are different). We call the property in Theorem 1 that there is a unique solution to $\boldsymbol{\alpha}^*$ and β^* to (7) if and only if $q_m = q_{m'}$ for all $m, m' \in \{1, \ldots, M\}$ the *consistency criterion*. We thank Jack Bowden for his insight and suggestions on terminology for interpreting Theorem 1.

As an example of applying Theorem 1, consider our numerical example above with $\boldsymbol{\gamma}^* = (1, 2, 3, 4)$ and $\boldsymbol{\Gamma}^* = (1, 2, 3, 8)$ and $U = 3$. Then, by Theorem 1 we have 3 sets $C_1 = \{1, 2\}$, $C_2 = \{1, 3\}$, $C_3 = \{2, 3\}$ with $q_1 = q_2 = q_3 = 1$. Hence, $\boldsymbol{\gamma}^*$ and $\boldsymbol{\Gamma}^*$ satisfy the consistency criterion of Theorem 1 and we have a unique solution $\boldsymbol{\alpha}^*$ and β^* to (7). In contrast, if $\boldsymbol{\gamma}^* = (1, 2, 3, 4)$ and $\boldsymbol{\Gamma}^* = (1, 2, 6, 8)$, we would have two sets $C_1 = \{1, 2\}$, $C_2 =$

$\{3, 4\}$ with $q_1 = 1$ and $q_2 = 2$, respectively. These $\boldsymbol{\gamma}^*$ and $\boldsymbol{\Gamma}^*$ do not satisfy the consistency criterion of Theorem 1 because $q_1 \neq q_2$ and there are no unique solutions $\boldsymbol{\alpha}^*$ and β^* to (7). Further discussion of this particular example is discussed in supplementary materials along with discussion of the implications of Theorem 1 when the additional linearity and normality assumptions of the classical linear simultaneous/structural equation model (Koopmans, Rubin, and Leipnik 1950) are considered.

Checking the consistency criterion can be computationally difficult, especially if U is large; it requires looking at $\binom{L}{L-U+1}$ possible subsets of $\{1, \ldots, L\}$ and the constants q_m associated with $\boldsymbol{\Gamma}^*$ and $\boldsymbol{\gamma}^*$. Corollary 1 says that the consistency criterion is automatically satisfied if $U \leq L/2$ (i.e., if 50% of the total candidate of L instruments are invalid) regardless of the values of $\boldsymbol{\gamma}^*$ and $\boldsymbol{\Gamma}^*$.

Corollary 1. If $U \leq L/2$, there is always a unique solution to (7)

In addition to the computational benefits, compared to Theorem 1, Corollary 1 is simpler to interpret. For example, for a geneticist, without knowing the entire biology of genetic instruments, specifically knowing which instruments are valid and invalid, as long as the number of invalid instruments is less than 50% of the total instruments, then the geneticist can rest assured that the parameters can always be identified. If this is not the case, the geneticist can always check the consistency criterion stated in Theorem 1.

We would like to mention two final points about Theorem 1. First, Theorem 1 is a statement about uniqueness of solutions for the parameters $\boldsymbol{\alpha}^*$, and β^* in Equation (7). A natural question to ask is whether the uniqueness is guaranteed for just β^*, the causal effect of interest, at the expense of nonuniqueness of $\boldsymbol{\alpha}^*$. In the proof of Theorem 1, we show that this cannot be the case. Specifically, regardless of the condition on s, the parameter β^* is a unique solution to (7) if and only if the parameter $\boldsymbol{\alpha}^*$ is a unique solution to (7). Second, Theorem 1 supposes the existences of the sets C_m and proceeds to compare their corresponding q_m. However, one may ask whether these sets C_m even exist in the first place. In the proof of Theorem 1, we provide a rigorous argument that, indeed, under model (3) and $s < U$, at least one set C_m has to exist.

3.2 Estimation of the Causal Effect of Exposure on Outcome

Given the model (3) and $s < U$, Theorem 1 lays out the sufficient and necessary condition for finding a unique solution to the moment Equation (6). Specifically, if the model is identified, the moment Equation (6) is zero at exactly one value, the true value of $\boldsymbol{\alpha}^*$ and β^*. Naturally then, a method to estimate the one true value is to find the values of $\boldsymbol{\alpha}^*$ and β^* that minimize (6) subject to the parameter constraint that $s < U$. Formally, we can write this estimation strategy as

$$\underset{\boldsymbol{\alpha}, \beta}{\text{argmin}} \frac{1}{2} \|\mathbf{P}_\mathbf{Z}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha} - \mathbf{D}\beta)\|_2^2, \quad \text{s.t.} \quad \|\boldsymbol{\alpha}\|_0 < U, \quad (8)$$

where $\|\boldsymbol{\alpha}\|_0$ is the number of nonzero entries of $\boldsymbol{\alpha}$ and by Definition 1, $s = \|\boldsymbol{\alpha}\|_0$. Equation (8) is similar to the moment selection

criterion (MSC) in Andrews (1999). However, both the moment selection criterion in Andrews (1999) and (8) are computationally infeasible in the sense that both require going through all subsets of size less than U and this type of problem has been shown to be NP-hard (Natarajan 1995). Instead, a computationally tractable version of estimation strategies like (8) has been proposed in the literature using a convex surrogate of the ℓ_0 norm (Candes and Tao 2005; Tropp 2006; Donoho 2006). Specifically, the computationally feasible version of the estimation strategy in (8) can be written as

$$\underset{\boldsymbol{\alpha}, \beta}{\text{argmin}} \frac{1}{2} \|\mathbf{P}_\mathbf{Z}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha} - \mathbf{D}\beta)\|_2^2, \quad \text{s.t.} \quad \|\boldsymbol{\alpha}\|_1 \leq t, \quad (9)$$

where the ℓ_0 norm is replaced by the convex norm ℓ_1 and U is replaced by a user-specified tuning parameter $t > 0$. In this article, we propose the equivalent Lagrangian form as our estimator of the causal effect, called *some invalid some valid IV estimator*, or sisVIVE, as follows

$$(\hat{\boldsymbol{\alpha}}_\lambda, \hat{\beta}_\lambda) \in \underset{\boldsymbol{\alpha}, \beta}{\text{argmin}} \frac{1}{2} \|\mathbf{P}_\mathbf{Z}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha} - \mathbf{D}\beta)\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1 \quad (10)$$

for some tuning parameter $\lambda > 0$ where λ corresponds to t in (9). If $\lambda = 0$ in (10), then (10) is the popular two stage least squares (TSLS) estimator, which is equivalent to the GMM estimator when the $\boldsymbol{\epsilon}$ are assumed to be homoscedastic (Hansen 1982). Hence, sisVIVE can be viewed as a generalization of TSLS or GMM.

sisVIVE also bears some resemblance to the traditional ℓ_1 penalization procedure, in particular the Lasso (Tibshirani 1996) or the recent ℓ_1 penalty procedures in IV estimation by Gautier and Tsybakov (2011) and Belloni et al. (2012). However, there are a few important differences. First, with regards to traditional Lasso and the procedure proposed by Gautier and Tsybakov (2011), our procedure in (10) only penalizes $\boldsymbol{\alpha}^*$. The estimator (10) does not penalize β^*, the causal effect of the exposure on the outcome, because the causal effect may be far from zero. In contrast, the prior works we mentioned penalize all the parameters in the model. Second, the traditional Lasso only considers regression with all exogenous regressors, which are regressors that are assumed to be independent of the error term or assumed to be fixed. The regressors in our model (3) are not all exogenous; specifically, model (3) contains one random endogenous variable, D_i, which is dependent on the error term. Third, Gautier and Tsybakov (2011) and Belloni et al. (2012) assume that either all the L instruments are valid or we know exactly which subset of them are valid. In contrast, our procedure does not assume this.

