1-1-2015

Instrumental Variables and Mendelian Randomization With Invalid Instruments

Hyunseung Kang
University of Pennsylvania, khyuns@wharton.upenn.edu

Instrumental Variables and Mendelian Randomization With Invalid Instruments

Abstract

Instrumental variables (IV) methods have been widely used to determine the causal effect of a treatment, exposure, policy, or an intervention on an outcome of interest. The IV method relies on having a valid instrument, a variable that is (A1) associated with the exposure, (A2) has no direct effect on the outcome, and (A3) is unrelated to the unmeasured confounders associated with the exposure and the outcome. However, in practice, finding a valid instrument, especially those that satisfy (A2) and (A3), can be challenging. For example, in Mendelian randomization studies where genetic markers are used as instruments, complete knowledge about instruments' validity is equivalent to complete knowledge about the involved genes' functions.

The dissertation explores the theory, methods, and application of IV methods when invalid instruments are present. First, when we have multiple candidate instruments, we establish a theoretical bound whereby causal effects are only identified as long as less than 50% of instruments are invalid, without knowing which of the instruments are invalid. We also propose a fast penalized method, called sisVIVE, to estimate the causal effect. We find that sisVIVE outperforms traditional IV methods when invalid instruments are present both in simulation studies as well as in real data analysis.

Second, we propose a robust confidence interval under the multiple invalid IV setting. This work is an extension of our work on sisVIVE. However, unlike sisVIVE which is robust to violations of (A2) and (A3), our confidence interval procedure provides honest coverage even if all three assumptions, (A1)-(A3), are violated.

Third, we study the single IV setting where the one IV we have may actually be invalid. We propose a nonparametric IV estimation method based on full matching, a technique popular in causal inference for observational data, that leverages observed covariates to make the instrument more valid. We propose an estimator along with inferential results that are robust to mis-specifications of the covariate-outcome model. We also provide a sensitivity analysis should the instrument turn out to be invalid, specifically violate (A3).

Fourth, in application work, we study the causal effect of malaria on stunting among children in Ghana. Previous studies on the effect of malaria and stunting were observational and contained various unobserved confounders, most notably nutritional deficiencies. To infer causality, we use the sickle cell genotype, a trait that confers some protection against malaria and was randomly assigned at birth, as an IV and apply our nonparametric IV method. We find that the risk of stunting increases by 0.22 (95% CI: 0.044,1) for every malaria episode and is sensitive to unmeasured confounders.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group
Statistics

First Advisor
Dylan S. Small

Second Advisor
Tony Cai

Keywords
Causal Inference, Econometrics, Instrumental Variables, Invalid Instruments, Mendelian Randomization

Subject Categories
Biostatistics | Economics | Statistics and Probability

INSTRUMENTAL VARIABLES AND MENDELIAN RANDOMIZATION WITH INVALID INSTRUMENTS

Hyunseung Kang

A DISSERTATION

in

Statistics

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2015

Supervisor of Dissertation	Co-Supervisor of Dissertation
Dylan S. Small Professor of Statistics	T. Tony Cai Dorothy Silberberg Professor Professor of Statistics

Graduate Group Chairperson

Eric T. Bradlow
K. P. Chao Professor
Professor of Marketing, Statistics, and Education

Dissertation Committee

Dylan S. Small, Professor of Statistics
T. Tony Cai, Dorothy Silberberg Professor, Professor of Statistics
Paul R. Rosenbaum, Robert G. Putzel Professor, Professor of Statistics
Benjamin F. Voight, Assistant Professor of Pharmacology and Genetics
Nancy R. Zhang, Associate Professor of Statistics

INSTRUMENTAL VARIABLES AND MENDELIAN RANDOMIZATION WITH INVALID INSTRUMENTS

ACKNOWLEDGEMENT

First, I want to thank my advisors, Tony, and Dylan. Tony's brilliant intuition and depth of understanding in theoretical problems, specifically those concerning high dimensional sparse regression and minimax problems, are unparalleled. Dylan's almost-infinite background in statistical, medical, and anything-academic literature, especially those concerning causal inference and instrumental variables in economics and medicine, along with his keen insight connecting seemingly-unrelated fields is exceptional. I couldn't have asked for a more patient and kind advisors helping me understand the rich set of materials related to my thesis. To this day, I am so fortunate that in my second year, both advisors who were coming from different backgrounds in statistics were willing to work on a project regarding high dimensional IVs, which turned into my current thesis on IVs with invalid instruments.

Second, I want to thank my dissertation committee members, Paul, Ben, and Nancy. Without exaggeration, my mind was blown (in a good way) during every discussion with Paul on observational studies, specifically concerning sensitivity analysis. Ben's never-ending enthusiasm and very valuable contextual information, specifically concerning the application of Mendelian randomization methods, his own work on HDL and heart attack, and the importance of summary data, provided a solid basis for my thesis and will pave my future research in applications of Mendelian randomization methods. Finally, Nancy has been a kind, generous, and constant support since my freshmen year undergraduate summer research days in 2007 and I am fortunate to have learned a great deal from her, especially regarding statistical genetics, in the latter years of my graduate career.

Third, my thesis could not have been possible without my collaborators throughout the years, including Benno, Ohene, Jürgen, Nandita, Ralf, Anru, and Afshin. I wanted to particularly thank Benno for providing the malaria data and for sharing his wealth of knowledge about malaria epidemiology in Ghana. I also want to thank Anru for his friendship, genius-level intellect, most notably his ability to prove the convergence of the sisVIVE estimator presented in Chapter 2 in two days, and boundless enthusiasm during our several conversations about life as graduate students. As we both start the next chapter in our career, I am very thrilled to share that journey with him at the same institution for our tenure-track positions.

Fourth, I want to thank the staff in the department, Noelle, Adam, Carol, Andrew, Sarin, Tanya, and Anand, for their unwavering and boundless support during my graduate career, including providing support for TAs, helping with my own personal finances and the headache-inducing reimbursement process, solving various logistical issues related to the doctoral program, troubleshooting software and even personal computer issues, finding administrative contacts for my postdoc, alerting me of free lunches after faculty meetings, tracking mail and packages, and so much more. They were always going above and beyond with any kind of help that I needed and I am so happy that I had the pleasure of working with them. They are, by far, the best administrators that I have ever met.

Fifth, I want to thank friends who I had the pleasure of meeting throughout my doctoral career, including some awesome undergrad students, my cohort-mates (Yang, Josh, Asaf, Raja, and Anru), and many Ph.D. students who I have met over the years (unfortunately, too many to name here). Your friendship (or a combination of your company and memorable concoctions) kept me sane and grounded throughout the ups and downs of my five years here and for that, I am forever grateful.

Finally, I want to thank my parents and my sister. I cannot be here today without their constant encouragement, support, and love throughout the years. You were always there and always a phone call away. You always listened and always knew the right things to say. Thank you for showing me that there is always hope and that there is light at the end of every tunnel. Thank you for being the best parents and sister I could ever wish for.

The counterfactual outcome of my Ph.D. if I met different advisors, faculty, friends, staff, and family is unknown and a “fundamental problem” in causal inference. But, based on observational data about my life, I can say with great confidence that my career, thus far, achieved oracle rates of convergence. All of them were instrumental to my career and that this conclusion is as insensitive to unmeasured confounding as the conclusion about smoking causing lung cancer from observational data.

ABSTRACT

INSTRUMENTAL VARIABLES AND MENDELIAN RANDOMIZATION WITH INVALID INSTRUMENTS

Hyunseung Kang

Dylan S. Small

T. Tony Cai

Instrumental variables (IV) methods have been widely used to determine the causal effect of a treatment, exposure, policy, or an intervention on an outcome of interest. The IV method relies on having a *valid* instrument, a variable that is (A1) associated with the exposure, (A2) has no direct effect on the outcome, and (A3) is unrelated to the unmeasured confounders associated with the exposure and the outcome. However, in practice, finding a valid instrument, especially those that satisfy (A2) and (A3), can be challenging. For example, in Mendelian randomization studies where genetic markers are used as instruments, complete knowledge about instruments' validity is equivalent to complete knowledge about the involved genes' functions.

The dissertation explores the theory, methods, and application of IV methods when invalid instruments are present. First, when we have multiple candidate instruments, we establish a theoretical bound whereby causal effects are only identified as long as less than 50% of instruments are invalid, without knowing which of the instruments are invalid. We also propose a fast penalized ℓ_1 method, called sisVIVE, to estimate the causal effect. We find that sisVIVE outperforms traditional IV methods when invalid instruments are present both in simulation studies as well as in real data analysis.

Second, we propose a robust confidence interval under the multiple invalid IV setting. This work is an extension of our work on sisVIVE. However, unlike sisVIVE which is robust to violations of (A2) and (A3), our confidence interval procedure provides honest coverage even if all three assumptions, (A1)-(A3), are violated.

Third, we study the single IV setting where the one IV we have may actually be invalid. We propose a nonparametric IV estimation method based on full matching, a technique popular in causal inference for observational data, that leverages observed covariates to make the instrument more valid. We propose an estimator along with inferential results that are robust to mis-specifications of the covariate-outcome model. We also provide a sensitivity analysis should the instrument turn out to be invalid, specifically violate (A3).

Fourth, in application work, we study the causal effect of malaria on stunting among children in Ghana. Previous studies on the effect of malaria and stunting were observational and contained various unobserved confounders, most notably nutritional deficiencies. To infer causality, we use the sickle cell genotype, a trait that confers some protection against malaria and was randomly assigned at birth, as an IV and apply our nonparametric IV method. We find that the risk of stunting increases by 0.22 (95% CI: $0.044, 1$) for every malaria episode and is sensitive to unmeasured confounders.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1 : Introduction

1.1. Association Versus Causation in Observational Studies and the Problem of Unmeasured Confounding

"Association does not imply causation." It is an old mantra taught in introductory statistical courses, usually accompanied with comical examples from contemporary news articles such as the "causal" relationship between global average temperatures and the number of pirates (Anderson, 2012) or more serious issues such as the "causal" relationship between childhood vaccinations and autism in children based on a methodologically flawed association analysis (Institute of Medicine of the National Academies: Immunization Safety Review Committee, 2004). With the latter, the deadly mistake of equating association with causation and the fear the study generated have led to a resurgence in preventable childhood diseases in the United States during the 21st century (Omer et al., 2009) along with countless and wasteful public resources dedicated to debunking this myth (Institute of Medicine of the National Academies: Immunization Safety Review Committee, 2004). In fact, often the goal in a scientific inquiry is causal. But, scientists, for costs or other reasons, are left with associational (observational) data to draw causal conclusions. Since association does not imply causation, is all hope of drawing causal conclusions from associational data lost? Are we bound to make the same faulty causal conclusions like the ones discussed above?

The statistical theory of observational studies seeks to provide principles and methods for designing and analyzing associational studies with the aim of connecting association and causation (see Rosenbaum (2002) and Rubin (2005) for an overview). Using the tools developed in observational studies, associational data has often contributed to important findings such as the finding of the 1964 Surgeon's General report that "cigarette smoking is causally related to lung cancer in men," which was based on observational studies with associational data and had a huge impact on public health (United States Surgeon General, 1988). A central problem in observational studies is how to deal with unmeasured confounding. To illustrate, consider a study where we were given observational data about children in sub-Saharan Africa, specifically their malarial infections and height. The goal of the study was to determine whether malarial infections caused a child to have stunted height, i.e. abnormally short height. The problem was that there were other potential explanations for stunted height besides malaria that were unmeasured for in the data, such as the child's daily diet, which can impact his/her growth as well as his/her immune system, making him/her more susceptible to malarial infections. In short, the child's diet was an unmeasured confounder that confounded the causal relationship between malaria and stunted height. Successfully dealing with unmeasured confounding is a central goal in the theory of observational studies.

1.2. A Potential Solution for Unmeasured Confounding: Instrumental Variables and Mendelian Randomization

One method, instrumental variables (IV), has remained a popular tool in statistics for overcoming the problem of unmeasured confounding. IV methods have been widely used in many fields outside of statistics, including economics (Angrist and Krueger, 2001), genomics and epidemiology (Davey Smith and Ebrahim, 2003), sociology (Bollen, 2012), psychology (Gennetian et al., 2008), political science (Sovey and Green, 2011), and countless others. For example, the malaria study mentioned above used one of our proposed IV matching methods in Chapter 4 to conclude, from observational data, that there is a causal effect between repeated malarial episodes and stunting where the risk of stunting increases by 0.22 for every malaria episode (p-value: 0.011, 95% confidence interval: 0.04, 1, see Chapter 5 for more details on this study).

The popularity of IV methods can be attributed to the fact that they alleviate the requirement to conduct a randomized experiment to determine a causal effect. Randomized experiments are the gold standard in determining causal effects. But, they are often expensive and sometimes unethical. For example, with our malarial study, a randomized clinical trial would involve randomly assigning children to receive the malarial parasite at the whim of a coin flip, a highly unethical task. IV methods avoid the need for a traditional randomized experiment by finding an instrument where the instrument is (A1) related to the exposure, (A2) has no direct pathway to the outcome, and (A3) is not related to unmeasured confounders that affect the exposure and the outcome. Recently, IV methods have been applied to genetic data where instruments are genes and the field is known as Mendelian randomization (MR)(Davey Smith and Ebrahim, 2003, 2004). For example, in our malaria study to be discussed in Chapter 5, we used the sickle cell trait, one of the genotypes that determines the shape of a red blood cell, as an instrument (see Figure 1). The sickle cell genotype has been shown to provide protection against malaria, satisfying (A1) (Friedman, 1978; Hill et al., 1991). For satisfying (A2) and (A3), prior studies (Ashcroft et al., 1978; Rehan, 1981) from non-malaria endemic areas, but where the sickle cell trait was present, provided support for the two assumptions (see Chapter 5 for more details).

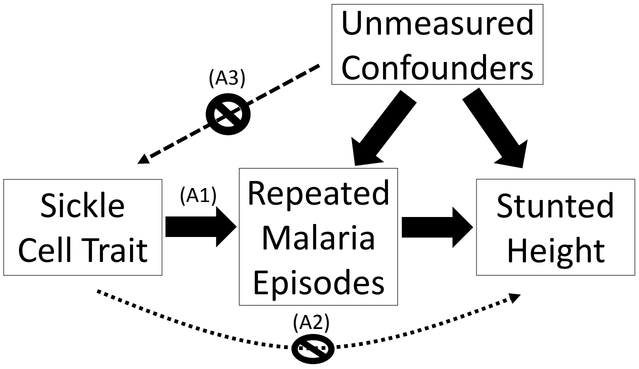

Figure 1: Diagram of instrumental variables assumptions in the malaria study. Arrows represent associations between variables. Absence of arrows indicates no relationship. Numbers (A1), (A2), and (A3) indicate different instrumental variables assumptions.

1.3. A Major Challenge in Instrumental Variables: Finding Valid Instruments

One of the biggest challenges in IV methods is finding an instrument that satisfies the conditions (A1)-(A3). Specifically, satisfying assumption (A2), also known as the no direct effect assumption, has been problematic in many IV studies. For example, if the instruments are genes, as is the case in Mendelian randomization, satisfying (A2) would imply that the gene/instrument's only biological function is to affect the exposure only, i.e. the gene is not pleiotropic. However, this assumption is unreasonable for many genetic markers as they often have multiple functions (Solovieff et al., 2013); in fact, our malaria study explained in Section 1.2 is no exception to this problem.

Many epidemiologists who use genetic instruments are aware of this problem (Davey Smith and Ebrahim, 2003; Lawlor et al., 2008); Lawlor et al. (2008) also describes other types of biological phenomena such as linkage disequilibrium and population stratification, which are unique to IV studies where instruments are genetic, and may violate (A2) and (A3). Unfortunately, without complete biological knowledge of the gene and its plethora of functions or effects by linkage disequilibrium, all IV methods in Mendelian randomization are plagued by possible violations of (A2) and (A3). Also, economists who use IV methods face similar problems, specifically concerning whether their candidate instruments violate (A2) and (A3) (Murray, 2006). Previous IV methods have assumed that there is at least one known valid instrument satisfying (A1)-(A3). However, in many applications, one may have many candidate instruments, but is not sure about the validity of any of them.

Also, in some cases, we may not have many candidate instruments and we may end up with only one candidate instrument. With the one instrument, we have to do our best to make sure that this instrument is valid and to assess the impact on our statistical analysis should this instrument turn out to be invalid despite our best efforts. For example, in our malaria example, the data only provided one instrument, the sickle cell genotype, for us to infer the causal effect of malaria and stunting.

In short, all IV analysis, to varying degree, suffer from the "invalid instrument problem." For example, if we are given multiple candidate instruments, we are never certain whether all of them are valid, that is satisfy (A2) and (A3); it is probably the case that some of them are invalid. As another example, if we are left with only a single candidate instrument, we have to find ways to make the instrument "more valid" and, more importantly, to assess the sensitivity of the statistical analysis should the instrument fail to be valid, despite our best efforts.

1.4. Our Contributions and Outline of Dissertation

Broadly speaking, the thesis tackles the invalid instrument problem into two cases, the case with multiple instruments and the case with one instrument. In Chapter 2, we consider the multiple instrument case where we aren't sure whether these instruments satisfy conditions (A2) and (A3). We show that key parameters in the data generating model can still be identified even without knowing which candidate instruments are valid or invalid by providing both a necessary and sufficient condition for identification. In particular, the causal effect of the treatment on the outcome can always be identified if the number of invalid instruments (denoted by s) is strictly less than 50% of the total candidate instruments L (i.e. $s < L/2$), even if one does not know which of the L instruments are valid and invalid, a priori. If more than 50% of the total candidate instruments may be invalid, then the scientist can check the necessary and sufficient conditions to see whether the parameters in the model are identified.