Finally, a careful reader may have recognized that there may be multiple minimizers to the Equation (10), specifically $\hat{\beta}_\lambda$, because $\|\boldsymbol{\alpha}\|_1$ is not strictly convex and hence, we use the set notation instead of the equality sign in (10). This might seem to be a concern as there are multiple estimates of β^*. However, as we will show in Section 3.4, all minimizers of (10) are close to the true values β^*. Also, if the entries of the matrix $\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{Z}$ where $\hat{\mathbf{D}} = \mathbf{P}_\mathbf{Z}\mathbf{D}$ (i.e., the predicted value of the exposure given the instruments) are drawn from a continuous distribution, then the solution to (10) is unique (Tibshirani 2013).

Without loss of generality, we assume that the columns of $\mathbf{Z}$ are scaled to unit length. This allows all L instruments to have

identical units so no columns of $\mathbf{Z}$ gets unfairly penalized by the penalty term in (10) simply due to their original units.

3.3 Choice of λ

Like many penalization procedures, the choice of the tuning parameter λ affects the performance of the estimation procedure and this is certainly the case with sisVIVE. High values of λ force heavy penalization on $\boldsymbol{\alpha}$, which will put most elements of $\hat{\boldsymbol{\alpha}}_\lambda$ to zero and most instruments will be estimated as valid instruments. In contrast, low values of λ will put few elements of $\hat{\boldsymbol{\alpha}}_\lambda$ to zero and most instruments will be estimated as invalid instruments. In short, the optimal choice of λ depends on knowing the exact number of invalid and valid instruments, something not implied by the condition $s < U$.

In practice, cross-validation is a popular data-driven method to choose λ. In the same spirit, we use a K-fold cross-validation where we minimize the estimating equation $||\mathbf{P}_{\mathbf{Z}}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha} - \mathbf{D}\beta)||_2$ instead of the predictive error $||(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha} - \mathbf{D}\beta)||_2$. We minimize the estimating equation instead of the predictive error since the parameter of interest is the causal effect β^* that sets the expected value of the estimating equation to zero (see Equation (6), Sections 3.1 and 3.2). We use the "one standard error" rule used in most cross-validation procedures (Hastie et al. 2009) and choose the smallest λ that is no more than one standard error above the minimum of the estimating equation. In Section 4, we discuss the performance of $\hat{\beta}_{\lambda_{cv}}$, where λ_{cv} is the cross-validated λ based on the estimating equation through various simulation studies. Also, in the supplementary materials, we discuss another method of choosing λ, in particular, choosing λ based on the theoretical guidance from Theorem 2 and Corollary 2. In short, the supplementary materials show that for better estimation performance of $\hat{\beta}_\lambda$, it is important not to incorrectly set invalid IVs to be valid (i.e., let $\hat{\alpha}_j$ to be zero when the true α_j^* is not zero), while the reverse is not as important. This observation argues for choosing λ that tends to set relatively few elements of $\hat{\boldsymbol{\alpha}}_\lambda$ to be zero and in the supplementary materials, we demonstrate that cross-validation achieves this goal in a wide variety of settings.

3.4 Estimation Performance

How well does sisVIVE estimate the causal effect β^*? To analyze the performance of sisVIVE, we first introduce some basic notations and definitions.

Definition 2. For any matrix $\mathbf{M}$, the upper and lower restricted isometry property (RIP) constants of order k, denoted as $\delta_k^+(\mathbf{M})$ and $\delta_k^-(\mathbf{M})$, respectively, are the smallest $\delta_k^+(\mathbf{M})$ and largest $\delta_k^-(\mathbf{M})$ such that

$$\delta_k^-(\mathbf{M})\|\boldsymbol{\alpha}\|_2^2 \leq \|\mathbf{M}\boldsymbol{\alpha}\|_2^2 \leq \delta_k^+(\mathbf{M})\|\boldsymbol{\alpha}\|_2^2 \tag{11}$$

holds for all k-sparse vectors $\boldsymbol{\alpha}$.

RIP conditions have been widely used in the literature on compressed sensing and high-dimensional linear regression. See Cai and Zhang (2013) and the references therein. The following theorem characterizes the performance of sisVIVE in finite samples using the RIP conditions. Note that this characterizes all the minimizers $\hat{\beta}_\lambda$ from sisVIVE in (10).

Theorem 2 (Estimation performance of sisVIVE). Suppose we have the model given in (3). Let $\hat{\mathbf{D}} = \mathbf{P}_{\mathbf{Z}}\mathbf{D}$. Let the restricted isometry constants $\delta_{2s}^-(\mathbf{Z})$, $\delta_{2s}^-(\mathbf{Z})$, $\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z})$ be defined as in (11), where s is the number of invalid instruments. Suppose

$$2\delta_{2s}^-(\mathbf{Z}) > \delta_{2s}^+(\mathbf{Z}) + 2\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z}) \tag{12}$$

holds, then the estimate $\hat{\beta}_\lambda$ given by (10) with tuning parameter $\lambda \geq 3\|\mathbf{Z}^T \mathbf{P}_{\hat{\mathbf{D}}^\perp}\boldsymbol{\epsilon}\|_\infty$ has the following performance guarantee

$$|\hat{\beta}_\lambda - \beta^*| \leq \frac{|\hat{\mathbf{D}}^T \boldsymbol{\epsilon}|}{\|\hat{\mathbf{D}}\|_2^2}$$
$$+ \frac{1}{\|\hat{\mathbf{D}}\|_2} \left(\frac{(4/3\sqrt{5})\lambda\sqrt{s\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z})}}{2\delta_{2s}^-(\mathbf{Z}) - \delta_{2s}^+(\mathbf{Z}) - 2\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z})} \right). \tag{13}$$

Condition (12) includes the RIP constants, $\delta_{2s}^-(\mathbf{Z})$, $\delta_{2s}^+(\mathbf{Z})$, and $\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z})$. Unfortunately, these RIP constants in (12) are difficult to evaluate. Hence, in some applications, it is more convenient to use a slightly stronger but much simpler and interpretable condition called the "mutual incoherence property" (MIP). Specifically, let $\hat{\mathbf{D}} = \mathbf{P}_{\mathbf{Z}}\mathbf{D}$ and $\|\mathbf{Z}_{.j}\|_2 = 1$ for all $j = 1, \ldots, L$. Define the constants μ and ρ as

$$\mu = \max_{i \neq j} |\mathbf{Z}_{.i}^T \mathbf{Z}_{.j}| \quad \text{and} \quad \rho = \max_j |\hat{\mathbf{D}}^T \mathbf{Z}_{.j}|/\|\hat{\mathbf{D}}\|_2. \tag{14}$$

First, the constant μ measures the maximum correlation between any two columns of the matrix of instruments $\mathbf{Z}$. This is related to Assumption (a) in Section 3.1 where a full rank $\mathbf{Z}$ means the columns of $\mathbf{Z}$ are linearly independent. In fact, if $\mu < 1/(L-1)$, $\mathbf{Z}$ is full rank. Second, the constant ρ measures the maximum strength of individual instruments. A high ρ doesn't necessarily imply that all L instruments are individually strong; it just implies that one of the L instruments is strong (i.e., has a high correlation to $\mathbf{D}$); it's possible that the rest of the $L - 1$ instruments are weak. This notion of strength by ρ is slightly different than the concentration parameter, which measures the overall strength of all the L instruments (see, Section 4 for details). Also, ρ stands in contrast to Condition (b) in Theorem 1 which looks at the individual values of γ_j, $j = 1, \ldots, L$, instead of the maximum of γ_j's.

Given the two MIP constants μ and ρ, we have the following result on estimation performance. Like Theorem 2, Corollary 2 characterizes all the minimizers $\hat{\beta}_\lambda$ from sisVIVE in (10).

Corollary 2 (Estimation performance of sisVIVE under MIP). Let the MIP constants μ and ρ be given in (14). If the number of invalid instruments, s, satisfies

$$s < \min\left(\frac{1}{12\mu}, \frac{1}{10\rho^2} \right) \tag{15}$$

the estimate $\hat{\beta}_\lambda$ given by (10) with tuning parameter $\lambda \geq 3\|\mathbf{Z}^T \mathbf{P}_{\hat{\mathbf{D}}^\perp}\boldsymbol{\epsilon}\|_\infty$ has the following performance guarantee

$$|\hat{\beta}_\lambda - \beta^*| \leq \frac{|\hat{\mathbf{D}}^T \boldsymbol{\epsilon}|}{\|\hat{\mathbf{D}}\|_2^2} + \frac{1}{\|\hat{\mathbf{D}}\|_2} \left(\frac{4\sqrt{105}/9\lambda s\rho}{1 - s(5\rho^2 + 6\mu)} \right). \tag{16}$$

We make the following remarks. First, in the supplementary materials, we show the condition in Equation (15) directly implies the condition in Equation (11). We also provide an example of a matrix of instruments $\mathbf{Z}$ where the RIP condition is satisfied, but the MIP condition is not satisfied. Second, the constraint on the number of invalid instruments, s, in Corollary 2 is strict, but is required to precisely characterize the bound on estimation performance. As two reviewers pointed out, if the instruments are even slightly correlated at $\mu = 0.1$, $s < 10/12$, no invalid instruments are allowed, and Corollary 2 is not useful in characterizing the performance of sisVIVE. In Section 4 and in the supplementary materials, we study the behavior of sisVIVE when this constraint in (15) may not hold. Third, in the case where all the instruments are uncorrelated with each other so that $\mu = 0$, a small ρ provides a less restrictive upper bound on s. At first glance, this may be counterintuitive since a small ρ implies that all the instruments' individual correlation to the exposure is weak and, therefore, having weak instruments allow one to have more invalid instruments. However, we note that the denominator of the bound (16), specifically $\|\hat{\mathbf{D}}\|_2^2$ is a function of the correlation of the instruments, and having a small ρ would translate to having a small $\|\hat{\mathbf{D}}\|_2^2$. Hence, even though the condition (15) allows for more invalid instruments, the upper bound (16) becomes worse and our estimator $\hat{\beta}_\lambda$ will be far from β^*. Finally, we emphasize that the conditions in both Theorem 2 and Corollary 2 are sufficient, but not necessary conditions for the performance bounds to hold. In particular, a violation of these conditions does not imply that sisVIVE will perform badly (see Section 4 and the supplementary materials).

3.5 Fast Numerical Algorithm

In addition to the theoretical guarantees on estimation performance, in practice, a fast, scalable numerical algorithm for estimation is desirable, especially for MR where genetic data can be large. Theorem 3 outlines a two-step numerical method whose solution is identical to sisVIVE in (10), but is as fast as OLS.

Theorem 3 (Fast two-step numerical algorithm). Let $\mathbf{P}_{\hat{\mathbf{D}}}$ be the projection matrix onto the vector $\hat{\mathbf{D}}$ and $\mathbf{P}_{\hat{\mathbf{D}}^\perp} = \mathbf{I} - \mathbf{P}_{\hat{\mathbf{D}}}$. We propose the two-step algorithm as follows.

Step 1. For a given $\lambda > 0$, solve

$$\hat{\boldsymbol{\alpha}}_\lambda \in \underset{\boldsymbol{\alpha}}{\mathrm{argmin}} \ \frac{1}{2}||\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{P}_{\mathbf{Z}}\mathbf{Y} - \mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{Z}\boldsymbol{\alpha}||_2^2 + \lambda||\boldsymbol{\alpha}||_1.$$

Step 2. Use $\hat{\boldsymbol{\alpha}}_\lambda$ from Step 1 to estimate $\hat{\beta}_\lambda$ by

$$\hat{\beta}_\lambda = \frac{\hat{\mathbf{D}}^T(\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\alpha}}_\lambda)}{||\hat{\mathbf{D}}||_2^2}.$$

The solution to the two-step algorithm is identical to the solution to sisVIVE in (10)

In the two-step algorithm, Step 1 is the standard Lasso problem with outcome $\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{P}_{\mathbf{Z}}\mathbf{Y}$ and $\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{Z}$; remember, sisVIVE in (10) is not the standard Lasso problem as discussed in Section 3.2. Fast algorithms for the Lasso exist, most notably LARS (Efron et al. 2004). In fact, LARS is able to solve $\hat{\boldsymbol{\alpha}}_\lambda$ for all values of $\lambda > 0$ at the same computational efficiency as OLS. Step

2 is also numerically efficient, requiring a simple dot product operation between $\hat{\mathbf{D}}$ and $\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\alpha}}_\lambda$. Thus, the proposed two-step algorithm is, practically speaking, as fast as OLS. Best of all, the estimate from this two-step algorithm is identical to sisVIVE.

4. SIMULATION STUDY

We conduct various simulation studies to study the estimation performance, measured by $|\hat{\beta} - \beta^*|$, for different methods. Specifically, we compare sisVIVE with TSLS, the most popular estimator in IV and MR, and OLS under various settings that vary the instruments' absolute/overall and relative strength, their validity and correlation among each other, and endogeneity.

Let there be $n = 2000$ individuals and $L = 10$ potential candidate instruments. The observations $(Y_i, D_i, \mathbf{Z}_{i.})$, $i = 1, \ldots, n$ are generated by

$$Y_i = \pi^* + \mathbf{Z}_{i.}^T \boldsymbol{\alpha}^* + D_i \beta^* + \epsilon_i,$$
$$D_i = \gamma_0^* + \mathbf{Z}_{i.}^T \boldsymbol{\gamma}^* + \xi_i,$$
$$\begin{pmatrix} \epsilon_i \\ \xi_i \end{pmatrix} \overset{\text{iid}}{\sim} N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{\epsilon\xi}^* \\ \sigma_{\epsilon\xi}^* & 1 \end{bmatrix} \right),$$

where $\mathbf{Z}_{i.}$ is drawn from a multivariate normal with mean $\mathbf{0}$ and covariance matrix where the diagonals are all one. Throughout the simulation, the parameters π^*, β^*, and γ_0^* are fixed. However, we vary (i) the endogeneity parameter $\sigma_{\epsilon\xi}^*$, (ii) the direct effect parameter $\boldsymbol{\alpha}^* = (1, 1, \ldots, 0, 0)$ where we change s in $\|\boldsymbol{\alpha}^*\|_0 = s$, (iii) the pairwise correlation between instruments, that is, μ in Equation (14), (iv) the absolute/overall strength of instruments, and (v) the relative strength of instruments, the latter two by changing the parameter $\boldsymbol{\gamma}^*$.