In line with the identification result, Chapter 2 also proposes a method to estimate the causal effect of the treatment on the outcome if some instruments are invalid, without knowing which instruments are invalid. Our proposed estimator, sisVIVE, is a penalized ℓ_1 estimator, which has theoretical guarantees on performance under certain regularity conditions. Also, in simulation studies and a real data analysis, we show that sisVIVE dominates the most popular IV method, the two stage least squares, whenever invalid instruments are present.

Chapter 3 extends the work in Chapter 2 by providing a robust confidence interval under the settings described in Chapter 2. In particular, we propose a simple and general method to construct confidence intervals that are theoretically guaranteed to provide honest coverage in the presence of invalid instruments.

In Chapter 4, we consider the case where we are only given one candidate instrument.

In this work, we attempt to make the IV assumptions more plausible, specifically (A3), by controlling for measured covariates. Conditional on these covariates, the instrument behaves as if it was a result of random assignment and hence, is unassociated with the unmeasured confounders. We incorporate this idea of conditioning by full matching, which has been shown to have some advantages compared to other methods that condition on covariates (Stuart, 2010). A matching algorithm generates matched sets by grouping individuals in the data who are similar to each other, except for the value of the instrument. For example, if the instrument is binary and is denoted by Z, the matching algorithm may generate I matched sets with each set containing n_k individuals of which m_k have Zs equal to 1 and $n_k - m_k$ have Zs equal to 0.

Once we obtain matched sets, we propose a nonparametric estimator of the causal effect of the exposure on the outcome where we do not assume a parametric model between the outcome Y_i and the covariates $\mathbf{X}_i$. We prove some desirable theoretical properties concerning our nonparametric estimator. We also derive a general formula for computing efficiency of any IV matching-based estimators. Finally, we propose sensitivity analysis if the instrument does violate (A3) even after controlling for the covariates using our nonparametric matching method.

In Chapter 5, we apply the nonparametric full matching technique developed in Chapter 4 by analyzing the malaria example mentioned in Section 1.2. Specifically, the goal in the data analysis is to provide an estimate of the causal effect of malaria episodes on stunted growth in children from Ghana. The novel idea in this work is the use of the sickle cell trait as an instrument. The trait has been known to confer some level of protection against malaria, thereby satisfying (A1). But, it's possible that (A2) and (A3) may be violated. We use the method in Chapter 4 to alleviate some of these concerns and provide an estimate of the causal effect.

CHAPTER 2 : Instrumental Variables With Possibly Invalid Instruments: Theory and Point Estimation

This is joint work with Anru Zhang, Tony Cai, and Dylan Small.

2.1. Motivating Examples of Invalid Instruments in Mendelian Randomization

As mentioned before, the goal in Mendelian randomization (MR) is to estimate the causal effect of an exposure on an outcome by using genetic markers, specifically single nucleotide polymorphisms (SNPs), as instruments (Davey Smith and Ebrahim, 2003, 2004; Lawlor et al., 2008; Wehby et al., 2008). However, there is always concern as to whether these SNPs satisfy the IV assumptions. 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 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 pleoitropic 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 2.1. The author also proposed an estimation strategy, called the moment selection criteria (MSC), to correctly select the valid instruments, which is similar to equation (2.10) in Section 2.3.3. Unfortunately, as we discuss in Section 2.3.3, MSC is computationally infeasible when the number of instruments is large. Kolesár et al. (2013) 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. (2013) 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 pre-defined 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.

We add to the prior literature as follows. First, we show that it is indeed possible to identify and estimate the causal effect without a known pre-defined 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 are 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. 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 IV estimation procedure. 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 a health-related quality of life (HRQL) measure using our new method.

2.2. Causal Model for Instrumental Variables With Invalid Instruments

2.2.1. 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 i.i.d. 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, i.e. the number of non-zero 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 non-zero elements of the vector $\boldsymbol{\alpha}$, i.e. $j \in \text{supp}(\boldsymbol{\alpha})$ if and only if $\alpha_j \neq 0$. A vector $\boldsymbol{\alpha}$ is called s-sparse if it has no more than s non-zero entries. Also, for a vector $\boldsymbol{\alpha} \in \mathbb{R}^L$ and a 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 $\mathbf{P}_{\mathbf{M}}$ be the n by n orthogonal projection matrix onto the column space of $\mathbf{M}$, specifically $\mathbf{P}_{\mathbf{M}} = \mathbf{M}(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T$; it is assumed that $\mathbf{M}^T\mathbf{M}$ has a proper inverse, unless otherwise noted. Let $\mathbf{P}_{\mathbf{M}^\perp}$ be the residual projection matrix, specifically $\mathbf{P}_{\mathbf{M}^\perp} = \mathbf{I} - \mathbf{P}_{\mathbf{M}}$ where $\mathbf{I}$ is an n by n identity matrix.

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

2.2.2. 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

$$Y_i^{(d',\mathbf{z}')} - Y_i^{(d,\mathbf{z})} = (\mathbf{z}' - \mathbf{z})^T \boldsymbol{\phi}^* + (d' - d)\beta^* \tag{2.1}$$

$$E(Y_i^{(0,\mathbf{0})} | \mathbf{Z}_{i.}) = \mathbf{Z}_{i.}^T \boldsymbol{\psi}^* \tag{2.2}$$

where $\boldsymbol{\phi}^*, \boldsymbol{\psi}^* \in \mathbb{R}^L$, and $\beta^* \in \mathbb{R}$ are unknown parameters. In equation (2.1), the parameter β^* represents the causal parameter of interest, the causal effect of changing the exposure by one unit on the outcome. Also in equation (2.1), 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 (2.2), 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 (2.2), they are related via $\boldsymbol{\psi}^*$.

Let $\boldsymbol{\alpha}^* = \boldsymbol{\phi}^* + \boldsymbol{\psi}^*$. When we combine equations (2.1) and (2.2) along with the definition of ϵ_i, we have the observed data model

$$Y_i = \mathbf{Z}_{i.}^T \boldsymbol{\alpha}^* + D_i \beta^* + \epsilon_i, \quad E(\epsilon_i | \mathbf{Z}_{i.}) = 0 \tag{2.3}$$

We make the following remarks regarding the model (2.3). 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 $\mathbf{X}$, where $\mathbf{X}$ is the n by p matrix of covariates, e.g. replace $\mathbf{Y}$ by $(\mathbf{I} - \mathbf{P_X})\mathbf{Y}$ (Wang and Zivot, 1998). The results in this paper 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 non-linear 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 non-linear 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 Jr. (1985), Björklund and Moffitt (1987), and Small (2007), we can incorporate heterogeneous effects as follows. Suppose, instead of equation (2.1), the potential outcomes model for the outcome is

$$Y_i^{(d',\mathbf{z}')} - Y_i^{(d,\mathbf{z})} = (\mathbf{z}' - \mathbf{z})^T \boldsymbol{\phi}^* + (d' - d)\beta_i^* \tag{2.4}$$

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 (2.4) as follows.

$$Y_i = \mathbf{Z}_{i.}^T \boldsymbol{\alpha}^* + D_i\beta^* + (\beta_i^* - \beta^*)D_i + \epsilon_i, \quad E(\epsilon_i | \mathbf{Z}_{i.}) = 0 \tag{2.5}$$

If $(\beta_i^* - \beta^*)$ is independent of D_i given $\mathbf{Z}_{i.}$, the heterogeneous model in (2.5) is identical to model (2.3) and our result for Theorem 2.1 in Section 2.3.1 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 non-zero $\boldsymbol{\alpha}^*$ (see Angrist and Imbens (1995) and Small (2007) for details). For simplicity of exposition, we will focus on a model with a constant linear effect β^*.

2.2.3. Definition of Valid Instruments

Based on the observed model in (2.3), the parameter $\boldsymbol{\alpha}^*$ combines both the direct effect, represented by $\boldsymbol{\phi}^*$, and the effect of confounders on the $\mathbf{Z}_{i\cdot}$ 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 2.1. Suppose we have the models in (2.1) -(2.3) 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 2.1 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 (2.3).

In the case of only one instrument (i.e. $L = 1$), Definition 2.1 of a valid instrument matches with the informal definition (A2) and (A3) in Section 1.2 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 (2.1) 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 (2.2) 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 a valid instrument discussed in Angrist et al. (1996) where our model assumes an additive, linear, and constant treatment effect β^*.

For more than one instruments (i.e. $L > 1$), our model (2.1)-(2.3) and definition of valid IVs can be viewed as a generalization of Holland (1988). It is important to note that in this generalization, Definition 2.1 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 $\mathbf{Z}_{\cdot j}$ may be associated with or causally affect another IV $\mathbf{Z}_{\cdot j'}$, $j \neq j'$ where $\alpha_{j'}^* \neq 0$.

2.3. Estimation of Causal Effect With Invalid Instruments

2.3.1. Identifiability of Model

We first address whether the model in equation (2.3) 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, i.e. $\gamma_j^* \neq 0$ 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 et al., 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 (2.3) implies the following moment condition.

$$E(\mathbf{Z}^T(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha}^* - \mathbf{D}\beta^*)) = 0 \tag{2.6}$$

Suppose assumptions (a) and (b) hold. Then, the moment equation in equation (2.6) simplifies to

$$\boldsymbol{\Gamma}^* = \boldsymbol{\alpha}^* + \boldsymbol{\gamma}^* \beta^* \tag{2.7}$$

where $\boldsymbol{\Gamma}^* = E(\mathbf{Z}^T\mathbf{Z})^{-1}E(\mathbf{Z}^T\mathbf{Y})$. Since both $\boldsymbol{\Gamma}^*$ and $\boldsymbol{\gamma}^*$ 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 $\boldsymbol{\alpha}^*, \beta^*$ and $\boldsymbol{\Gamma}^*, \boldsymbol{\gamma}^*$, i.e. a unique solution of $\boldsymbol{\alpha}^*, \beta^*$ 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 (2.7) 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 et al., 1950) and is commonly used in the traditional IV literature.

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 (2.7) 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 (2.7). 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 (2.7).

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 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, i.e. 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 (2.7) can exist and this is stated in Theorem 2.1.

Theorem 2.1 (Uniqueness of Solution)**.** *Suppose we assume assumptions (a) and (b) and the modeling assumption* (2.3)*. Let $s \in \{0, 1, \dots, L\}$ with $s < U$ where $U = 1, \dots, L$. Consider all sets $C_m \subseteq \{1, \dots, L\}, m = 1, \dots, M$ 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 (2.7) *if and only if $q_m = q_{m'}$ for all $m, m' \in \{1, \dots, M\}$.*

To understand Theorem 2.1, note that if the valid instruments are those in the set C_m, then the causal effect $\beta^* = q_m$. More specifically, Theorem 2.1 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 2.1 that there is a unique solution to $\boldsymbol{\alpha}^*$ and β^* to (2.7) if and only if $q_m = q_{m'}$ for all $m, m' \in \{1, \dots, M\}$ the *consistency criterion.* We thank Jack Bowden for his insight and suggestions on terminology for interpreting Theorem 2.1.

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 $\{1, \dots, L\}$ and the constants q_m associated with $\boldsymbol{\Gamma}^*$ and $\boldsymbol{\gamma}^*$. Corollary 2.1 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 2.1. *If $U \leq L/2$, there is always a unique solution to* (2.7)

In addition to the computational benefits, compared to Theorem 2.1, Corollary 2.1 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 2.1.

We would like to mention two final points about Theorem 2.1. First, Theorem 2.1 is a statement about uniqueness of solutions for the parameters $\boldsymbol{\alpha}^*$, and β^* in equation (2.7). A natural question to ask is whether the uniqueness is guaranteed for just β^*, the causal effect of interest, at the expense of non-uniqueness of $\boldsymbol{\alpha}^*$. In the proof of Theorem 2.1 in the Appendix, we show that this cannot be the case. Specifically, regardless of the condition on s, the parameter β^* is a unique solution to (2.7) if and only if the parameter $\boldsymbol{\alpha}^*$ is a unique solution to (2.7). Second, Theorem 2.1 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 2.1 in the Appendix, we provide a rigorous argument that, indeed, under model (2.3) and $s < U$, at least one set C_m has to exist.

2.3.2. Some Examples of Identified Models Using Theorem 2.1

To illustrate the nature of identified models with invalid instruments, specifically in relation to Theorem 2.1, we consider a couple of examples. First, let us revisit the earlier numerical example in Section 2.3.1 with $\boldsymbol{\gamma}^* = (1,2,3,4)$ and $\boldsymbol{\Gamma}^* = (1,2,3,8)$. Suppose our prior knowledge on the upper bound on s is 3, i.e. $U = 3$. Then, by Theorem 2.1 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 2.1 and we have a unique solution $\boldsymbol{\alpha}^*$ and β^* to (2.7). 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 2.1 because $q_1 \neq q_2$ and there are no unique solutions $\boldsymbol{\alpha}^*$ and β^* to (2.7).

One of the reviewers, however, mentioned an extension of this numerical example where the setup is identical except $\boldsymbol{\Gamma}^*$ is perturbed by $\epsilon > 0$ such that $\tilde{\boldsymbol{\Gamma}}^* = (1,2,6,8+\epsilon)$. With $\tilde{\boldsymbol{\Gamma}}^*$, there is only one set $C_1 = \{1,2\}$ where $q_1 = 1$ and we have identification for any ϵ. However, we can shrink ϵ to be arbitrary small such that $\boldsymbol{\Gamma}^*$ and $\tilde{\boldsymbol{\Gamma}}^* = (1,2,6,8+\epsilon)$, are arbitrarily close to each other.

However, consider the identical setup as before, except $\boldsymbol{\Gamma}^* = (1,2,7,9)$. Then, there is only one subset $C_1 = \{1,2\}$ where $q_1 = 1$ and identification is achieved. Furthermore, any small perturbation of $\boldsymbol{\Gamma}^*$ by $\delta > 0$ and $\epsilon > 0$, i.e. $\tilde{\boldsymbol{\Gamma}}^* = (1,2,7+\delta,9+\epsilon)$, will still produce only subset $C_1 = \{1,2\}$ and identification is maintained.

The two numerical examples with $\boldsymbol{\Gamma}^* = (1,2,6,8)$ and $\boldsymbol{\Gamma}^* = (1,2,7,9)$ illustrate what we call the *identification boundary.* The vector $\boldsymbol{\Gamma}^* = (1,2,6,8)$ lies just at the identification boundary where any small perturbation can render the model unidentified or identified. In contrast, for $\boldsymbol{\Gamma}^* = (1,2,7,9)$, the vector $\boldsymbol{\Gamma}^*$ lies far from the identification boundary and any small perturbation can still make the model identifiable. Exploration of the identification boundary for different values of $\boldsymbol{\Gamma}^*$ and $\boldsymbol{\gamma}^*$ is a topic for future research.

As a second example of identification using Theorem 2.1, we consider the classical linear simultaneous/structural equations model in econometrics (Koopmans et al., 1950). To do so, we impose two additional modeling assumptions which are not needed for identification, but are part of the classical ecnoometrics model, and discuss the identification result in 2.3.1 under this context. The first additional modeling assumption is that the relationship between D_i and $\mathbf{Z}_{i.}$ is linear

$$D_i = \mathbf{Z}_{i.}^T \boldsymbol{\gamma}^* + \xi_i, \quad E(\xi_i | \mathbf{Z}_{i.}) = 0 \tag{2.8}$$

where $\boldsymbol{\gamma}^*$ relates the instruments to the exposure. The second additional assumption is that the error terms are bivariate Normal

$$(\epsilon_i, \xi_i) \overset{\text{iid}}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}) \tag{2.9}$$

Under these assumptions in (2.8) and (2.9), the distributions of Y_i and D_i conditional on $\mathbf{Z}_{i.}$ are fully characterized by finite-dimensional parameters $\boldsymbol{\alpha}^*, \beta^*, \boldsymbol{\gamma}^*$, and $\boldsymbol{\Sigma}$ known as "structural" parameters in econometrics (Wooldridge, 2010). Let $\epsilon_i' = \beta^* \xi_i + \epsilon_i$. Then, we have the "reduced forms" (Wooldridge, 2010)

$$\begin{aligned} Y_i &= \mathbf{Z}_{i.}^T \boldsymbol{\Gamma}^* + \epsilon_i' \\ D_i &= \mathbf{Z}_{i.}^T \boldsymbol{\gamma}^* + \xi_i \end{aligned}$$

where $\boldsymbol{\Gamma}^* = \boldsymbol{\alpha}^* + \beta^* \boldsymbol{\gamma}^*$ and the covariance matrix of (ϵ_i', ξ_i) is $\boldsymbol{\Sigma}' = \mathbf{M} \boldsymbol{\Sigma} \mathbf{M}^T$ with

$$M = \begin{pmatrix} 1 & \beta^* \\ 0 & 1 \end{pmatrix}$$

We see that the distributions of Y_i and D_i are also fully characterized by the reduced form parameters $\boldsymbol{\Gamma}^*, \boldsymbol{\gamma}^*$ and $\boldsymbol{\Sigma}'$. By Rothenberg (1971), the reduced form parameters, $\boldsymbol{\Gamma}^*, \boldsymbol{\gamma}^*$, and $\boldsymbol{\Sigma}'$, are globally identified. Also, by Rothenberg (1971), the structural parameters, $\boldsymbol{\alpha}^*$, β^*, $\boldsymbol{\gamma}^*$, and $\boldsymbol{\Sigma}$, are identified if and only if the mapping between the reduced form parameters, $\boldsymbol{\Gamma}^*, \boldsymbol{\gamma}^*, \boldsymbol{\Sigma}'$, and the structural parameters, $\boldsymbol{\alpha}^*$, β^*, $\boldsymbol{\gamma}^*, \boldsymbol{\Sigma}$, represented by equations $\boldsymbol{\Sigma}' = \mathbf{M} \boldsymbol{\Sigma} \mathbf{M}^T$, $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}^*$, and $\boldsymbol{\Gamma}^* = \boldsymbol{\alpha}^* + \beta^* \boldsymbol{\gamma}^*$, is bijective. We see that $\mathbf{M}$ is an invertible matrix for any β^* and hence there is a bijective map between $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}'$. For $\boldsymbol{\gamma}^*$, it maps onto itself between the structural and reduced form parameters. Consequently, whether there is a bijection between the structural parameters and reduced form parameters is determined only by whether there is a unique solution $\boldsymbol{\alpha}^*$ and β^* to equation 2.7 given $\boldsymbol{\gamma}^*$ and $\boldsymbol{\Gamma}^*$. Theorem 2.1 states that a unique solution $\boldsymbol{\alpha}^*$ and β^* of (2.7) exists if and only if the consistency criterion holds, that $q_m = q_{m'}$ for all $m, m' \in \{1, \ldots, M\}$. Hence, with the modeling assumptions (2.8) and (2.9), we have identification of the structural parameters if and only if the consistency criterion in Theorem 2.1 holds.

2.3.3. Estimation of the Causal Effect of Exposure on Outcome

Given the model (2.3) and $s < U$, Theorem 2.1 lays out the sufficient and necessary condition for finding a unique solution to the moment equation (2.6). Specifically, if the model is identified, the moment equation (2.6) 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 (2.6) subject to the parameter constraint that $s < U$. Formally, we can write this estimation strategy as

$$\underset{\boldsymbol{\alpha},\beta}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{P_Z}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha} - \mathbf{D}\beta)\|_2^2, \quad s.t. \quad ||\boldsymbol{\alpha}||_0 < U \tag{2.10}$$

where $||\boldsymbol{\alpha}||_0$ is the number of non-zero entries of $\boldsymbol{\alpha}$ and by Definition 2.1, $s = ||\boldsymbol{\alpha}||_0$. Equation (2.10) is similar to the moment selection criterion (MSC) in Andrews (1999). However, both the moment selection criterion in Andrews (1999) and (2.10) 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 (2.10) 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 (2.10) can be written as

$$\underset{\boldsymbol{\alpha},\beta}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{P_Z}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha} - \mathbf{D}\beta)\|_2^2, \quad s.t. \quad ||\boldsymbol{\alpha}||_1 \leq t \tag{2.11}$$

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 paper, we propose the equivalent Lagrangian form as our estimator of the causal effect, called *some invalid some valid IV estimator*, or sisVIVE, as follows

$$(\hat{\boldsymbol{\alpha}}_\lambda, \hat{\beta}_\lambda) \in \underset{\boldsymbol{\alpha},\beta}{\operatorname{argmin}} \ \frac{1}{2}\|\mathbf{P_Z}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha} - \mathbf{D}\beta)\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1 \tag{2.12}$$

for some tuning parameter $\lambda > 0$ and λ corresponds to t in (2.11). If $\lambda = 0$ in (2.12), then (2.12) is the popular two stage least squares (TSLS) estimator, which is equivalent to the generalized method of moments (GMM) estimator when $\boldsymbol{\epsilon}$ in Section 2.2.2 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 the traditional Lasso and the procedure proposed by Gautier and Tsybakov (2011), our procedure in (2.12) only penalizes $\boldsymbol{\alpha}^*$. The estimator (2.12) 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 (2.3) are not all exogenous; specifically, model (2.3) 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 equation (2.12), 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 (2.12). This might seem to be a concern as there are multiple estimates of β^*. However, as we will show in Section 2.3.5, all minimizers of (2.12) are close to the true values β^*. Also, if the entries of the matrix $\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{Z}$ where $\hat{\mathbf{D}} = \mathbf{P_Z D}$ (i.e. the predicted value of the exposure given the instruments) are drawn from a continuous distribution, then the solution to (2.12) 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 (2.12) simply due to their original units.

2.3.4. 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 $||\mathbf{P_Z}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha} - \mathbf{D}\beta)||_2$ 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 (2.6), Sections 2.3.1 and 2.3.3). 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 2.4.1, 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 Kang et al. (2015), we discuss another method of choosing λ, in particular, choosing λ based on the theoretical guidance from Theorem 2.2 and Corollary 2.2. In short, we 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 $\hat{\alpha}_j$ 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 we demonstrate that cross validation achieves this goal in a wide variety of settings.

2.3.5. Estimation Performance

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

Definition 2.2. 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

$$\delta_k^-(\mathbf{M})\|\boldsymbol{\alpha}\|_2^2 \leq \|\mathbf{M}\boldsymbol{\alpha}\|_2^2 \leq \delta_k^+(\mathbf{M})\|\boldsymbol{\alpha}\|_2^2 \tag{2.13}$$

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 (2013a) 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 (2.12).

Theorem 2.2 (Estimation performance of sisVIVE under RIP)**.** *Suppose we have the model given in* (2.3)*. 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* (2.13)*, where s is the number of invalid instruments. Suppose*

$$2\delta_{2s}^-(\mathbf{Z}) > \delta_{2s}^+(\mathbf{Z}) + 2\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z}) \tag{2.14}$$

holds. Then, the estimate $\hat{\beta}_\lambda$ *given by* (2.12) *with tuning parameter* $\lambda \geq 3\|\mathbf{Z}^T\mathbf{P}_{\hat{\mathbf{D}}^\perp}\boldsymbol{\epsilon}\|_\infty$ *has the following performance guarantee*

$$|\hat{\beta}_\lambda - \beta^*| \leq \frac{|\hat{\mathbf{D}}^T\boldsymbol{\epsilon}|}{\|\hat{\mathbf{D}}\|_2^2} + \frac{1}{\|\hat{\mathbf{D}}\|_2}\left(\frac{(4/3\sqrt{5})\lambda\sqrt{s\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z})}}{2\delta_{2s}^-(\mathbf{Z}) - \delta_{2s}^+(\mathbf{Z}) - 2\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z})}\right). \tag{2.15}$$

Condition (2.14) 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 (2.14) 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

$$\mu = \max_{i \neq j} |\mathbf{Z}_{.i}^T \mathbf{Z}_{.j}| \quad \text{and} \quad \rho = \max_{j} |\hat{\mathbf{D}}^T \mathbf{Z}_{.j}| / \|\hat{\mathbf{D}}\|_2. \tag{2.16}$$

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 2.3.1 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 2.4.1 for more discussion). Also, ρ stands in contrast to Condition (b) in Theorem 2.1 which looks at the individual values of $\gamma_j, j = 1, \ldots, L$, instead of the maximum of γ_js.

Given the two MIP constants μ and ρ, we have the following result on estimation performance. Like Theorem 2.2, Corollary 2.2 characterizes all the minimizers $\hat{\beta}_\lambda$ from sisVIVE in (2.12).

Corollary 2.2 (Estimation performance of sisVIVE under MIP)**.** *Let the MIP constants μ and ρ be given in* (2.16)*. If the number of invalid instruments, s, satisfies*

$$s < \min(\frac{1}{12\mu}, \frac{1}{10\rho^2}) \tag{2.17}$$

the estimate $\hat{\beta}_\lambda$ given by (2.12) *with tuning parameter $\lambda \geq 3\|\mathbf{Z}^T \mathbf{P}_{\hat{\mathbf{D}}^\perp} \boldsymbol{\epsilon}\|_\infty$ has the following* *performance guarantee*

$$|\hat{\beta}_\lambda - \beta^*| \leq \frac{|\hat{\mathbf{D}}^T \boldsymbol{\epsilon}|}{\|\hat{\mathbf{D}}\|_2^2} + \frac{1}{\|\hat{\mathbf{D}}\|_2} \left(\frac{4\sqrt{105}/9\lambda s\rho}{1 - s(5\rho^2 + 6\mu)} \right). \tag{2.18}$$

We make the following remarks. First, in the Appendix, we show the condition in equation (2.17) directly implies the condition in equation (2.13). Note that the converse is not true. For example, suppose the matrix of instruments $\mathbf{Z}$ is an n by L matrix where each entry Z_{ij} are from i.i.d. standard Normal. Based on Theorem 5.2 in Baraniuk et al. (2008), when $n \geq Cs\log(L/s)$ for some C not dependent on L and s, we are able to ensure the RIP condition $2\delta_{2s}^-(\mathbf{Z}) > 3\delta_{3s}^+(\mathbf{Z})$ with high probability. Here, $2\delta_{2s}^-(\mathbf{Z}) > 3\delta_{3s}^+(\mathbf{Z})$ is a stronger condition than $2\delta_{2s}^-(\mathbf{Z}) > \delta_{3s}^+(\mathbf{Z}) + 2\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z})$, the RIP condition we need for Theorem 2.2. However, based on Theorem 8 in Cai et al. (2013), to guarantee our MIP condition $\mu < \frac{1}{12s}$, we need $n \geq Cs^2 \log L$ for some C not dependent on L and s. In short, when the order of n is between $s\log(L/s)$ and $s^2 \log L$, $\mathbf{Z}$ meet the RIP condition but not the MIP condition, with high probability.

Second, the constraint on the number of invalid instruments, s, in Corollary 2.2 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 2.2 is not useful in characterizing the performance of sisVIVE. In Section 2.4.2, we study the behavior of sisVIVE when this constraint in (2.17) 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 (2.18), 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 (2.17) allows for more invalid instruments, the upper bound (2.18) becomes worse and our estimator $\hat{\beta}_\lambda$ will be far from β^*.

Finally, we emphasize that the conditions in both Theorem 2.2 and Corollary 2.2 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 2.4.2).

2.3.6. 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 2.3 outlines a two-step numerical method whose solution is identical to sisVIVE in (2.12), but is as fast as ordinary least squares.

Theorem 2.3 (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:*

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

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 (2.12)

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 design matrix $\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{Z}$; remember, sisVIVE in (2.12) is not the standard Lasso problem as discussed in Section 2.3.3. 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 ordinary least squares. 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 ordinary least squares. Best of all, the estimate from this two-step algorithm is identical to sisVIVE.

2.4. Simulation

2.4.1. General Setup

We conduct various simulation studies to study the estimation performance between sisVIVE,two stage least squares (TSLS), the most popular estimator in IV and MR, and ordinary least squares (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 potential candidate instruments. The observations $(Y_i, D_i, \mathbf{Z}_{i.}), i = 1, \ldots, n$ are generated by

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

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 the following parameters

(i) the number of invalid instruments (L)

(ii) the endogeneity parameter ($\sigma_{\epsilon\xi}^*$)

(iii) the direct effect parameter $\boldsymbol{\alpha}^* = (1, 1, \ldots, 0, 0)$ (s in $\|\boldsymbol{\alpha}^*\|_0 = s$)

(iv) the pairwise correlation between instruments (μ in equation (2.16))

(v) the correlation structure between instruments (covariance matrix of $\mathbf{Z}_{i.}$)

(vi) the absolute/overall strength of instruments (concentration parameter)

(vii) the relative strength of all instruments (individual elements of $\boldsymbol{\gamma}^*$)

(viii) the relative strength between invalid and valid instruments ($\boldsymbol{\gamma}^*_{A^*}$ and $\boldsymbol{\gamma}^*_{(A^*)^C}$ where $A* = \text{supp}(\boldsymbol{\alpha}^*)$)

In particular, for (i), we let $L = 10$ and $L = 100$. For (ii), we vary $\sigma^*_{\epsilon\xi}$ from 0 to 0.9. For (iii), we vary s from 0 to 9. For (iv), we set μ at four different values, $0, 0.25, 0.5$, and 0.75. For (v), we consider three types of correlation structures between instruments. The first case is where all the pairwise correlation between instruments is set to μ, i.e. the off-diagonal elements of the covariance matrix for $\mathbf{Z}_{i.}$ is set to μ. The second case is where only the pairwise correlation between valid instruments is set to μ and the pairwise correlation between invalid instruments is set to μ. However, there is no correlation between any pair consisting of one valid and one invalid instrument. The third case is where the pairwise correlation between a valid instrument and an invalid instrument is set to μ. However, there is no pairwise correlation between any pair of valid instruments or any pair of invalid instruments.

For (vi), 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 is 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 et al., 2002). Based on Table 1 in Stock et al. (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 a set of instruments with a concentration parameter (scaled by the number of valid instruments) of around 100 is considered strong in the absolute/overall sense. We use these concentration parameters, 10 and 1000, to var the absolute, overall strength of the instruments. For (vii), 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, say 10, 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 (viii), we look at two cases, the case where the invalid instruments are "stronger" than the valid instruments and the case where the valid instruments are "stronger" than the invalid instruments. To simulate these two new cases, we first fix the concentration parameter from the setup in (vi). Then, for the case when the invalid instruments are "stronger" than the valid instruments, we find $\boldsymbol{\gamma}^*$ where $\gamma_j^* = 2 * \gamma_k^*$ for $j \in$supp$(\boldsymbol{\alpha}^*)$ (i.e. set of invalid instruments) and $k \in$supp$(\boldsymbol{\alpha}^*)^C$ (i.e. set of invalid instruments). In other words, the γ_j^*s associated with invalid instruments have twice the magnitude of the γ_j^*s associated with the valid instruments. For the case when the valid instruments are "stronger" than the invalid instruments, we flip the roles of j and k where j now belongs to supp$(\boldsymbol{\alpha}^*)^C$ and k belongs to supp$(\boldsymbol{\alpha}^*)$.

For each simulation setting, we repeat the simulation either 500 or 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 2.3.4. We also compute estimates 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, i.e. 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).

2.4.2. Simulation Setup 1: $L = 10$, Pairwise Correlation Between All IVs and Uniform IV Strength Between Valid and Invalid IVs

This setup has 10 candidate instruments (i.e. $L = 10$ in (i)), there is pairwise correlation between all instruments (i.e. the first case in (v)), and there is no distinction between invalid and valid IVs with regards to strength (i.e. we ignore (viii)). All other parameters described in the previous section are varied.

Figure 2 shows the estimation error for β^* when endogeneity is varied (i.e. vary (ii)). 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 various values of $j \neq k$) 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 an 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 for β^* when the number of invalid instruments is

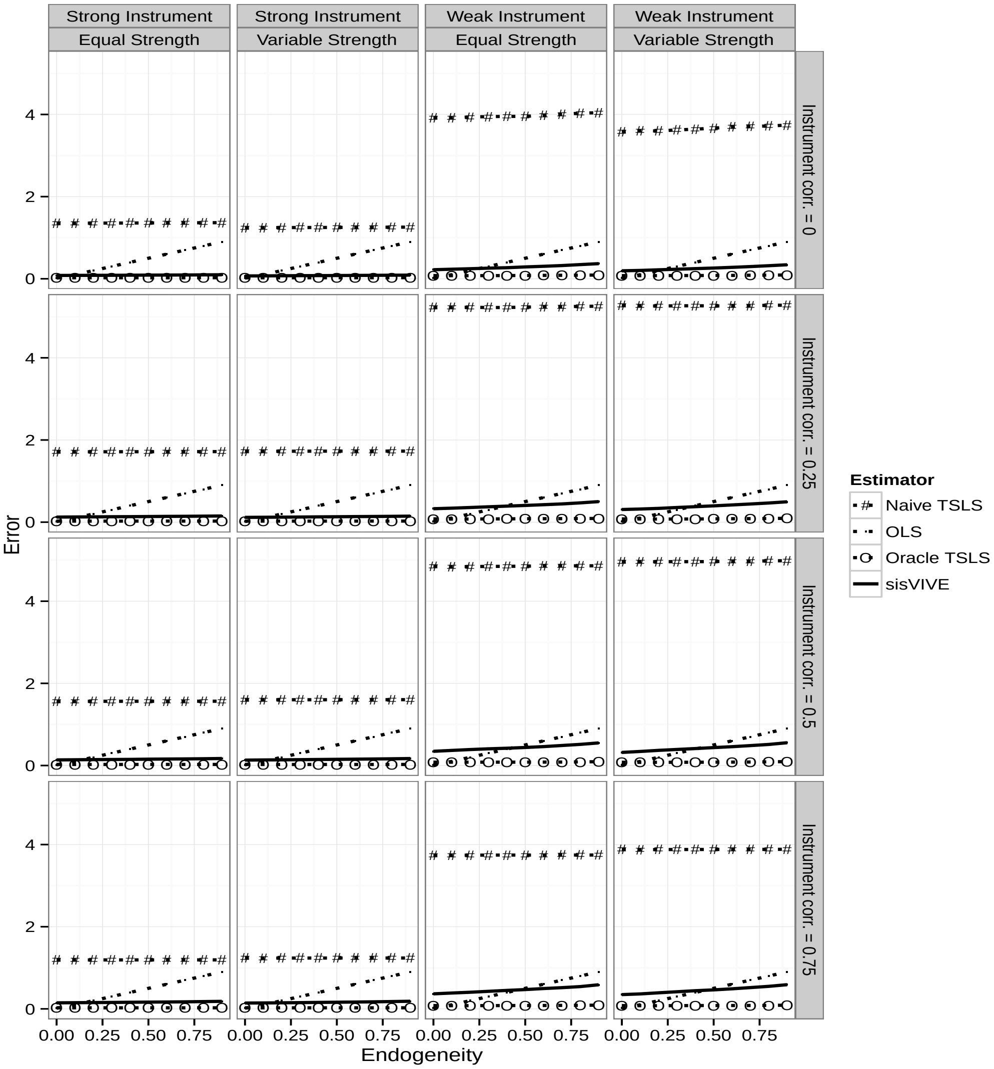

Figure 2: sisVIVE Simulation Study of β^* With Different Endogeneity and Where Correlation Exists Between All IVs (Setup 1). There are ten ($L = 10$) instruments. Each line represents median absolute estimation error ($|\beta^* - \hat{\beta}|$) after 1000 simulations. We fix the number of invalid instruments 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 $\boldsymbol{\gamma}^*$ while holding the absolute strength (i.e. concentration parameter) fixed. Each row corresponds to the maximum correlation between instruments.