In particular, for (i), we vary $\sigma_{\epsilon\xi}^*$ from 0 to 0.9. For (ii), we vary s from 0 to 9. For (iii), we set μ at four different values, 0, 0.25, 0.5, and 0.75, by setting all the off-diagonal elements of the covariance matrix of $\mathbf{Z}_{i.}$ to this value. For (iv), we vary the absolute/overall instrument strength by the concentration parameter. The concentration parameter is a popular measure for instrument strength; high values of the concentration parameter indicate the overall strength of all L instruments are strong and vice versa. The concentration parameter is also the population value of the first stage F statistic for the instruments when the exposure is regressed on them; this first stage F statistic is often used to check instrument strength (Stock, Wright, and Yogo 2002). Based on Table 1 in Stock, Wright, and Yogo (2002), a set of instruments with a concentration parameter (scaled by the number of valid instruments) of around 10 is considered weak in the absolute/overall sense and instruments with a concentration parameter (scaled by the number of valid instruments) of around 100 is considered strong in the absolute/overall sense. Finally for (v), we vary the relative instrument strength by changing the individual entries of the vector $\boldsymbol{\gamma}^*$ while keeping the concentration parameter fixed. Specifically, for a particular concentration parameter, we consider instruments to have equal relative strength if $\gamma_j^* = \gamma_k^*$ for all $j \neq k$ and variable relative strength if $\gamma_j^* = 2 * \gamma_k^*$ for various values of $j \neq k$.

For each simulation setting, we repeat the simulation 1000 times. For each repetition, we compute sisVIVE's estimate of the causal effect, $\hat{\beta}_\lambda$, where λ is chosen by 10-fold cross-validation outlined in Section 3.3. We also compute es-

Table 1. Summary of instruments in the data analysis

Instruments	Major alleles	Heterozygote	Minor alleles	MAF (SE)
rs1421085	1281 (34.5%; TT)	1818 (49.0%; CT)	613 (16.5%; CC)	0.39 (0.0057)
rs1501299	1950 (52.5%; CC)	1502 (40.5%; AC)	260 (7.0%; AA)	0.24 (0.0049)
rs2241766	2956 (79.6%; TT)	719 (19.4%; TG)	37 (1.0%; GG)	0.10 (0.0036)
rs6265	2437 (65.7%; GG)	1112 (30.0%; AG)	163 (4.4%; AA)	0.19 (0.0046)

NOTE: MAF stands for minor allele frequency.

timates from TSLS and OLS. For TSLS, we run two types of TSLS. First, we run the "naive" TSLS as if all the instruments are valid. This is quite common in MR studies where all the instruments are assumed to be valid and the causal estimate is computed using TSLS. When some of the instruments are in fact invalid, naive TSLS should give biased estimates. Second, we run TSLS as if we knew exactly which instruments are valid, that is, the "oracle" TSLS. Specifically, we use the knowledge of the support of $\boldsymbol{\alpha}^*$ and run TSLS controlling for the invalid instruments that are in the support of $\boldsymbol{\alpha}^*$ as covariates. Finally, we run OLS with $\mathbf{Z}$ and $\mathbf{D}$ as our regressors and $\mathbf{Y}$ as our outcome. We expect OLS to perform poorly when there is substantial endogeneity by $\mathbf{D}$ since OLS cannot control for endogenous variables. But, OLS should be more efficient than IV methods if there is no endogeneity (Richardson and Wu 1971).

Figure 2 shows the estimation error when endogeneity is varied. The number of invalid instruments is fixed at $s = 3$ and we consider 16 different sets of instruments based on their absolute and relative strength as well as their pairwise correlations. For example, the top lefthand plot of Figure 2 corresponds to instruments whose overall strength is strong (i.e., scaled concentration parameter is around 100) , their relative strength is equal (i.e., γ_j^* are identical for all $j = 1, \ldots, L$), and their pairwise correlations are 0. In contrast, the bottom right plot of Figure 2 corresponds to instruments whose their overall strength is weak (i.e., scaled concentration parameter is around 10), their relative strength is variable (i.e., $\gamma_j^* = 2 * \gamma_k^*$ for $j \neq k$) and their pairwise correlations are equal to 0.75.

As expected, OLS dominates naive TSLS, oracle TSLS, and sisVIVE when the endogeneity is small and close to zero, with the dominance being greater for weak instruments. Once there is a sufficient amount of endogeneity, oracle TSLS, which knows exactly which instruments are valid and invalid, does best. However, sisVIVE, which is a feasible rather than infeasible oracle estimator, is close to the oracle TSLS; the gap between oracle TSLS and sisVIVE gets larger as the instruments' absolute strength gets weaker. Regardless of instrument strength, naive TSLS, which assumes all the L instruments are valid, has a high error since it cannot take into account the bias introduced by invalid instruments.

Figure 3 shows the estimation error when the number of invalid instruments is varied. The endogeneity, $\sigma_{\epsilon\xi}^*$, is fixed at 0.8. Like Figure 2, we consider the same 16 sets of instruments. We first see that at $s = 0$, that is, when there are no invalid instruments, sisVIVE's performance is nearly identical to naive and oracle TSLS. However, sisVIVE does not use the knowledge that one knows exactly which instruments are valid while the two TSLS estimators do. Also, sisVIVE's performance degrades slightly for instruments with weak absolute strength when the correlation between instruments increases.

When $s < L/2 = 5$, sisVIVE's performance is comparable to oracle TSLS and better than naive TSLS. However, for instruments with weak absolute strength, sisVIVE does slightly worse compared to the oracle TSLS than for instruments with strong absolute strength. Once we reach the identification boundary in Corollary 1, $s < L/2 = 5$, sisVIVE's performance becomes similar to naive TSLS. This is the case regardless of the instruments' absolute and relative strength. Finally, for any s, oracle TSLS performs much better than all the other estimators.

Also, in all 16 sets of instruments, we compute the ρ and μ found in the condition for Corollary 2 from the simulated data and this is detailed in supplementary materials. For example, the top lefthand plot of Figure 2 has ρ of approximately 0.31 and $\mu = 0$. Based on this, the upper bound on s in Corollary 2 is 1.04. However, since $s = 3$ for the simulations in Figure 2, the condition (15) in Corollary 2 is violated and cannot be used to characterize the behavior of sisVIVE. Regardless, in our simulation study presented in this section, sisVIVE performs just as well as the oracle TSLS.

In supplementary materials, we expand the simulation study to cover different types of instrument strength, correlation structure between instruments, and total number of potential instruments. We also explore different metrics of error, such as the proportion of correctly selected valid instruments and invalid instruments, to analyze the relationship between these proportion-based error metrics and the median bias error metric used in this Section. In addition, we also compute the conditions for Corollary 2, specifically ρ, μ, and λ required to achieve the performance bound. The supplementary materials show that in every case considered, sisVIVE performs no worse than the next best alternative, naive TSLS. In fact, in most cases, sisVIVE beats naive TSLS and performs similarly to the oracle TSLS. The only case where sisVIVE's performance deviated greatly from the oracle TSLS was when the invalid instruments were weaker than the valid instruments and $s = 4$. In addition, supplementary materials show that a good estimate of β^* depends strongly on correctly selecting the invalid instruments more than correctly selecting the valid instruments and choosing λ based on cross-validation seems to favor this situation. We also find that choosing λ based on Corollary 2 leads to a higher λ than one based on cross-validation. Finally, we find that sisVIVE based on λ chosen by cross-validation always performed at least as well as sisVIVE based on λ chosen by Corollary 2. In fact, in most cases, sisVIVE with a cross-validated λ performs better than sisVIVE with a λ chosen by Corollary 2.