Instrument Corr. (μ)	Strong Instrument, Equal Strength	Strong Instrument, Variable Strength	Weak Instrument, Equal Strength	Weak Instrument, Variable Strength
0	0.31	0.39	0.20	0.22
0.25	0.54	0.58	0.36	0.37
0.5	0.72	0.73	0.53	0.53
0.75	0.87	0.87	0.73	0.73

Table 1: Values of ρ in Corollary 2.2 for sisVIVE Simulation Study (Setup 1)

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 $s = 0$, i.e. 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 2.1, $s < L/2 = 5$, 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 ρ and μ found in the condition for Corollary 2.2 from the simulated data. Specifically, we computed ρ from each simulated data set and take the median value of it after 1000 simulations. For μ, we use the true values of the correlation of $\mathbf{Z}_{i.}$, specifically $\mu = 0, 0.25, 0.5$, and 0.75. Table 1 shows the results.

Using Table 1, we see that the top lefthand plot of Figure 2 in our simulation study has ρ of approximately 0.31 and $\mu = 0$. Based on this, the upper bound on s in Corollary 2.2 is 1.04. However, since $s = 3$ for the simulations in Figure 2, the condition (2.17) in Corollary

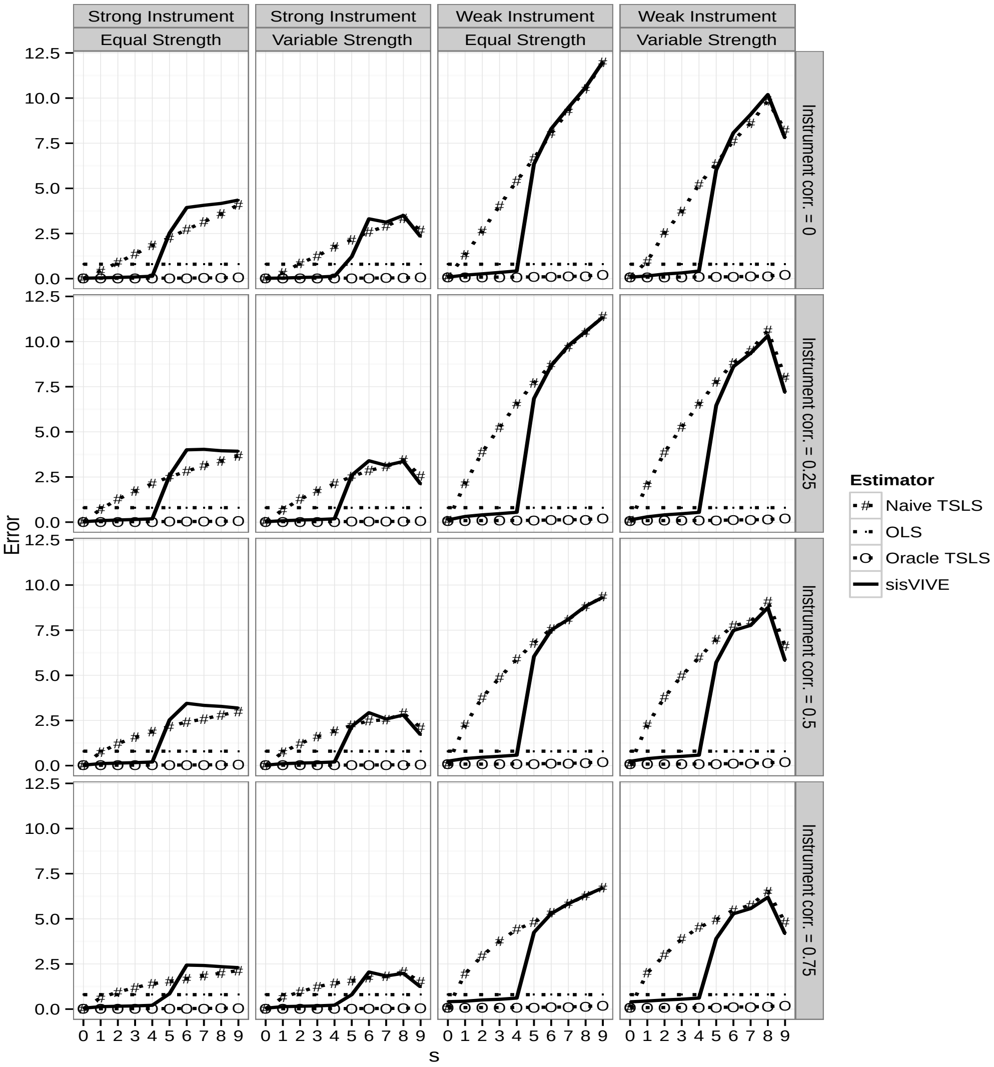

Figure 3: sisVIVE Simulation Study of β^* With Different Number of Invalid IVs and Where Correlation Exists Between All IVs (Setup 1). There are ten ($L = 10$) instruments. Each line represents median absolute estimation error ($|\beta^* - \hat{\beta}|$) after 1000 simulations. We fix the endogeneity $\sigma^*_{\epsilon\xi}$ 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 $\boldsymbol{\gamma}^*$ while holding the absolute strength fixed. Each row corresponds to maximum correlation between instruments.

2.2 is violated and cannot be used to characterize the behavior of sisVIVE.

Table 2 shows the condition required by Corollary 2.2, specifically the upper bound on s in (2.17), for all values of ρ and μ in Table 1. Based on Table 2, the condition for Corollary 2.2 is only satisfied when $s = 0$, i.e. when there are no invalid instruments, for vast majority of cases. For example, when instrument are correlated and $\mu > 0$, Corollary 2.2 cannot be used to characterize the performance of sisVIVE if invalid instruments are present. Table 2 also re-illustrates the point in Section 2.3.5 that the condition for Corollary 2.2, even though it's interpretable, are strict and that Theorem 2.2 is a generalization of Corollary 2.2 at the expense of interpretability.

Instrument Corr. (μ)	Strong Instrument, Equal Strength	Strong Instrument, Variable Strength	Weak Instrument, Equal Strength	Weak Instrument, Variable Strength
0	1.04	0.66	2.50	2.07
0.25	0.33	0.33	0.33	0.33
0.5	0.17	0.17	0.17	0.17
0.75	0.11	0.11	0.11	0.11

Table 2: Condition on s in Corollary 2.2 for sisVIVE Simulation Study (Setup 1)

Overall, in this setting, we find that in terms of absolute estimation error, $|\beta^* - \hat{\beta}_\lambda|$, sisVIVE dominates TSLS whenever there are invalid IVs and its performance is similar to the oracle. Also, we find that Corollary 2.2, while interpretable, provides a poor theoretical characterization of sisVIVE's performance.

2.4.3. Simulation Setup 2: $L = 10$, Pairwise Correlation Between Subsets of IVs and Uniform IV strength Between Valid and Invalid IVs

This simulation setup is identical to Section 2.4.2, except we have two different types of pairwise correlation between subsets of instruments instead of having pairwise correlation between all instruments. Specifically, Figures 4 and 5 represent the setting where the pairwise correlation between valid instruments is set to μ and the pairwise correlation between invalid instruments is also set to μ. However, there is no correlation between any pair consisting of one valid and one invalid instrument. Figures 6 and 7 represent the setting where the pairwise correlation between a valid instrument and an invalid instrument is set to μ. However, there is no pairwise correlation between any pair of valid instruments or any pair of invalid instruments.

Figures 4 and 6 vary endogeneity, but the number of invalid instruments is fixed at $s = 3$. The behavior of all the estimators are similar to each other and to those in Section 2.4.2, specifically Figure 2. OLS dominates naive TSLS, oracle TSLS, and sisVIVE when endogeneity is small and close to zero, with the dominance being greater for weaker instruments. Once there is a sufficient amount of endogeneity, oracle TSLS, which knows exactly which instruments are valid and invalid, does best. sisVIVE also resembles the oracle in terms of performance. Naive TSLS, which assumes all the L instruments are valid, does worst since it assumes that all the L instruments are valid.

Similarly, Figures 5 and 7 vary the number of invalid instruments, s, but fix the endogeneity to 0.8. The estimators behave similarly across the two figures and to those in Section 2.4.2, specifically Figure 3. We first see that at $s = 0$, i.e. when there are no invalid instruments, sisVIVE's performance is nearly identical to naive and oracle TSLS, although it degrades slightly for instruments with weak absolute strength. Also, when $s < L/2 = 5$, sisVIVE's performance is comparable to oracle TSLS and better than naive TSLS. Once we reach the identification boundary, $s < L/2 = 5$, sisVIVE's performance becomes similar to naive TSLS. This is the case regardless of the instruments' absolute and relative strength.

Overall, in this setting, we find that the three correlation structures produce similar simulation results with regards to the estimation error $|\beta^* - \hat{\beta}_\lambda|$.

2.4.4. Simulation Setup 3: $L = 10$*, Pairwise Correlation Between All IVs and Non-Uniform IV Strength Between Valid and Invalid IVs*

This simulation setup is identical to Section 2.4.2, except we consider two other types of instrument strength, specifically the case where the invalid instruments are "stronger" than

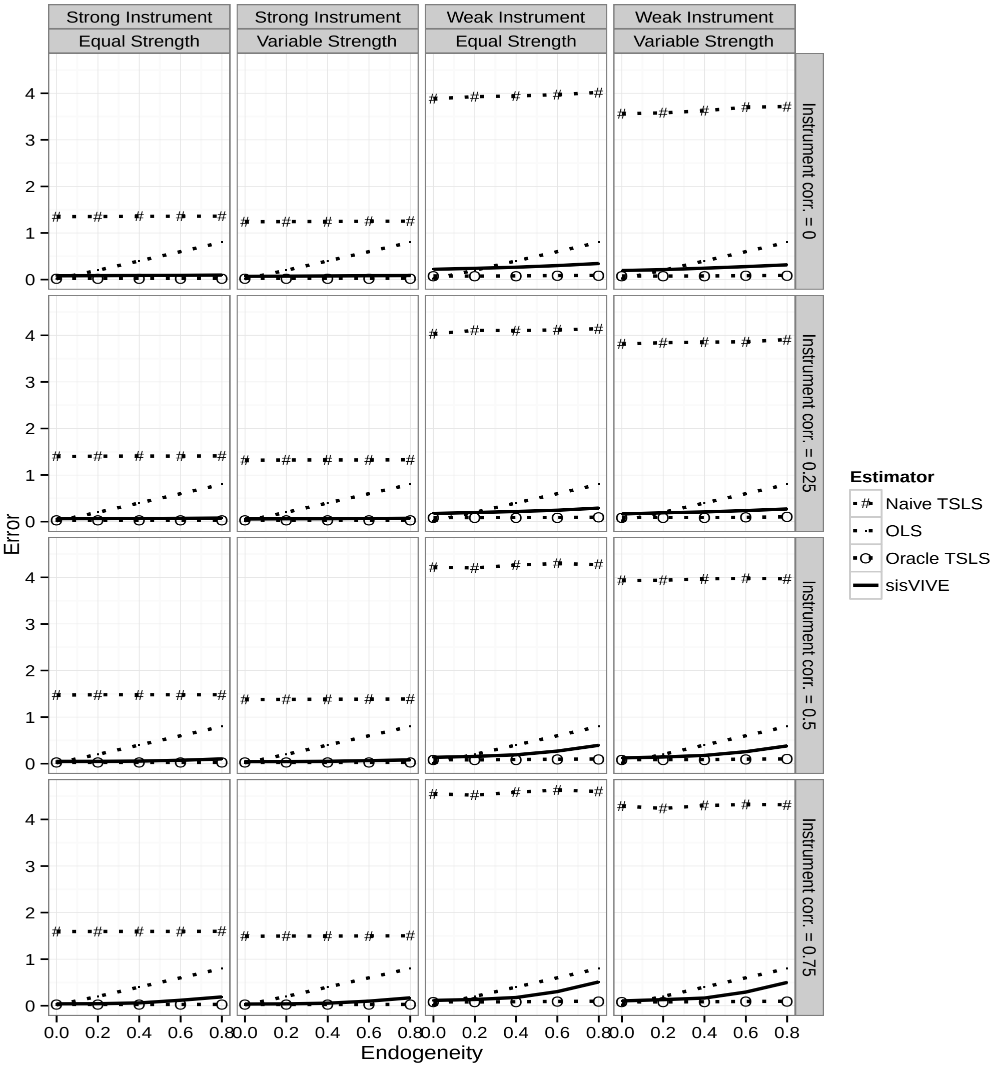

Figure 4: sisVIVE Simulation Study of β^* With Different Endogeneity and Where Correlation Only Exists Within Valid and Invalid IVs (Setup 2). There are ten ($L = 10$) instruments. Each line represents the median absolute estimation error ($|\beta^* - \hat{\beta}|$) after 500 simulations. We fix the number of invalid instruments to $s = 3$. Each column in the plot corresponds to 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 $\boldsymbol{\gamma}^*$ while holding the absolute strength (i.e. concentration parameter) fixed. Each row corresponds to the maximum correlation between instruments, but correlation only exists within valid and invalid instruments.

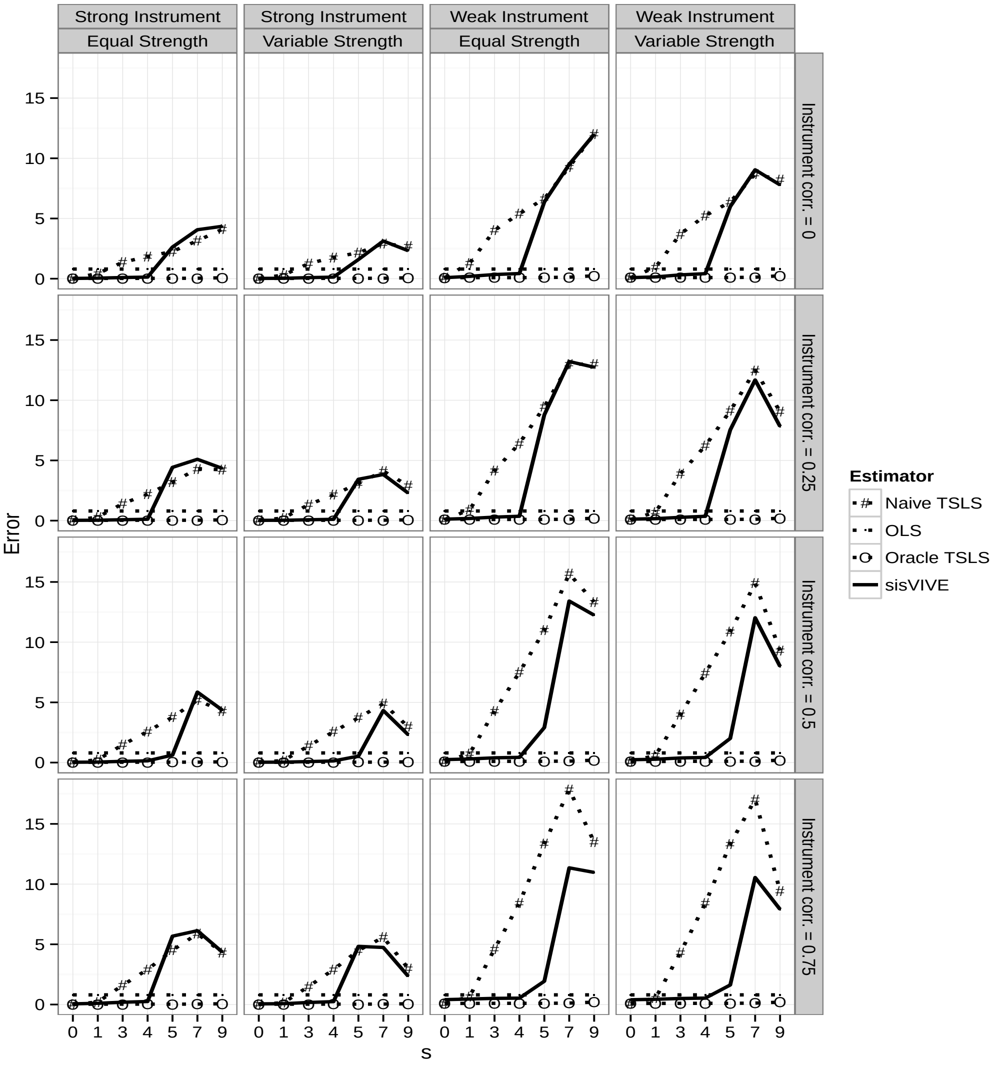

Figure 5: sisVIVE Simulation Study of β^* With Different Number of Invalid IVs and Where Correlation Only Exists Within Valid and and Invalid IVs (Setup 2). There are ten ($L = 10$) instruments. Each line represents the median absolute estimation error ($|\beta^* - \hat{\beta}|$) after 500 simulations. 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 $\boldsymbol{\gamma}^*$ while holding the absolute strength fixed. Each row corresponds to the maximum correlation between instruments, but correlation only exists within valid and invalid instruments.

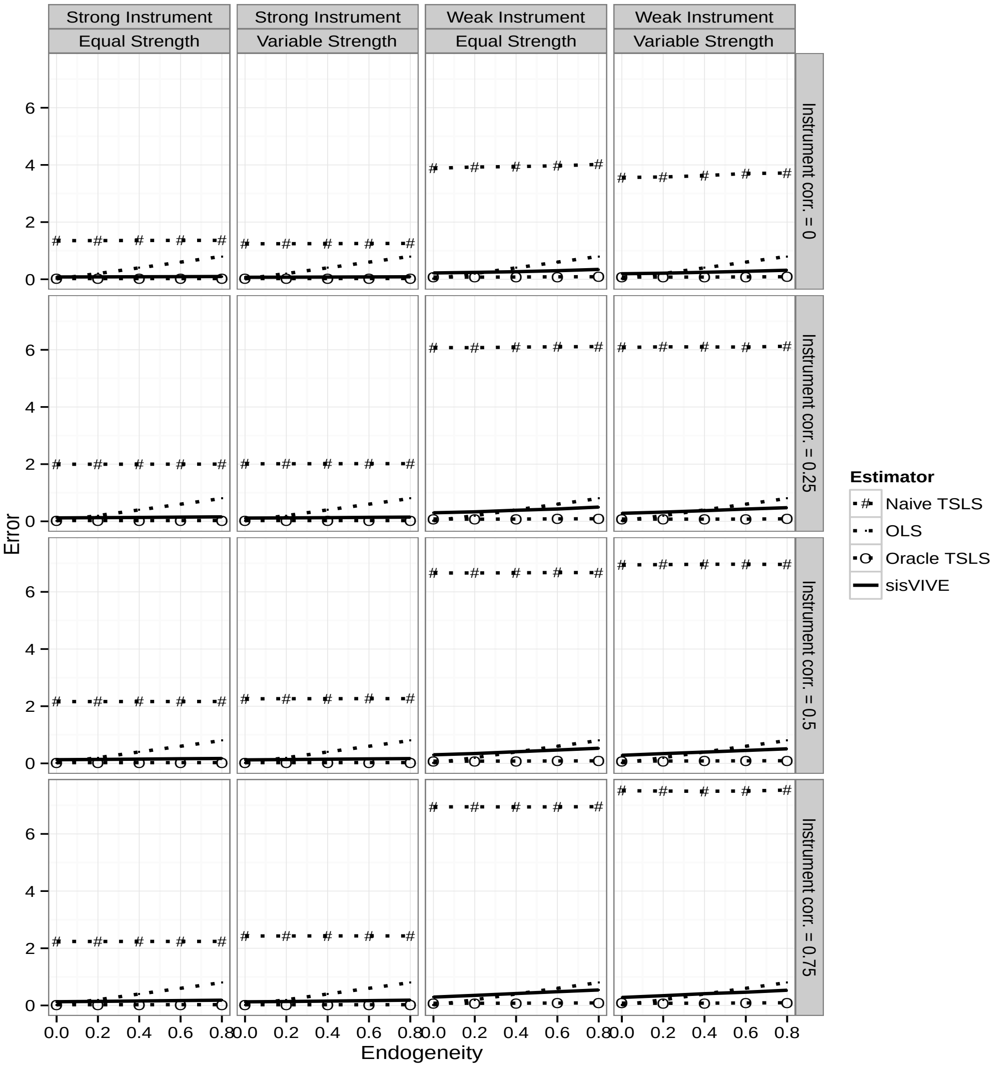

Figure 6: sisVIVE Simulation Study of β^* With Different Endogeneity and Where Correlation Only Exists Between Valid and Invalid IVs (Setup 2). Each line represents the median absolute estimation error ($|\beta^* - \hat{\beta}|$) after 500 simulations. We fix the number of invalid instruments to $s = 3$. 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 $\boldsymbol{\gamma}^*$ while holding the absolute strength (i.e. concentration parameter) fixed. Each row corresponds to the maximum correlation between instruments, but correlation only exists between valid and invalid instruments.

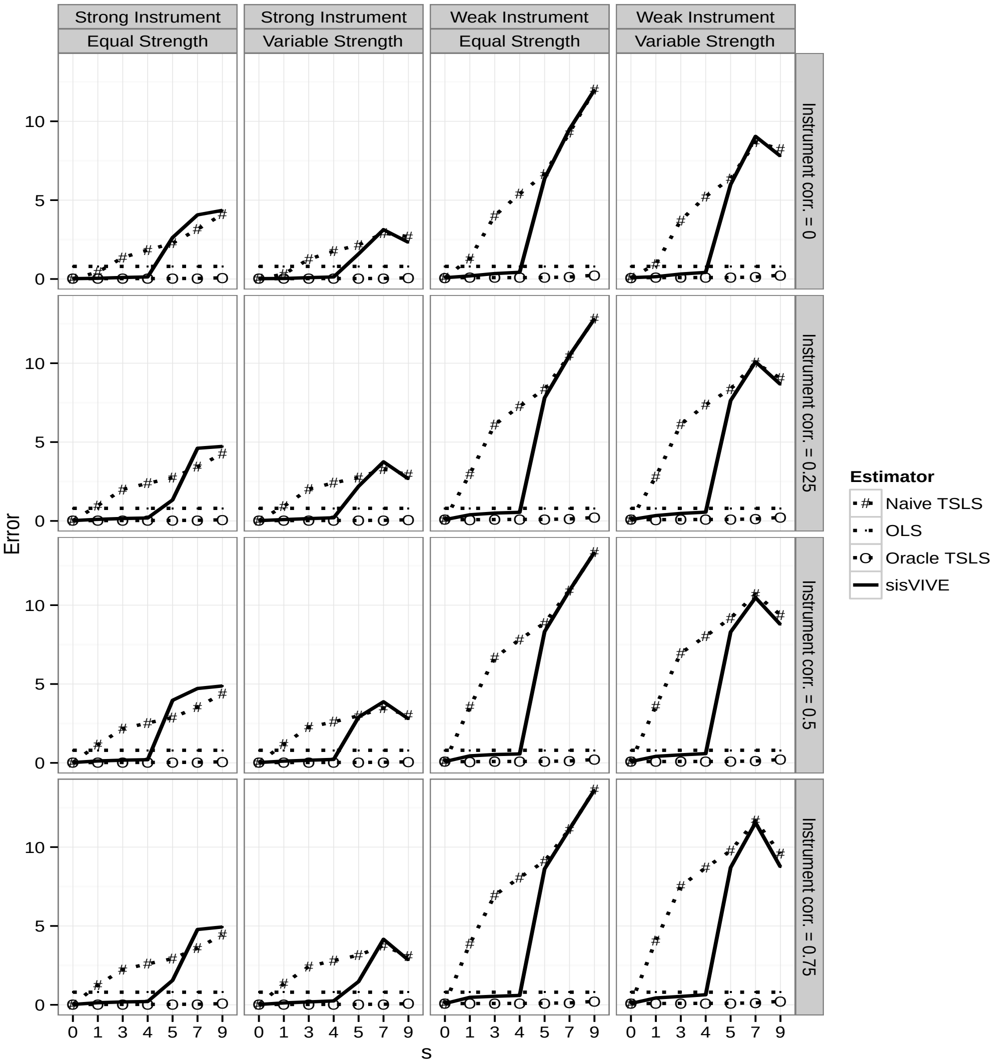

Figure 7: sisVIVE Simulation Study of β^* With Different Number of Invalid IVs and Where Correlation Only Exists Between Valid and and Invalid IVs (Setup 2). There are ten ($L = 10$) instruments. Each line represents the median absolute estimation error ($|\beta^* - \hat{\beta}|$) after 500 simulations. 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 $\boldsymbol{\gamma}^*$ while holding the absolute strength fixed. Each row corresponds to the maximum correlation between instruments, but correlation only exists between valid and invalid instruments.

the valid instruments and the case where the valid instruments are "stronger" than the invalid instruments, i.e. where the strength of IVs are non-uniform between valid and invalid IVs.

Figure 8 varies endogeneity, but fixes $s = 3$. In this case, sisVIVE performs as well as the oracle for strong instruments. For weak instruments, sisVIVE does better when the valid instruments are stronger than the invalid instruments (i.e. "Stronger Valid") than when the invalid instruments are stronger than the valid instruments (i.e. "Stronger Invalid"). Under any strength, sisVIVE does much better than the next best alternative, naive two stage least squares.

Figure 9 varies s, but fixes endogeneity to 0.8. In this case, sisVIVE deviates from the oracle at $s = 4$ for the case when the invalid instruments are stronger than the valid instruments (i.e. "Stronger Invalid") and at $s = 7$ for the case when the valid instruments are stronger than the invalid instruments (i.e. "Stronger Valid"). When sisVIVE deviates from oracle TSLS, sisVIVE's performance is no worse than naive TSLS.

In addition, for each of the simulation setups in this section (16 in total, each corresponding to 16 subfigures in Figures 8 and 9), we compute ρ and μ that appear in Corollary 2.2, similar to what we did in Section 2.4.2. Table 3 and 4 show the results. The column and row labels in the two tables are identical to those found in Section 2.4.2, except the new headings "Stronger Invalid" and "Stronger Valid."

Instrument Corr. (μ)	Strong Instrument, Stronger Invalid	Strong Instrument, Stronger Valid	Weak Instrument, Stronger Invalid	Weak Instrument, Stronger Valid
0	0.41	0.33	0.28	0.18
0.25	0.60	0.54	0.47	0.33
0.5	0.75	0.71	0.64	0.49
0.75	0.88	0.86	0.81	0.70

Table 3: Values of ρ in Corollary 2.2 for sisVIVE Simulation Study (Setup 3)

The simulation study in this section showed that in vast majority of cases, sisVIVE estimates

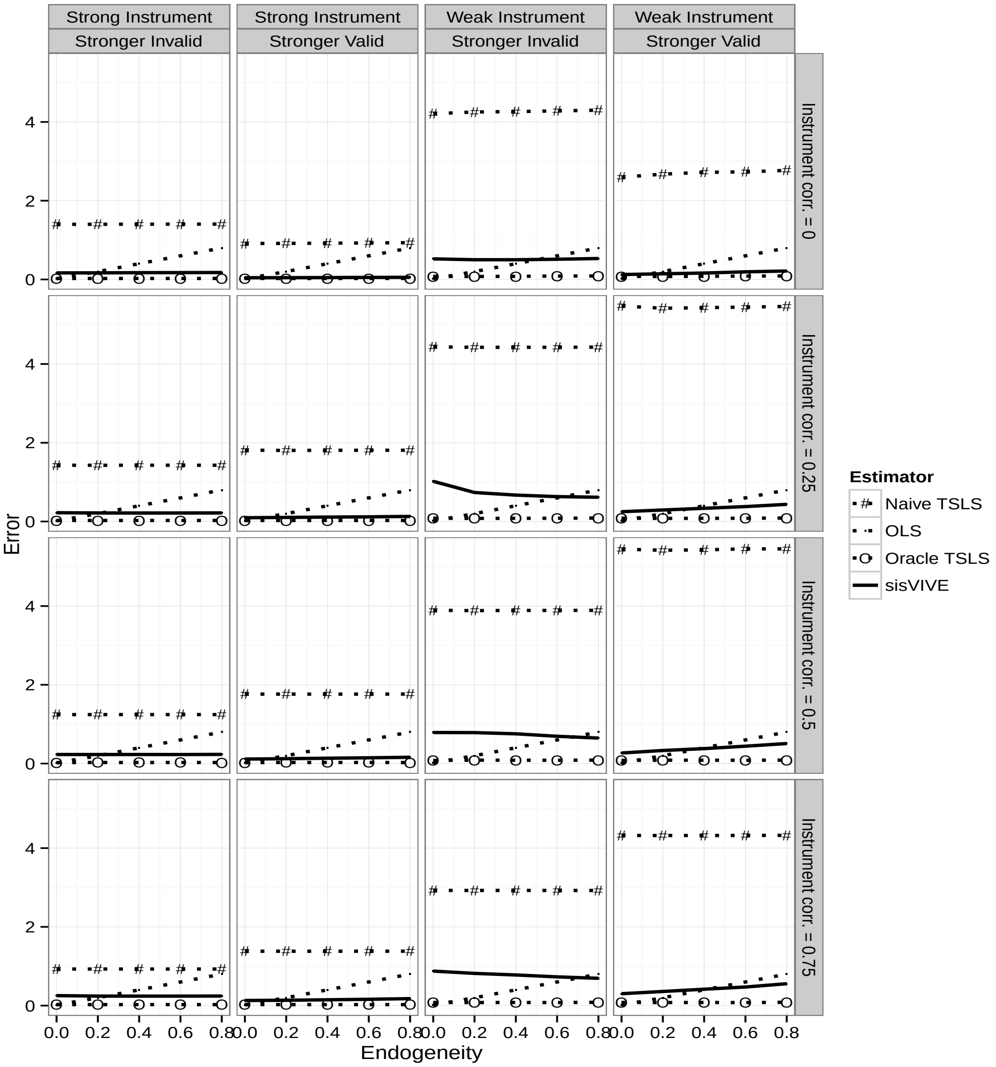

Figure 8: sisVIVE Simulation Study of β^* With Different Endogeneity and Where Correlation Exists Between All IVs (Setup 3). We also vary the instrument strength of valid and invalid instruments. There are ten ($L = 10$) instruments. Each line represents the median absolute estimation error ($|\beta^* - \hat{\beta}|$) after 500 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 strengths for valid and invalid instruments, "Stronger Invalid" and "Stronger Valid", determined by varying $\boldsymbol{\gamma}^*$ while holding the absolute strength fixed. Each row corresponds to the maximum correlation between instruments.

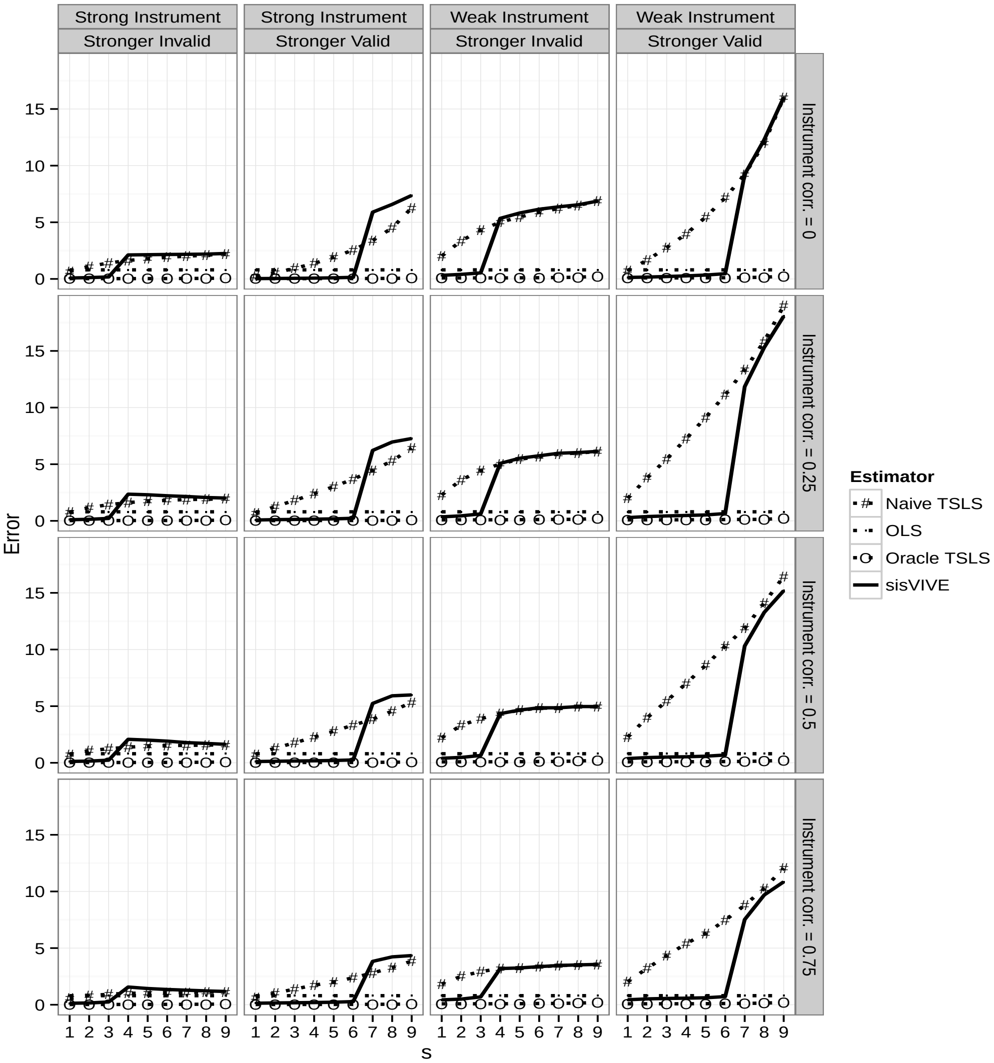

Figure 9: sisVIVE Simulation Study of β^* With Different Number of Invalid IVs and Where Correlation Exists Between All IVs (Setup 3). We also vary the instrument strength of valid and invalid instruments. There are ten ($L = 10$) instruments. Each line represents the median absolute estimation error ($|\beta^* - \hat{\beta}|$) after 500 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 strengths for valid and invalid instruments, "Stronger Invalid" and "Stronger Valid", determined by varying $\boldsymbol{\gamma}^*$ while holding the absolute strength fixed. Each row corresponds to the maximum correlation between instruments.

Instrument Corr. (μ)	Strong Instrument, Stronger Invalid	Strong Instrument, Stronger Valid	Weak Instrument, Stronger Invalid	Weak Instrument, Stronger Valid
0	0.60	0.90	1.27	3.02
0.25	0.28	0.33	0.33	0.33
0.5	0.17	0.17	0.17	0.17
0.75	0.11	0.11	0.11	0.11

Table 4: Condition on s in Corollary 2.2 for sisVIVE Simulation Study (Setup 3)

the causal effect of interest better than the next best alternative, naive TSLS, and in many cases, sisSIVE's performance is similar to the oracle. However, when the invalid instruments are stronger than the valid instruments (i.e. "Stronger Invalid"), sisVIVE's performance does not do as well as the oracle, even though by the identification result in Corollary 2.1, at $s = 4$, identification is guaranteed. The degradation in performance of sisVIVE may be due to a number of reasons. It may follow from the fact that the condition in Corollary 2.2 are not met since Table 4 shows that in the "Stronger Invalid" case, s has to be less than 1 or 2. It may be that we chose a bad tuning parameter λ (see Sections 2.4.7 and 2.4.8). A closer analysis of this particular case is a topic for future research. Regardless, even when sisVIVE's performance degrades compared to the oracle, sisVIVE does no worse than the next best alternative, naive TSLS.

2.4.5. Simulation Setup 4: $L = 10$, Pairwise Correlation Between Subsets of IVs and Non-Uniform IV Strength Between Valid and Invalid IVs

The simulation setup is identical to Section 2.4.3, except we consider the two types of strengths considered in Section 2.4.4.