Overall, sisVIVE using a cross-validated λ does much better than naive TSLS, the most frequently used estimator in MR
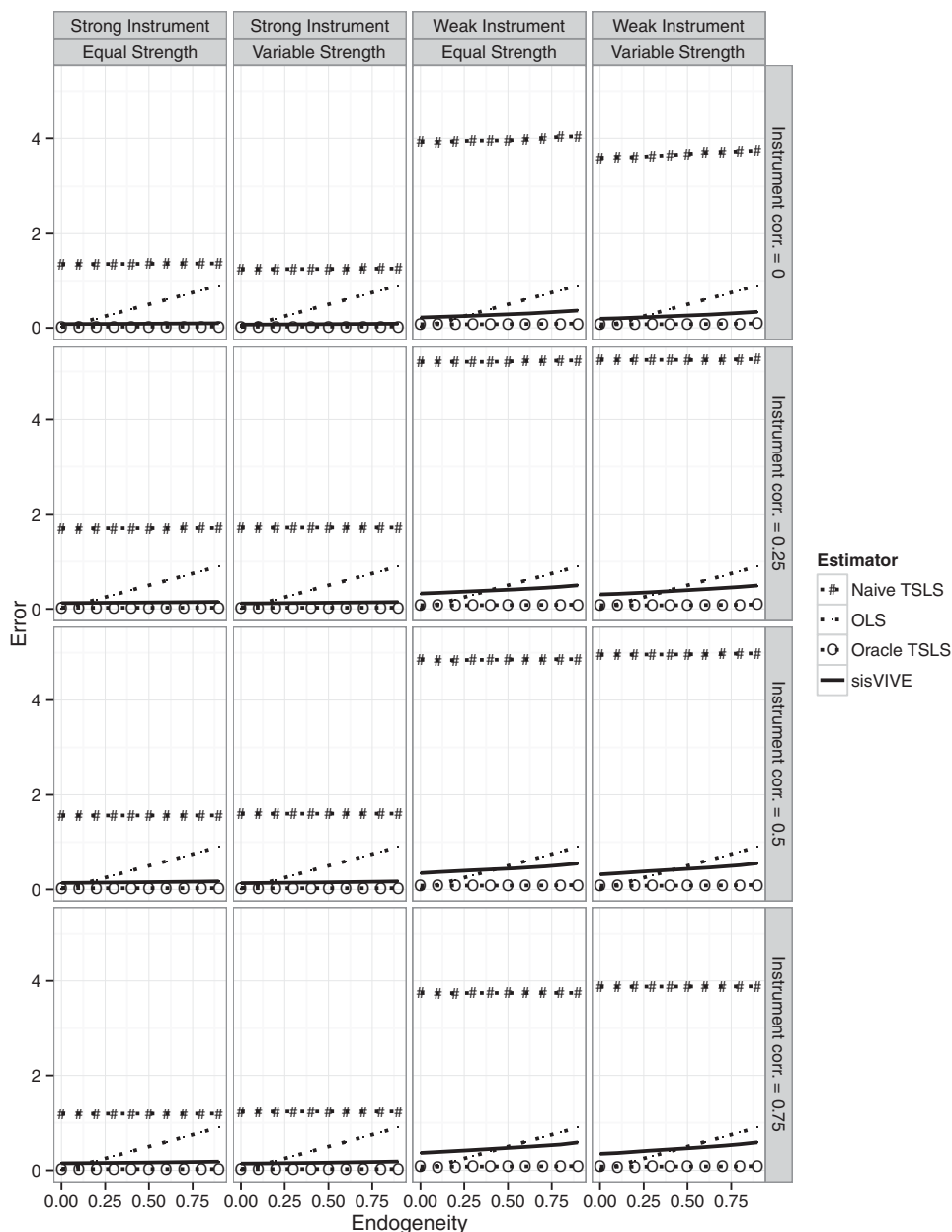
Figure 2. Simulation study of estimation performance varying endogeneity. There are 10 ($L = 10$) instruments. Each line represents median absolute estimation error ($|\beta^* - \hat{\beta}|$) after 1000 simulations. We fix the number of invalid instruments to $s = 3$. Each column in the plot corresponds to a different variation of instruments' absolute and relative strength. There are two types of absolute strengths, "Strong" and "Weak," measured by the concentration parameter. There are two types of relative strengths, "Equal" and "Variable," measured by varying γ^* while holding the absolute strength (i.e., concentration parameter) fixed. Each row corresponds to the maximum correlation between instruments.

and IV. In many cases, sisVIVE beats the naive TSLS and it is comparable to oracle TSLS. The promising simulation results suggest that sisVIVE should be used whenever there is concern about invalid instruments.

5. DATA ANALYSIS

We demonstrate the potential benefit of using sisVIVE in MR by analyzing the effect of obesity, the exposure, on HRQL, the outcome. An individual's quality of life is the general well-being of the individual; an individual's health quality of life is the subset of quality of life related to the individual's health (Torrance 1987). Previous non-MR studies by Trakas et al. (2001) and

Sach et al. (2007) have shown that there is a negative association between obesity and HRQL. However, a fundamental difficulty with these studies is that the outcome, HRQL, encompasses various factors about the individual, making it difficult to control for all possible confounders that may affect obesity and HRQL (Cawley and Meyerhoefer 2012). An MR approach offers the potential of controlling for unmeasured confounders.

For the analysis, we use the data from the Wisconsin Longitudinal Study (WLS), a well-known longitudinal study that has kept track of American high school graduates from Wisconsin since 1957. We look at graduates that were reinterviewed in 2003–2005 (Hauser 2005) and who have been genotyped. Similar to another analysis with the WLS genetic data, we remove
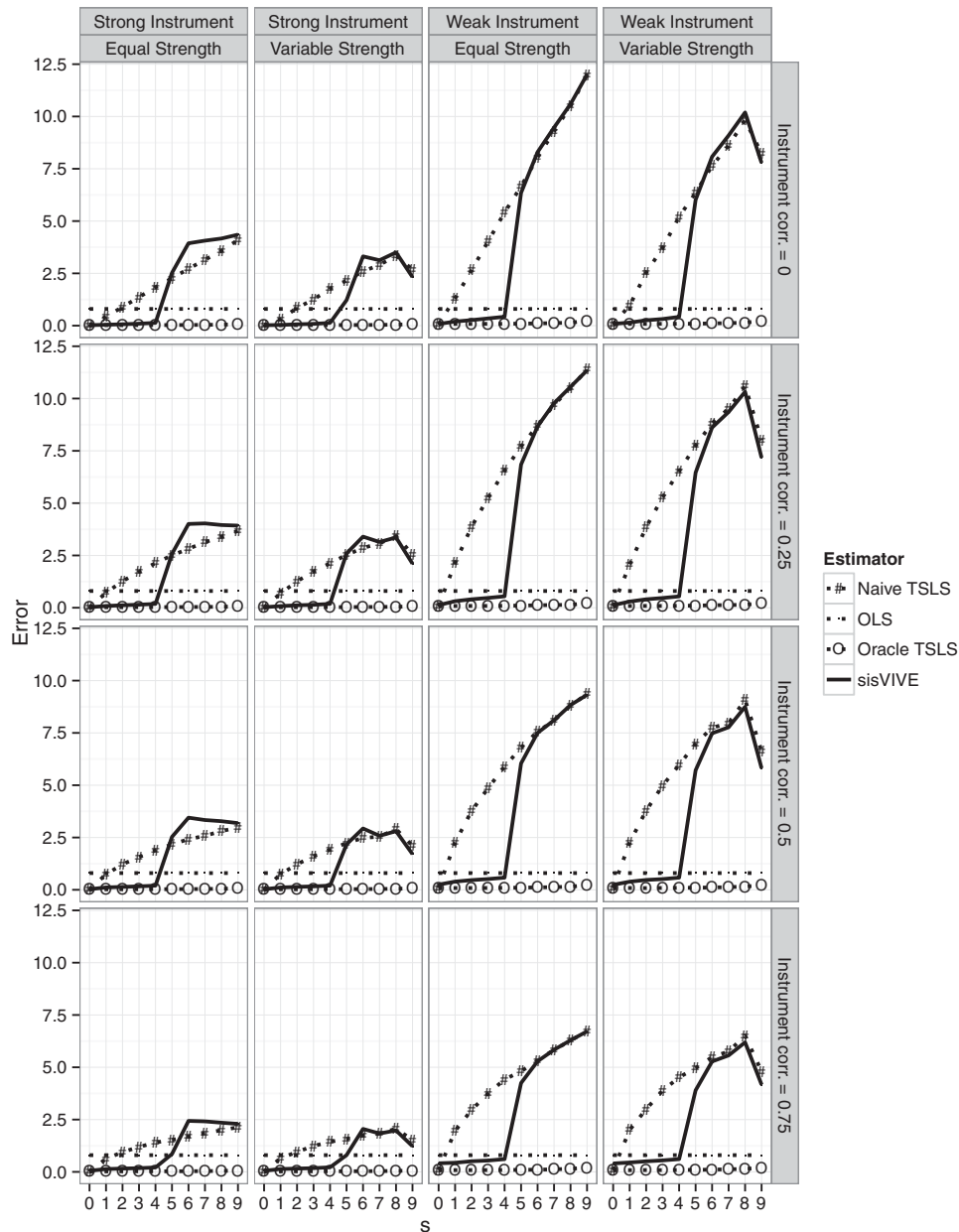
Figure 3. Simulation study of estimation performance varying the number of invalid instruments (s). There are 10 ($L = 10$) instruments. Each line represents median absolute estimation error ($|\beta^* - \hat{\beta}|$) after 1000 simulations. We fix the endogeneity $\sigma^*_{\epsilon\xi}$ to $\sigma^*_{\epsilon\xi} = 0.8$. Each column in the plot corresponds to a different variation of instruments' absolute and relative strength. There are two types of absolute strengths, "Strong" and "Weak," measured by the concentration parameter. There are two types of relative strengths, "Equal" and "Variable," measured by varying γ^* while holding the absolute strength fixed. Each row corresponds to maximum correlation between instruments.