Figures 10 and 11 vary endogeneity, but fix $s = 3$. Figure 10 is the case where there is no correlation between an invalid instrument and valid instrument and Figure 11 is the case where there is no correlation among invalid instruments and among valid instruments. In both scenarios, the behavior of the simulation is similar to Section 2.4.4. sisVIVE performs as well as the oracle for strong instruments. For weak instruments, sisVIVE does better

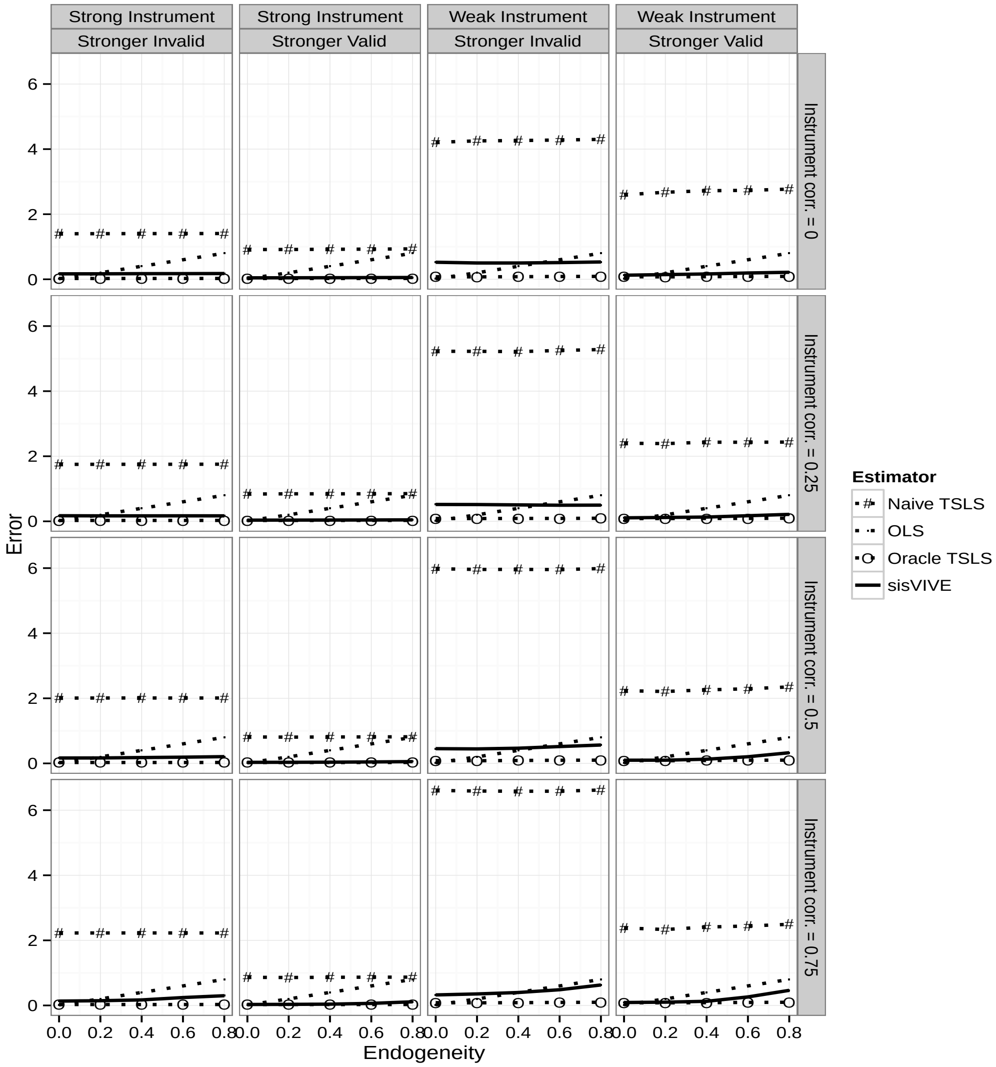

Figure 10: sisVIVE Simulation Study of β^* With Different Endogeneity and Where Correlation Only Exists Within Valid and Invalid IVs (Setup 4). We also vary the instrument strength of valid and invalid instruments. There are ten ($L = 10$) instruments. Each line represents the median absolute estimation error ($|\beta^* - \hat{\beta}|$) after 500 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 strengths for valid and invalid instruments, "Stronger Invalid" and "Stronger Valid", determined by varying $\boldsymbol{\gamma}^*$ while holding the absolute strength fixed. Each row corresponds to the maximum correlation between instruments, but correlation only exists within valid and invalid instruments.

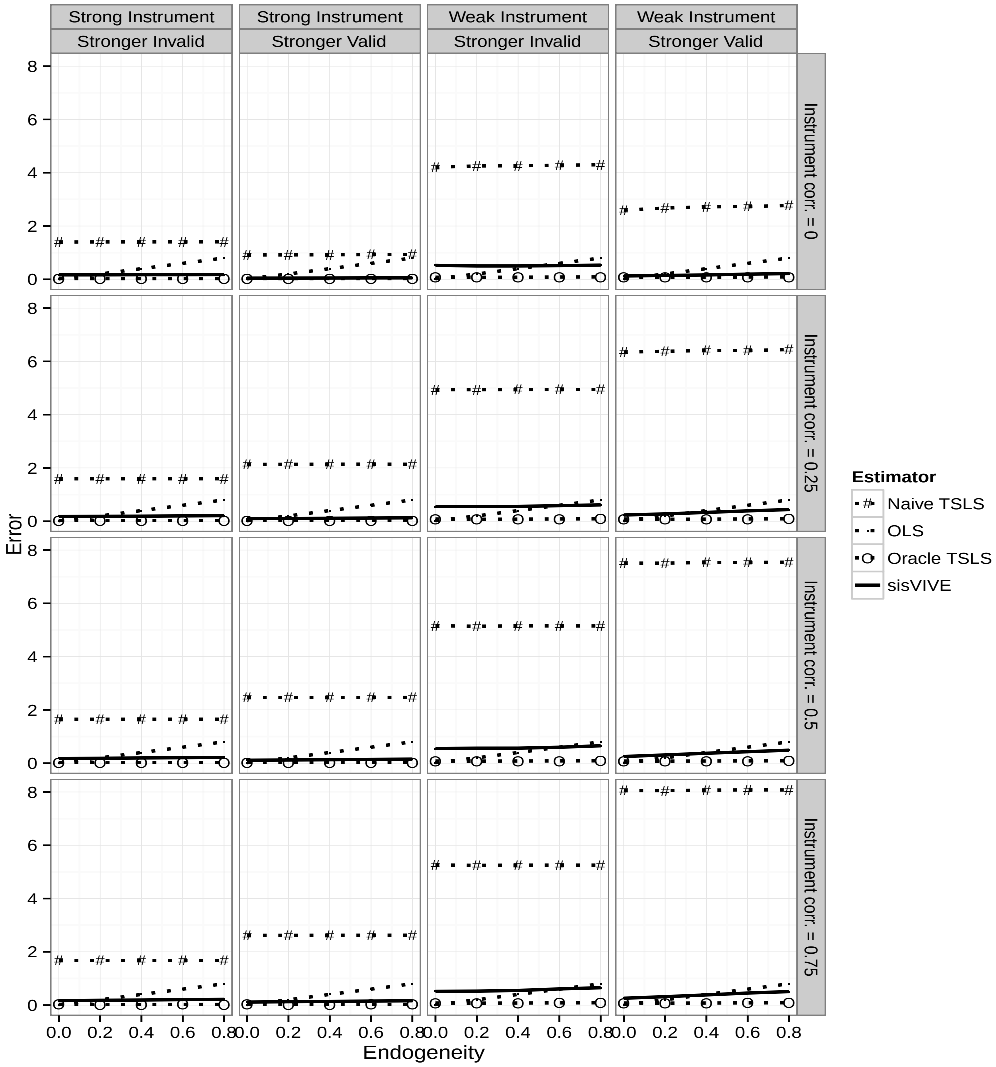

Figure 11: sisVIVE Simulation Study of β^* With Different Endogeneity and Where Correlation Only Exists Between Valid and Invalid IVs (Setup 4). We also vary the instrument strength of valid and invalid instruments. There are ten ($L = 10$) instruments. Each line represents the median absolute estimation error ($|\beta^* - \hat{\beta}|$) after 500 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 strengths for valid and invalid instruments, "Stronger Invalid" and "Stronger Valid", determined by varying $\boldsymbol{\gamma}^*$ while holding the absolute strength fixed. Each row corresponds to the maximum correlation between instruments, but correlation only exists between valid and invalid instruments.

when the valid instruments are stronger than the invalid instruments (i.e. "Stronger Valid") than when the invalid instruments are stronger than the valid instruments (i.e. "Stronger Invalid"). In both the strong and weak cases, sisVIVE does much better than the next best alternative, naive TSLS.

Figures 12 and 13 vary s, but fix endogeneity to 0.8. Figure 12 is the case where there is no correlation between an invalid instrument and valid instrument and Figure 13 is the case where there is no correlation among invalid instruments and among valid instruments. In both scenarios, the behavior of the simulation is similar to Section 2.4.4, sisVIVE deviates from the oracle at $s = 4$ for the case when the invalid instruments are stronger than the valid instruments (i.e. "Stronger Invalid") and at $s = 7$ for the case when the valid instruments are stronger than the invalid instruments (i.e. "Stronger Valid"). When sisVIVE deviates from oracle TSLS, sisVIVE's performance is no worse than naive TSLS.

Overall, similar to what we saw in Section 2.4.3, the three correlation structures produce similar results with respect to estimation error, $|\beta^* - \hat{\beta}_\lambda|$.

2.4.6. Simulation Setup 5: $L = 100$*, Pairwise Correlation Between All IVs and Uniform IV Strength Between Valid and Invalid IVs*

The simulation setup is identical to Section 2.4.2, except we increase the number of instruments to $L = 100$. We only consider instruments where all the pairwise correlation is set to μ since results from Sections 2.4.3 and 2.4.5 showed the three different structures of instrumental correlation produced similar results in terms of sisVIVE's estimation performance.

We note that in Mendelian randomization settings, it is rare to have 100 potential genetic instruments since all 100 of the genetic instruments must affect the exposure (see Sections 2.1 and 2.3.1 for more details). Usually, the number of potential instruments is far less than 100 (see Section 2.1 for some example MR studies). However, for completeness, we demonstrate sisVIVE's performance when $L = 100$ potential instruments are present.

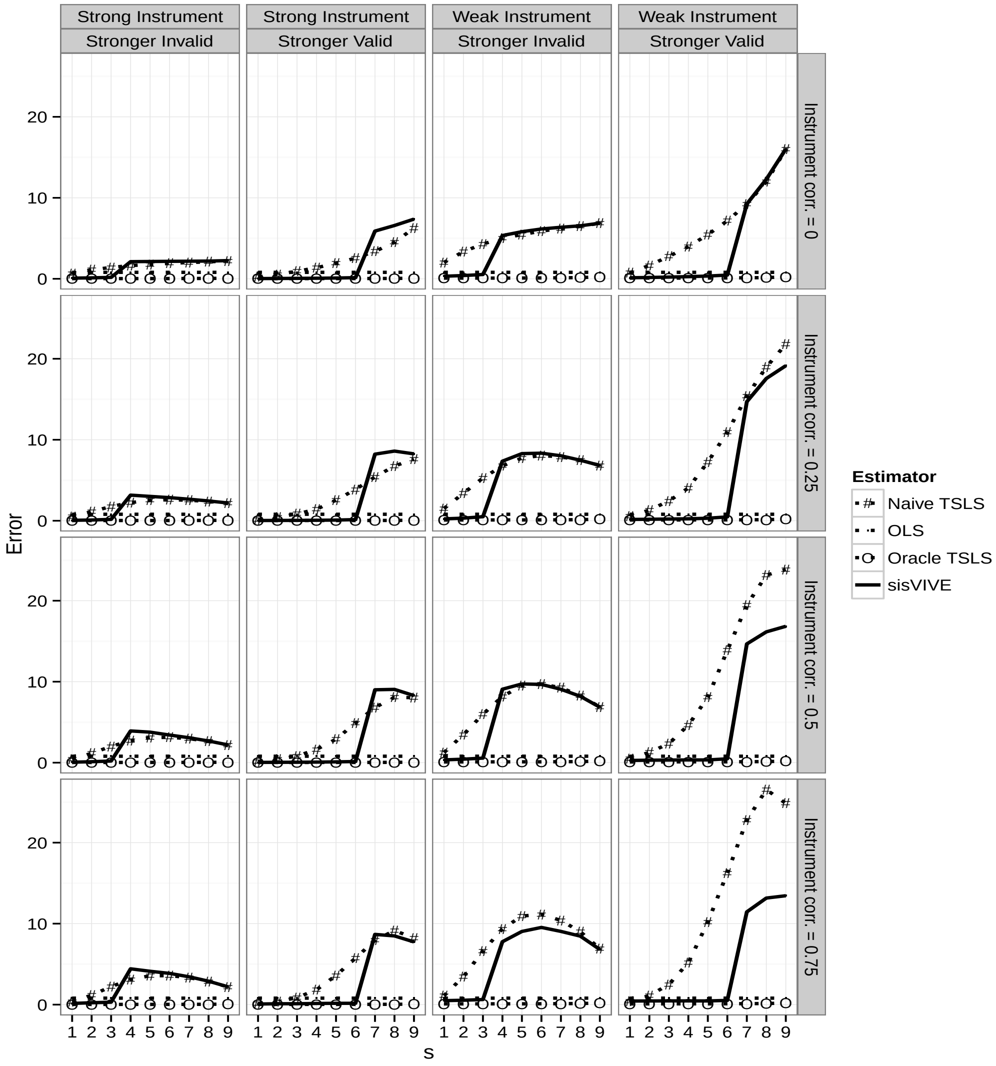

Figure 12: sisVIVE Simulation Study of β^* With Different Number of Invalid IVs and Where Correlation Only Exists Within Valid and and Invalid IVs (Setup 4). We also vary the instrument strength of valid and invalid instruments. There are ten ($L = 10$) instruments. Each line represents the median absolute estimation error ($|\beta^* - \hat{\beta}|$) after 500 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 strengths for valid and invalid instruments, "Stronger Invalid" and "Stronger Valid", determined by varying $\boldsymbol{\gamma}^*$ while holding the absolute strength fixed. Each row corresponds to the maximum correlation between instruments, but correlation only exists within valid and invalid instruments.

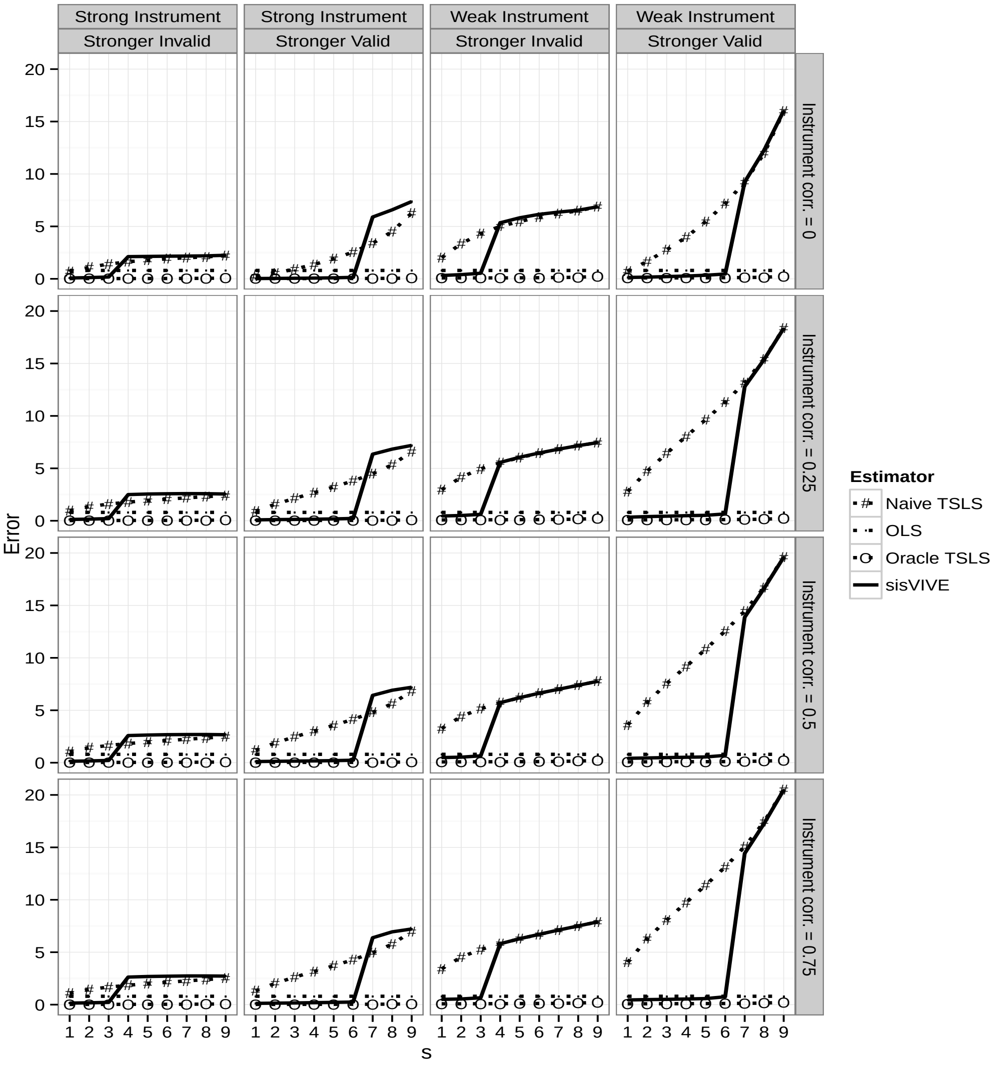

Figure 13: sisVIVE Simulation Study of β^* With Different Number of Invalid IVs and Where Correlation Only Exists Between Valid and Invalid IVs (Setup 4). We also vary the instrument strength of valid and invalid instruments. There are ten ($L = 10$) instruments. Each line represents median absolute estimation error ($|\beta^* - \hat{\beta}|$) after 500 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 strengths for valid and invalid instruments, "Stronger Invalid" and "Stronger Valid", determined by varying $\boldsymbol{\gamma}^*$ while holding the absolute strength fixed. Each row corresponds to the maximum correlation between instruments, but correlation only exists between valid and invalid instruments.

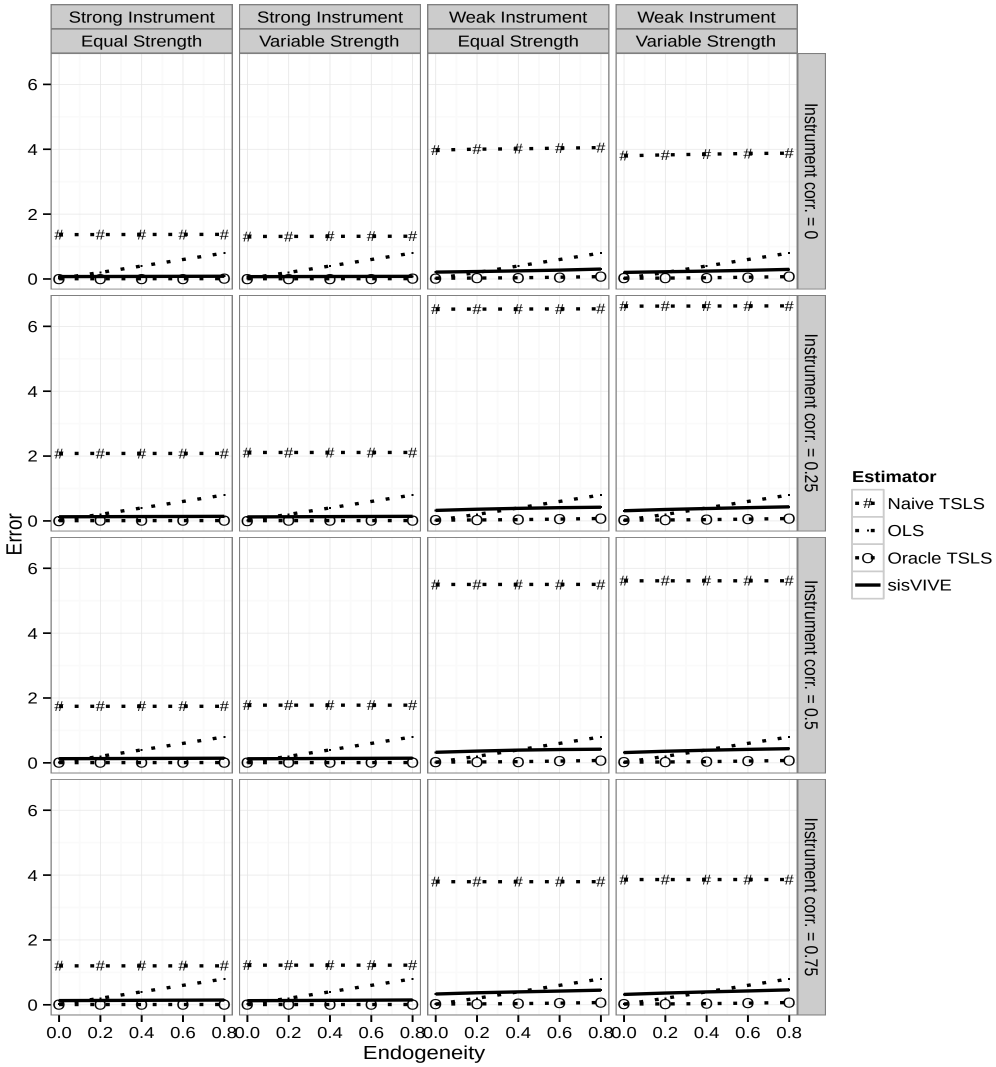

Figure 14: sisVIVE Simulation Study of β^* With Different Endogeneity and Where Correlation Exists Between All IVs (Setup 5). There are 100 ($L = 100$) instruments. Each line represents the median absolute estimation error ($|\beta^* - \hat{\beta}|$) after 500 simulations. We fix the number of invalid instruments to $s = 30$. 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 $\boldsymbol{\gamma}^*$ while holding the absolute strength (i.e. concentration parameter) fixed. Each row corresponds to the maximum correlation between instruments.

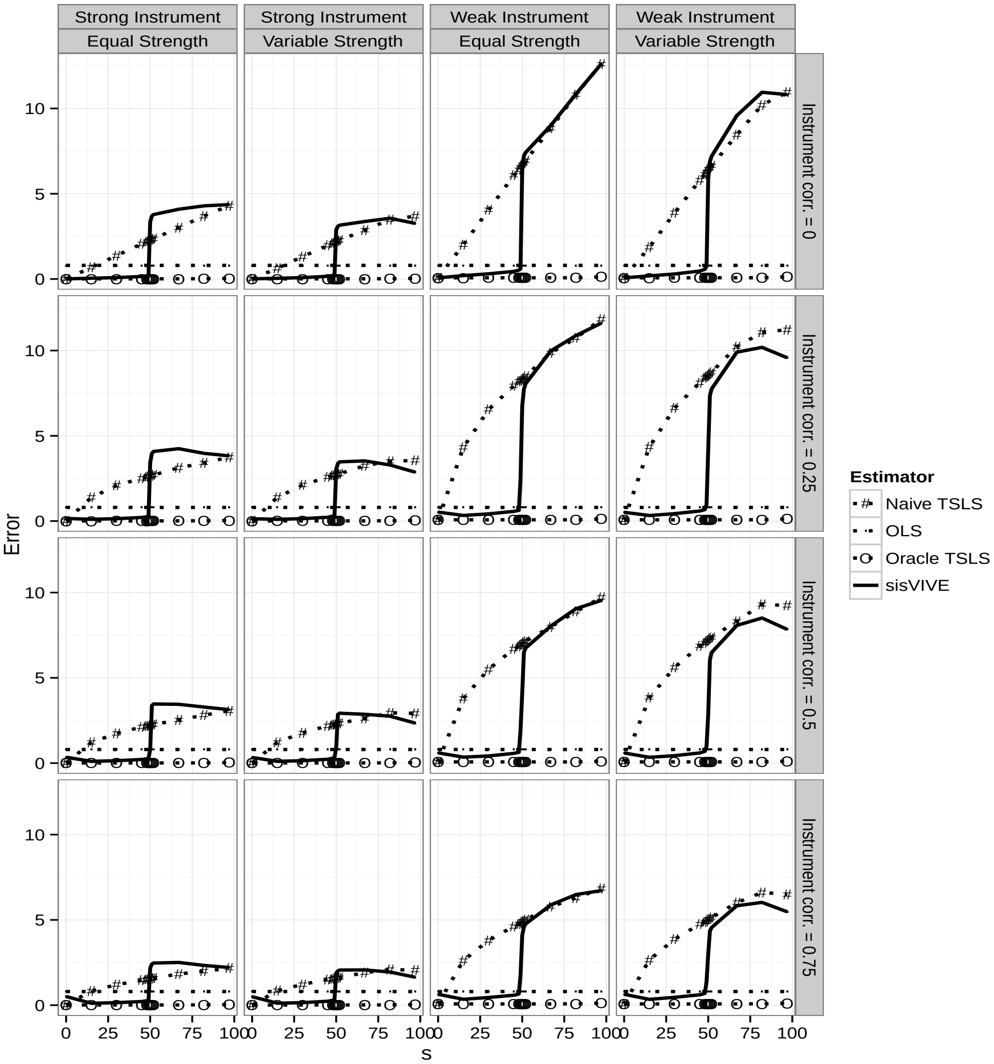

Figure 15: sisVIVE Simulation Study of β^* With Different Number of Invalid IVs and Where Correlation Exists Between All IVs (Setup 5). There are 100 ($L = 100$) instruments. Each line represents the median absolute estimation error ($|\beta^* - \hat{\beta}|$) after 500 simulations. We fix the endogeneity $\sigma^*_{\epsilon\xi}$ 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 $\boldsymbol{\gamma}^*$ while holding the absolute strength fixed. Each row corresponds to maximum correlation between instruments.

Figures 14 and 15 represent the results of estimation performance of β^* over 500 simulations. The behavior of all four estimators is similar to Figures 2 and 3 in Section 2.4.2. For example, when we vary endogeneity (Figure 14), sisVIVE tends to perform slightly worse when the overall strength of the instruments is weak. Also, when the number of invalid instruments, s, is varied (Figure 15), sisVIVE has a sharp peak at $s = 50$, similar to the sharp peak at $s = 5$ in Figure 3.

We also compute the ρ and μ resulting from this simulation study. Tables 5 and 6 show the results. Notice that again, Corollary 2.2, while interpretable, tends to give very stringent conditions on s in Table 6.

Instrument Corr. (μ)	Strong Instrument, Equal Strength	Strong Instrument, Variable Strength	Weak Instrument, Equal Strength	Weak Instrument, Variable Strength
0	0.15	0.17	0.16	0.17
0.25	0.54	0.54	0.53	0.53
0.5	0.73	0.73	0.53	0.73
0.75	0.87	0.87	0.88	0.87

Table 5: Values of ρ in Corollary 2.2 for sisVIVE Simulation Study (Setup 5)

Instrument Corr. (μ)	Strong Instrument, Equal Strength	Strong Instrument, Variable Strength	Weak Instrument, Equal Strength	Weak Instrument, Variable Strength
0	4.2	3.3	4.0	3.4
0.25	0.33	0.33	0.33	0.33
0.5	0.17	0.17	0.17	0.17
0.75	0.11	0.11	0.11	0.11

Table 6: Condition on s in Corollary 2.2 for sisVIVE Simulation Study (Setup 5)

Overall, the simulation study suggests that sisVIVE does scale as L increases and that its performance at large values of L is similar to its performance at smaller values of L, such as $L = 10$.

2.4.7. Measuring the Performance of sisVIVE's Estimation of $\boldsymbol{\alpha}^*$

This simulation setup is identical to Sections 2.4.2, 2.4.3, 2.4.4, 2.4.5, and 2.4.6, except we examine sisVIVE's estimation performance of $\boldsymbol{\alpha}^*$ instead of the estimation performance on β^*. As we noted before, in Mendelian randomization, the target of estimation is β^*, the causal effect of the exposure on the outcome, and our procedure, sisVIVE, was designed to estimate β^*. However, in the process of estimating β^*, sisVIVE does produce an estimate for $\boldsymbol{\alpha}^*$ and we explore the relationship between this intermediate estimate for $\boldsymbol{\alpha}^*$, $\hat{\boldsymbol{\alpha}}_\lambda$, and our desired estimate for β^*, $\hat{\beta}_\lambda$.

To evaluate the estimate $\hat{\boldsymbol{\alpha}}_\lambda$, we consider two metrics of error, (a) the proportion of correctly selected valid instruments and (b) the proportion of correctly selected invalid instruments. To illustrate the two proportion-based error metrics, consider the following numerical example. Suppose there are $L = 10$ instruments of which the first three instruments are invalid, i.e. $\alpha_j^* \neq 0$ for $j = 1, 2, 3$, and the last seven instruments are valid, i.e. $\alpha_j^* = 0$ for $j = 4, 5, \ldots, 10$. If sisVIVE estimates the first two instruments to be invalid, i.e. $\hat{\alpha}_j \neq 0$ for $j = 1, 2$, and the last eight to be valid, i.e. $\hat{\alpha}_j = 0$ for $j = 3, 4, \ldots, 10$, the proportion of correctly selected valid instruments is $7/7 = 1$ and sisVIVE makes no error in choosing the valid instruments. However, the proportion of correctly selected invalid instruments is $2/3$ and sisVIVE makes an error in choosing the invalid instruments.

First, we look at simulation setups in Sections 2.4.2 and 2.4.3. When we vary endogeneity but fix the number of invalid instruments to $s = 3$ (Figures 16, 18, and 20, each figure representing different correlation structures between IVs), the proportion of correctly selected invalid instruments is 1 and sisVIVE never makes a mistake in selecting the invalid instruments. However, sisVIVE does make mistakes in selecting the valid instruments as the proportion of correctly selected valid instruments is mostly below 1. Also, depending on the correlation structure between instruments, we get different behaviors for the proportion of correctly selected valid instruments. For example, when every pair of instruments has non-zero pairwise correlation (Figure 16), the proportion of correctly selected valid instruments remains roughly the same for different values of endogeneity. When there is only pairwise correlation within valid and invalid instruments (Figure 18), the proportion of correctly selected valid instruments decreases as endogeneity increases, most notably among weak instruments. Finally, when there is only pairwise correlation between valid and invalid instruments (Figure 20), the proportion of correctly selected valid instruments increases as endogeneity increases. Despite these differences in the proportion of correctly selected valid instruments between different correlation structures, as the simulations in Sections 2.4.3 and 2.4.5, sisVIVE's median absolute deviation from the truth, $|\hat{\beta}_\lambda - \beta^*|$, remains relatively small and constant for all values of the endogeneity, irrespective of the different correlation structures. Note that this constant behavior is also present in the proportion of correctly selected invalid instruments, which remains at 1 for all correlation structures. This suggests that there is a strong relationship between correctly selecting the invalid instruments and sisVIVE's median absolute deviation from β^* while there is at most a weak relationship between correctly selecting valid instruments and sisVIVE's median absolute deviation from β^*. In fact, it appears that correctly selecting invalid instruments is more important than valid instruments if a small median absolute deviation is desired.

When we vary the number of invalid instruments s, but fix the endogeneity (Figures 17, 19, and 21, each figure representing different correlation structures between IVs), the proportion of correctly selected invalid instrument decreases significantly at the $s = 5$ boundary, regardless of the correlation structure between instruments. For example, for strong instruments in the three figures, when $s < 5$, the proportion of correctly selected invalid instruments remain at 1. However, when $s \geq 5$, the proportion of correctly selected invalid instruments moves sharply away from 1. For weak instruments in the three figures, when $s < 5$, the proportion of correctly selected invalid instruments remains close to 1, although there is a slightly decrease in the proportion when s moves from $s = 3$ to $s = 4$ and when μ is away from zero. But, similar to the strong instruments, when $s \geq 5$, the proportion of correctly selected invalid instruments moves away from 1.

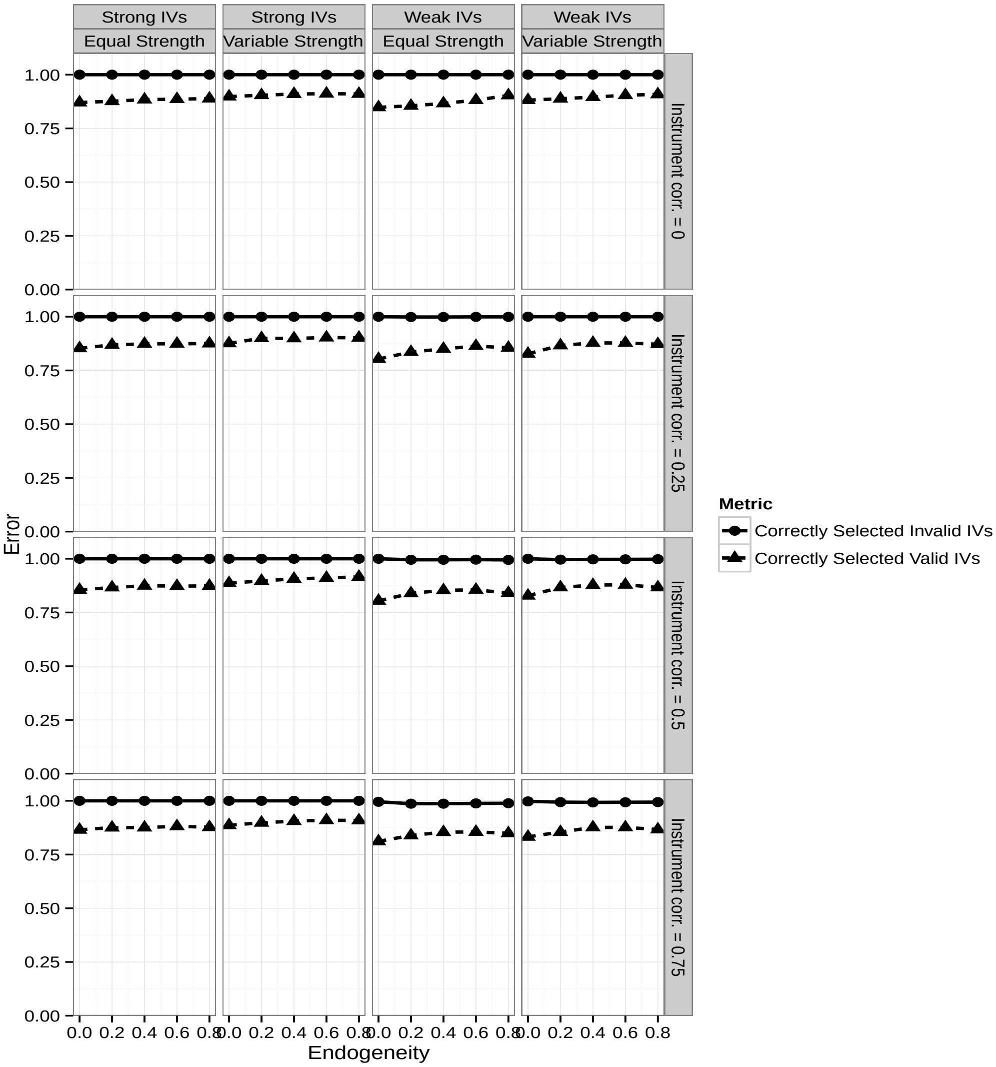

Figure 16: sisVIVE Simulation Study of $\boldsymbol{\alpha}^*$ With Different Endogeneity and Where Correlation Exists Between All IVs (Setup 1). There are ten ($L = 10$) instruments. Each line represents the average proportions of correctly selected valid instruments and correctly selected invalid instruments after 500 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 $\boldsymbol{\gamma}^*$ while holding the absolute strength (i.e. concentration parameter) fixed. Each row corresponds to the maximum correlation between all instruments.