individuals with more than 10% missing genotype data (Roetker et al. 2012). Our analysis of the dataset contains $n = 3712$ individuals with 1913 females and 1799 males born mostly between 1938 and 1940.

To measure HRQL, we use the Health Utility Index Mark 3 (HUI-3) which was also used in Trakas et al. (2001). HUI-3 is a composite score of utility between 0 and 1, with 1 indicating highest health state and 0 indicating a health state equivalent to death; negative utility is possible and indicates that the person is alive, but in a state worse than death. To measure obesity, we use the BMI and the U.S. National Institute of Health clinical guidelines (National Institute of Health 1998) that were also used in Trakas et al. (2001) and Sach et al. (2007) in their anal-

ysis. Specifically, we follow Trakas et al. (2001) and define the exposure by assigning individuals with BMI less than 30 (i.e., not obese) to be 0, individuals with BMI between 30 and 35 (i.e., obese class I) to be 1, individuals with BMI between 35 and 40 (i.e., obese class II) to be 2, and individuals with BMI greater than 40 (i.e., obese class III) to be 3 so that each value of the exposure corresponds to the increasing obese classes used in Trakas et al. (2001) and the U.S. National Institutes of Health clinical guidelines (National Institutes of Health 1998). For instance, exposure value of zero corresponds to nonobese individuals while exposure value of two corresponds to individuals in obese class II. Hence, the causal effect of interest is the effect of moving up in the obese class; specifically β^* in model

(1) will correspond to the effect of moving up one obese class on the HUI-3 index of HRQL. In supplementary materials, we explore different methods to quantify obesity and the resulting estimates from different methods.

For potential candidate instruments, we use the following SNPs in the WLS that have been previously shown to be associated with obesity: rs1421085, rs1501299, and rs2241766 (see, Table 1). rs1421085 is in the FTO gene and it has been shown to be strongly associated with obesity (Dina et al. 2007; Price, Li, and Zhao 2008). rs1501299 (i.e., +276G>T) is in the ADIPOQ gene that encodes adiponectin, a protein encoding for lipid metabolism, and has been associated with obesity (Bouatia-Naji et al. 2006; Yang et al. 2007). Finally, rs2241766 is also in the ADIPOQ gene that has been associated with obesity (Ukkola et al. 2003; Yang et al. 2003; Beckers et al. 2009). For all the SNPs, we follow an MR study done by Timpson et al. (2005) and assume an additive model. Although we have no particular reason to think any of the SNPs is an invalid IV, we are uncertain due to the lack of complete knowledge about the biological functions of the SNPs, a common scenario in MR studies. Our sisVIVE estimator will provide a good estimate as long as least two of the three SNPs are valid IVs.

A simple OLS analysis estimates that an increase in one obese class is associated with a 0.052 (SE: 0.0040) decrease in HUI-3 score. The reduced form estimates along with the first stage F statistics are summarized in the supplementary materials.

If we use TSLS, under the operating assumption that all the instruments are valid, the estimated causal effect is -0.00094 (SE: 0.081), that is, climbing up one obese class reduces your health utility quality of life by 0.00094. Our estimator, sisVIVE, which operates only under the assumption that a proportion of instruments are invalid, estimates -0.00094 as the causal effect, which is identical to the estimate by TSLS. Also, sisVIVE does not select any SNPs as an invalid IV. The overidentifying restrictions test and the implied structural correlation between D_i and the error term are summarized in the supplementary materials.

To further validate our method, we include another instrument, rs6265 (i.e., Val66Met). rs6265 is in the brain-derived neurotrophic factor BDNF gene and has been shown to not only be associated with BMI (Thorleifsson et al. 2008; Shugart et al. 2009), but also neurological and cognitive function (Hwang et al. 2006; Rybakowski et al. 2006). Hence, there is some reason to believe that rs6265 may be pleiotropic; rs6265 may impact obesity, but also affect HRQL through mechanisms other than obesity. sisVIVE should be able to pick up on this instrument being invalid in contrast to TSLS, which will always assume that all the instruments used are valid.

If we use TSLS under the operating assumption that all the four instruments are valid, the estimated effect is -0.0086 (SE:0.080). sisVIVE, on the other hand, estimates the causal effect to be -0.0037, which is closer to the estimates when we used three instruments. sisVIVE also throws out the instrument, rs6265, which we suspect to be invalid. The reduced form estimates and the overidentifying restrictions test are summarized in the supplementary materials.

In both data analyses, sisVIVE operates under the assumption of possibly invalid instruments, which are typical in MR studies, while TSLS operates under the assumption of all valid instruments. In the first data analysis where there was no reason to believe that the instruments were invalid, sisVIVE provides the same answer as TSLS, but without assuming that all the instruments were valid. In the second data analysis where one instrument was suspect, sisVIVE removed the suspected instrument. In both cases, sisVIVE was robust to possibly invalid instruments compared to TSLS.

6. DISCUSSION

This article demonstrates that proper estimation of causal effects using the IV method is possible without knowledge of all the instruments' validity. Our results show that simply knowing a proportion of the instrument is valid, without knowing which are valid, is sufficient and we construct the sisVIVE estimator that dominates the naive TSLS in almost every aspect while performing similarly to the oracle TSLS. Both the simulation result and data analysis show that sisVIVE is a robust alternative to TSLS in the presence of possibly invalid instruments.

Future work could involve generalizing the model considered. In particular, the current article discusses a model in which treatment effects are constant. Angrist, Imbens, and Rubin (1996) discusses the setting in which the treatment effects are not constant and individuals may select into treatment based on expected gains from treatment. Then, q_m and $q_{m'}$ in Theorem 1 might not be equal to each other for different sets of valid instruments and Theorem 1 does not apply. It would be useful to understand what sisVIVE is estimating under this setting of treatment effect heterogeneity. Other useful directions for future work are relaxing the conditions on Corollary 2 to encompass more invalid instruments s and deriving tests for identification. Also, we have focused on the applications of our method to Mendelian randomization. In economic applications, it is also common to have multiple candidate instruments and be concerned that some proportion of the instruments are invalid (Murray 2006). Our current work demonstrates that instrumental variable estimation is definitely possible even in the presence of possibly invalid instruments.

7. SUPPLEMENTARY MATERIALS

Proofs: The file contains additional information about the Wisconsin Longitudinal Data, the simulation study, further discussions, and all the technical details, including the proofs of Theorems 1, 2, and 3.

Supplementary zip file: The file contains the R-package "sisVIVE" which implements the method proposed in the article along with documentation and one technical report on arXiv and one article on NBER Working Article series, both of which are cited in the main manuscript.

[Received November 2013. Revised November 2014.]

REFERENCES

Andrews, W. K. D. (1999), "Consistent Moment Selection Procedures for Generalized Method of Moments Estimation," *Econometrica*, 67, 543–563. [133,136]

Angrist, J. D., and Imbens, G. W. (1995), "Two-Stage Least Squares Estimation of Average Causal Effects in Models With Variable Treatment Intensity," *Journal of the American Statistical Association*, 90, 431–442. [134]

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455. [134,142]

Beckers, S., Peeters, A. V., de Freitas, F., Mertens, I. L., Verhulst, S. L., Haentjens, D., Desager, K. N., Van Gaal, L. F., and Van Hul, W. (2009), "Association Study and Mutation Analysis of Adiponectin Shows Association of Variants in APM1 With Complex Obesity in Women," *Annals of Human Genetic*, 73, 492–501. [142]

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012), "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain," *Econometrica*, 80, 2369–2429. [136]

Björklund, A., and Moffitt, R. (1987), "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models," *The Review of Economics and Statistics*, 69, 42–49. [134]

Bouatia-Naji, N., Meyre, D., Lobbens, S., Séron, K., Fumeron, F., Balkau, B., Heude, B., Jouret, B., Scherer, P. E., Dina, C., Weill, J., and Froquel, P. (2006), "ACDC/Adiponectin Polymorphisms are Associated with Severe Childhood and Adult Obesity," *Diabetes*, 55, 545–550. [142]

Brennan, P. (2004), "Commentary: Mendelian Randomization and Gene–Environment Interaction," *International Journal of Epidemiology*, 33, 17–21. [132]

Cai, T. T., and Zhang, A. (2013), "Compressed Sensing and Affine Rank Minimization Under Restricted Isometry," *IEEE Transactions on Signal Processing*, 61, 3279–3290. [137]

Candes, E. J., and Tao, T. (2005), "Decoding by Linear Programming," *IEEE Transactions on Information Theory*, 51, 4203–4215. [136]

Cawley, J., and Meyerhoefer, C. (2012), "The Medical Care Costs of Obesity: An Instrumental Variables Approach," *Journal of Health Economics*, 31, 219–230. [140]

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. [132]

——— (2004), "Mendelian Randomization: Prospects, Potentials, and Limitations," *International Journal of Epidemiology*, 33, 30–42. [132]

Dina, C., Meyre, D., Gallina, S., Durand, E., Kórner, A., Jacobson, P., Carlsson, L. M. S., Kiess, W., Vatin, V., Lecoeur, C., Delplanque, J., Vaillant, C., Pattou, F., Ruiz, J., Weill, J., Levy-Marchal, C., Horber, F., Potoczna, N., Hercberg, S., Le Stunff, C., Bougneres, P., Kovacs, P., Marre, M., Balkau, B., Cauchi, S., Chevre, J. C., and Froguel, P. (2007), "Variation in FTO Contributes to Childhood Obesity and Severe Adult Obesity," *Nature Genetics*, 39, 724–726. [142]

Didelez, V., and Sheehan, N. (2007), "Mendelian Randomization as an Instrumental Variable Approach to Causal Inference," *Statistical Methods in Medical Research*, 16, 309–330. [135]

Donoho, D. L. (2006), "For Most Large Underdetermined Systems of Linear Equations the Minimal L_1-norm Solution is Also the Sparsest Solution," *Communications on Pure and Applied Mathematics*, 59, 797–829. [136]

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499. [138]

Gautier, E., and Tsybakov, A. B. (2011), "High-Dimensional Instrumental Variables Regression and Confidence Sets," arXiv:1105.2454 [math.ST]. [133,136]

Glymour, M. M., Tchetgen, E. T., and Robins, J. M. (2012), "Credible Mendelian Randomization Studies: Approaches for Evaluating the Instrumental Variable Assumptions," *American Journal of Epidemiology*, 175, 332–339. [135]

Hansen, L. P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 1029–1054. [136]

Hastie, T., Tibshirani, R., and Friedman, H. (2009), *The Elements of Statistical Learning* (2nd ed.), New York: Springer. [137]

Hauser, R. M. (2005), "Survey Response in the Long Run: The Wisconsin Longitudinal Study," *Field Methods*, 17, 3–29. [140]

Hernán, M. A., and Robins, J. M. (2006), "Instruments for Causal Inference: An Epidemiologist's Dream," *Epidemiology*, 17, 360–372. [132,135]

Heckman, J. J., and Robb, R. (1985), "Alternative Methods for Evaluating the Impact of Interventions: An Overview," *Journal of Econometrics*, 30, 239–267. [134]

Holland, P. W. (1988), "Causal Inference, Path Analysis, and Recursive Structural Equations Models," *Sociological Methodology*, 18, 449–484. [133,134,135]

Hwang, J.-P., Tsai, S.-J., Hong, C.-J., Yang, C.-H., Lirng, J.-F., and Yang, Y.-M. (2006), "The Val66Met Polymorphism of the Brain-Derived Neurotrophic-factor Gene is Associated With Geriatric Depression," *Neurobiology of Aging*, 27, 1834–1837. [142]

Katan, M. B. (1986), "Apolipoprotein E Isoforms, Serum Cholesterol, and Cancer," *Lancet*, 327, 507–508. [132]

Kolesár, M., Chetty, R., Friedman, J. N., Glaeser, E. L., and Imbens, G. W. (2011), "Identification and Inference With Many Invalid Instruments," NBER Working Paper No. 17519, National Bureau of Economic Research. [133]

Koopmans, T. C., Rubin, H., and Leipnik, R. B. (1950), "Measuring the Equation Systems of Dynamic Economics," in *Statistical Inference in Dynamic Economic Models*, New York: Wiley, pp. 54–237. [135]

Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N., and Smith, G. D. (2008), "Mendelian Randomization: Using Genes as Instruments for Making Causal Inferences in Epidemiology," *Statistics in Medicine*, 27, 1133–1163. [132]

Little, J., and Khoury, M. J. (2003), "Mendelian Randomisation: A New Spin or Real Progress?" *The Lancet*, 362, 930–931. [132]

Martínez-Calleja, A., Quiróz-Vargas, I., Parra-Rojas, I., Muñoz-Valle, J. F., Leyva-Vázquez, M. A., Fernández-Tilapa, G., Vences-Velázquez, A., Cruz, M., Salazar-Martínez, E., and Flores-Alfaro, E. (2012), "Haplotypes in the CRP Gene Associated With Increased BMI and Levels of CRP in Subjects With Type 2 Diabetes or Obesity From Southwestern Mexico," *Experimental Diabetes Research*, 2012, 1–7, Article ID 982683. [132]

Mealli, F., and Pacini, B. (2013), "Using Secondary Outcomes to Sharpen Inference in Randomized Experiments With Noncompliance," *The Journal of the American Statistical Association*, 108, 1120–1131. [133]

Murray, M. P. (2006), "Avoiding Invalid Instruments and Coping with Weak Instruments," *The Journal of Economic Perspectives*, 20, 111–132. [142]

Natarajan, B. K. (1995), "Sparse Approximate Solutions to Linear Systems," *SIAM Journal on Computing*, 24, 227–234. [136]

National Institutes of Health (1998), "Clinical Guidelines on the Identification, Evaluation, and Treatment of Overweight and Obesity in Adults: The Evidence Report," *Obesity Research*, 2, 51S–209S. [141]

Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments," *Statistical Science*, 5, 463–480. [133]

Price, R. A., Li, W.-D., and Zhao, H. (2008), "FTO Gene SNPs Associated With Extreme Obesity in Cases, Controls and Extremely Discordant Sister Pairs," *BMC Medical Genetics*, 9, 1–5. [142]

Richardson, D. H., and Wu, D. (1971), "A Note on the Comparison of Ordinary and Two-Stage Least Squares Estimators," *Econometrica*, 39, 973–981. [139]

Roetker, N. S., Yonker, J. A., Lee, C., Chang, V., Basson, J. J., Roan, C. L., Hauser, T. S., Hauser, R. M., and Atwood, C. S. (2012), "Multigene Interactions and Prediction of Depression in the Wisconsin Longitudinal Study," *British Medical Journal Open*, 2, e000944, 1–7. [141]

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701. [133]

Rybakowski, J. K., Borkowska, A., Skibinska, M., Szczepankiewicz, A., Kapelski, P., Leszczynska-rodziewicz, A., Czerski, P. M., and Hauser, J. (2006), "Prefrontal Cognition in Schizophrenia and Bipolar Illness in Relation to Val66Met Polymorphism of the Brain-Derived Neurotrophic Factor Gene," *Psychiatry and Clinical Neurosciences*, 60, 70–76. [142]

Sach, T. H., Barton, G. R., Doherty, M., Muir, K. R., Jenkinson, C., and Avery, A. J. (2007), "The Relationship Between Body Mass Index and Health-Related Quality of Life: Comparing the EQ-5D, EuroQol VAS and SF-6D," *International Journal of Obesity*, 31, 189–196. [140,141]

Shugart, Y. Y., Chen, L., Day, I. N., Lewis, S. J., Timpson, N. J., Yuan, W., Abdollahi, M. R., Ring, S. M., Ebrahim, S., Golding, J., Lawlor, D. A., and Smith, G. D. (2009), "Two British Women Studies Replicated the Association Between the Val66Met Polymorphism in the Brain-Derived Neurotrophic Factor (BDNF) and BMI," *European Journal of Human Genetics*, 17, 1050–1055. [142]

Small, D. S. (2007), "Sensitivity Analysis for Instrumental Variables Regression with Overidentifying Restrictions," *Journal of the American Statistical Association*, 102, 1049–1058. [134]

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. [133]

Stock, J. H., Wright, J. H., and Yogo, M. (2002), "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business and Economic Statistics*, 20, 518–520. [138]

Thomas, D. C., and Conti, D. V. (2004), "Commentary: The Concept of 'Mendelian Randomization'," *International Journal of Epidemiology*, 33, 21–25. [132]

Thorleifsson, G., Walters, G. B., Gudbjartsson, D. F., Steinthorsdottir, V., Sulem, P., Helgadottir, A., Styrkarsdottir, U., Gretarsdottir, S., Thorlacius, S., Jonsdottir, I., Jonsdottir, T., Olafsdottir, E. J., Olafsdottir, G. H., Jonsson, T., Jonsson, F., Borch-Johnsen, K., Hansen, T., Andersen, G., Jorgensen, T., Lauritzen, T., Aben, K. K., Verbeek, A. L., Roeleveld, N., Kampman, E., Yanek, L. R., Becker, L. C., Tryggvadottir, L., Rafnar, T., Becker, D. M., Gulcher, J., Kiemeney, L. A., Pedersen, O., Kong, A., Thorsteinsdottir, U.,

and Stefansson, K. (2008), "Genome-Wide Association Yields New Sequence Variants at Seven Loci that Associate With Measures of Obesity," *Nature Genetics*, 41, 18–24. [142]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society,* Series B, 58, 267–288. [136]

——— (2013), "The Lasso Problem and Uniqueness," *Electronic Journal of Statistics*, 7, 1456–1490. [136]

Timpson, N. J., Lawlor, D. A., Harbord, R. M., Gaunt, T. R., Day, I. N., Palmer, L. J., Hattersley, A. T., Ebrahim, S., Lowe, G., Rumley, A., and Smith, G. D. (2005), "C-reactive Protein and its Role in Metabolic Syndrome: Mendelian Randomisation Study," *The Lancet*, 366, 1954–1959. [132,133,142]

Torrance, G. W. (1987), "Utility Approach to Measuring Health-Related Quality of Life," *Journal of Chronic Disease*, 40, 593–600. [140]

Trakas, K., Oh, P. I., Singh, S., Risebrough, N., and Shear, N. H. (2001), "The Health Status of Obese Individuals in Canada," *International Journal of Obesity and Related Metabolic Disorders: Journal of the International Association for the Study of Obesity*, 25, 662–668. [140,141]

Tropp, J. A. (2006), "Just Relax: Convex Programming Methods for Identifying Sparse Signals in Noise," *IEEE Transactions on Information Theory*, 52, 1030–1051. [136]

Ukkola, O., Ravussin, E., Jacobson, P., Sjöström, L., and Bouchard, C. (2003), "Mutations in the Adiponectin Gene in Lean and Obese Subjects From the Swedish Obese Subjects Cohort," *Metabolism*, 52, 881–884. [142]

Wang, J., and Zivot, E. (1998), "Inference on Structural Parameters in Instrumental Variables Regression With Weak Instruments," *Econometrica*, 66, 1389–1404. [134]

Wehby, G. L., Ohsfeldt, R. L., and Murray, J. C. (2008), "'Mendelian Randomization' Equals Instrumental Variable Analysis With Genetic Instruments," *Statistics in Medicine*, 27, 2745–2749. [132]

Wooldridge, J. M. (2010), *Econometric Analysis of Cross Section and Panel Data* (2nd ed.), Cambridge, MA: MIT Press. [133,135]

Yang, W. S., Tsou, P. L., Lee, W. J., Tseng, D. L., Chen, C. L., Peng, C. C., Lee, K. C., Chen, M. J., Huang, C. J., Tai, T. Y., and Chuang, L. M. (2003), "Allele-Specific Differential Expression of a Common Adiponectin Gene Polymorphism Related to Obesity," *Journal of Molecular Medicine*, 81, 428–434. [142]

Yang, W. S., Yang, Y. C., Chen, C. L., Wu, I. L., Lu, J. Y., Lu, F. H., Tai, T. Y., and Chang, C.-J. (2007), "Adiponectin SNP276 is Associated with Obesity, the Metabolic Syndrome, and Diabetes in the Elderly," *The American Journal of Clinical Nutrition*, 86, 509–513. [142]