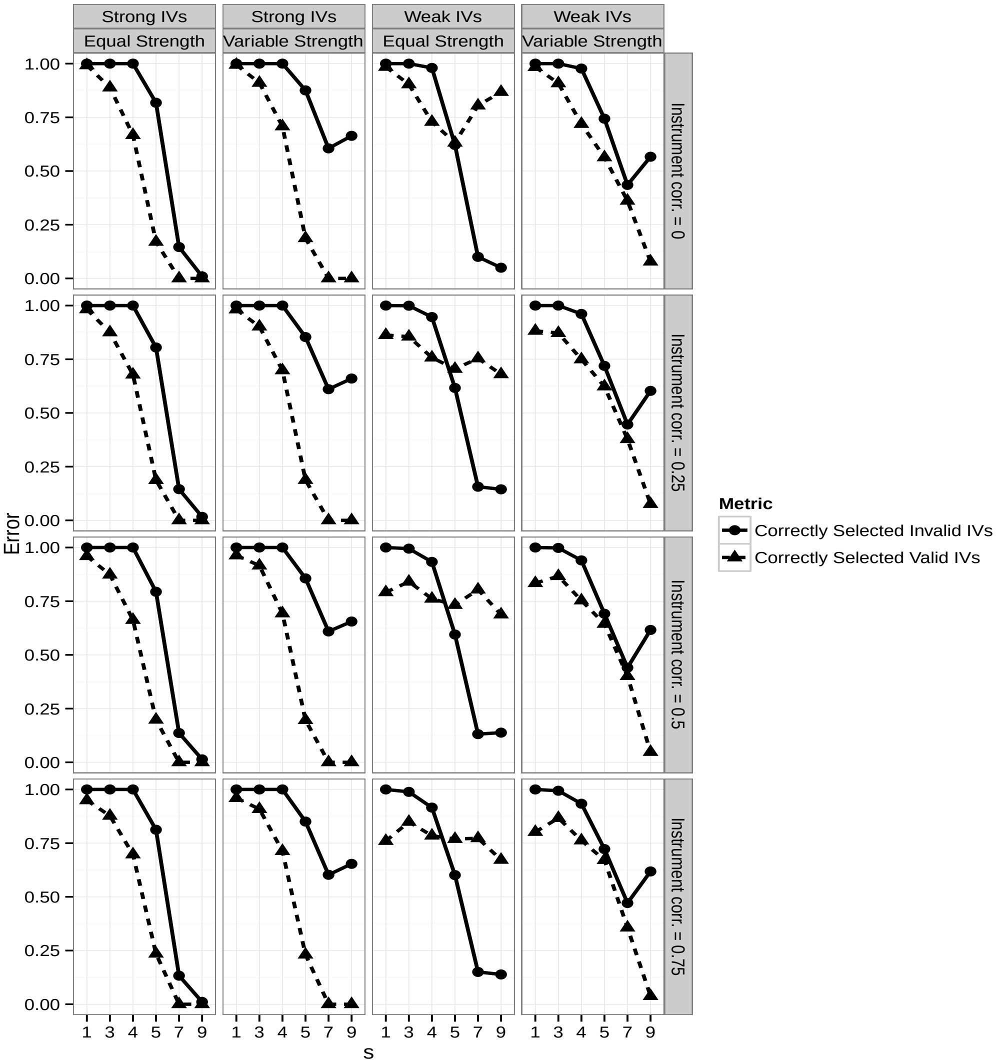

Figure 17: sisVIVE Simulation Study of $\boldsymbol{\alpha}^*$ With Different Number of Invalid IVs and Where Correlation Exists Between All IVs (Setup 1). There are ten ($L = 10$) instruments. Each line represents the average proportions of correctly selected valid instruments and correctly selected invalid instruments after 500 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 $\boldsymbol{\gamma}^*$ while holding the absolute strength fixed. Each row corresponds to the maximum correlation between all instruments.

The proportion of correctly selected valid instruments decreases steadily as s increases, regardless of the type of correlation structure between instruments. For strong instruments in the three figures, the decrease in the proportion of correctly selected valid instruments begins immediately after $s = 1$. For weak instruments in the three figures, there is considerable fluctuation of the proportion of correctly selected valid instruments. For example, for Figures 17 and Figures 19 under weak IVs, the proportion of correctly selected valid instruments generally decreases as s increase, with the notable exception in the first row, third column of both figures. But, for Figure 21 under weak IVs, the proportion of correctly selected valid instruments decreases when $s < 5$, but increases again after $s \geq 5$.

The behaviors of the proportions of correctly selected invalid and valid instruments from Figures 17, 19, and 21 reaffirm our observation that there is a strong association between the proportion of correctly selected invalid instruments and the median absolute deviation of $\hat{\beta}_\lambda$, $|\hat{\beta}_\lambda - \beta^*|$. In particular, from Figures 3, 5 and 7, when $s < 5$, sisVIVE's median absolute deviation is just as small as the oracle TSLS. However, when $s \geq 5$, sisVIVE's median absolute deviation is just as large as the naive TSLS. The proportion of correctly selected invalid instruments in Figures 17, 19, and 21 closely corresponds to this sharp change in behavior between $s < 5$ and $s \geq 5$. In contrast, the proportion of correctly selected valid instruments does not have this sharp behavior at $s = 5$ across all the figures.

Overall, in simulation setups in Sections 2.4.2 and 2.4.3, we find that for any type of correlation structure between instruments and different variations on endogeneity and s, sisVIVE deviates far from the truth if we incorrectly select the invalid instruments. Hence, it is much more important to correctly select invalid instruments at the expense of incorrectly selecting valid instruments for better estimation of β^*. This relationship makes sense since using invalid instruments creates bias whereas using at least one valid instrument and not using other valid instruments does not create bias, but just reduces efficiency. The relationship also suggests that when we choose the tuning parameter λ, which controls the number of non-zero $\hat{\boldsymbol{\alpha}}_\lambda$ and consequently, controls the proportion of correctly selected valid

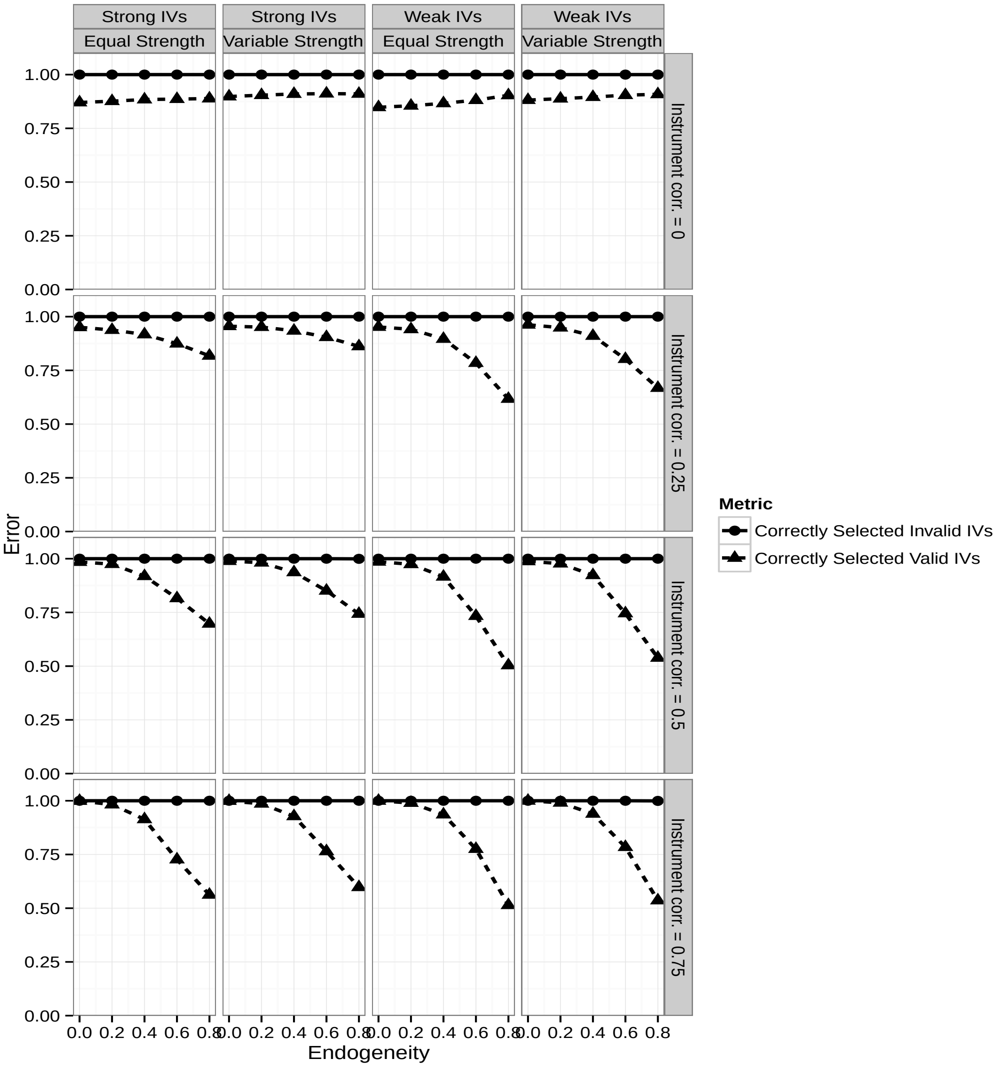

Figure 18: sisVIVE Simulation Study of $\boldsymbol{\alpha}^*$ with Different Endogeneity and Where Correlation Only Exists Within Valid and Invalid IVs (Setup 2). There are ten ($L = 10$) instruments. Each line represents the average proportions of correctly selected valid instruments and correctly selected invalid instruments after 500 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 $\boldsymbol{\gamma}^*$ while holding the absolute strength (i.e. concentration parameter) fixed. Each row corresponds to the maximum correlation between instruments, but correlation only exists within valid and invalid instruments.

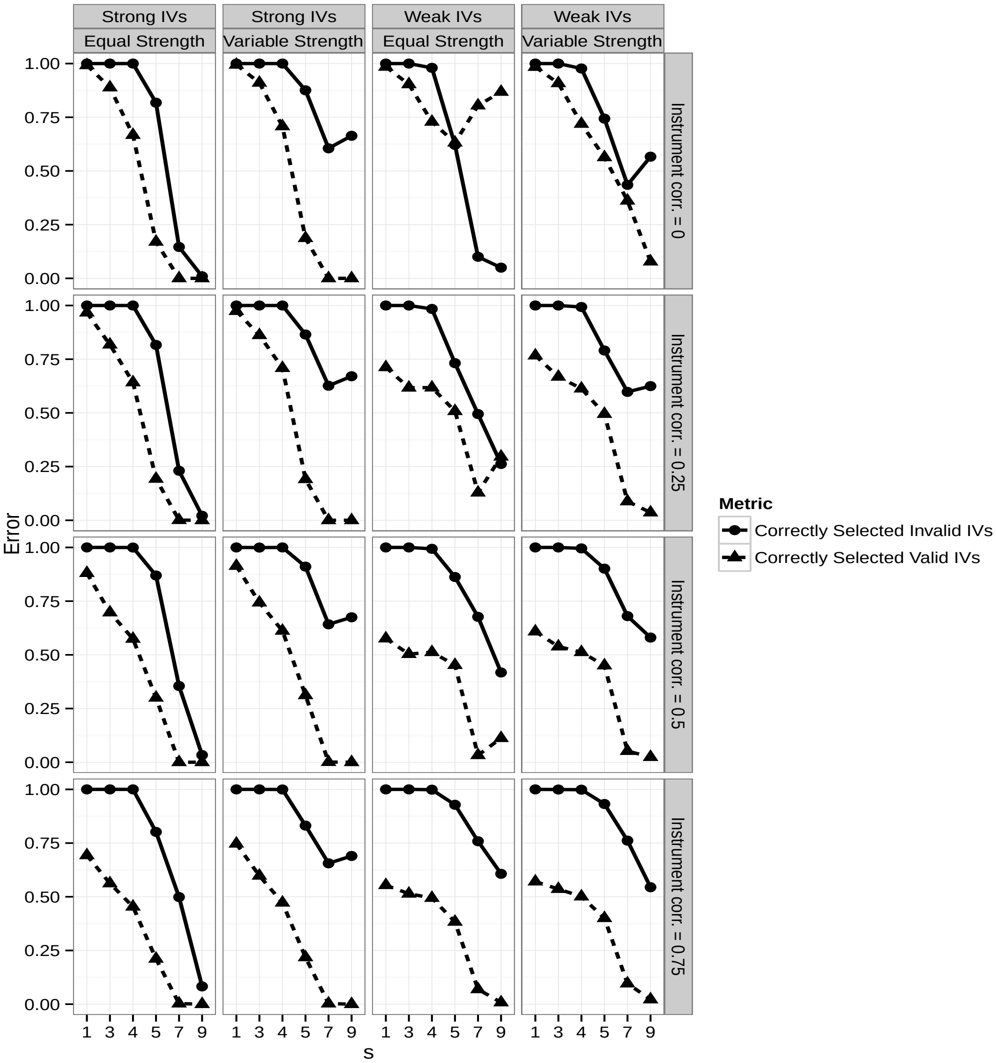

Figure 19: sisVIVE Simulation Study of $\boldsymbol{\alpha}^*$ with Different Number of Invalid IVs and Where Correlation Only Exists Within Valid and and Invalid IVs (Setup 2). There are ten ($L = 10$) instruments. Each line represents the average proportions of correctly selected valid instruments and correctly selected invalid instruments after 500 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 $\boldsymbol{\gamma}^*$ while holding the absolute strength fixed. Each row corresponds to the maximum correlation between instruments, but correlation only exists within valid and invalid instruments.

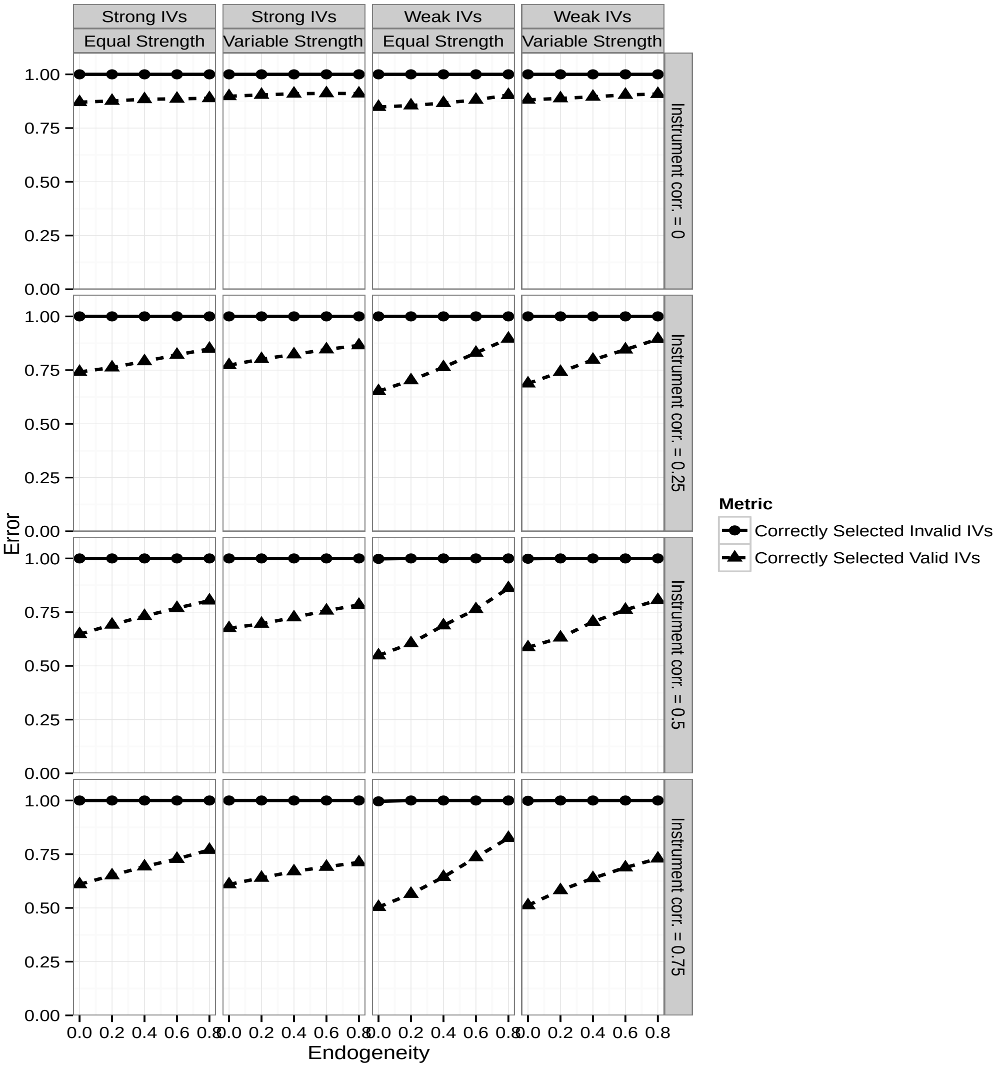

Figure 20: sisVIVE Simulation Study With Different Endogeneity and Where Correlation Only Exists Between Valid and Invalid IVs (Setup 2). Each line represents the average proportions of correctly selected valid instruments and correctly selected invalid instruments after 500 simulations. We fix the number 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 $\boldsymbol{\gamma}^*$ while holding the absolute strength (i.e. concentration parameter) fixed. Each row corresponds to the maximum correlation between instruments, but correlation only exists between valid and invalid instruments.

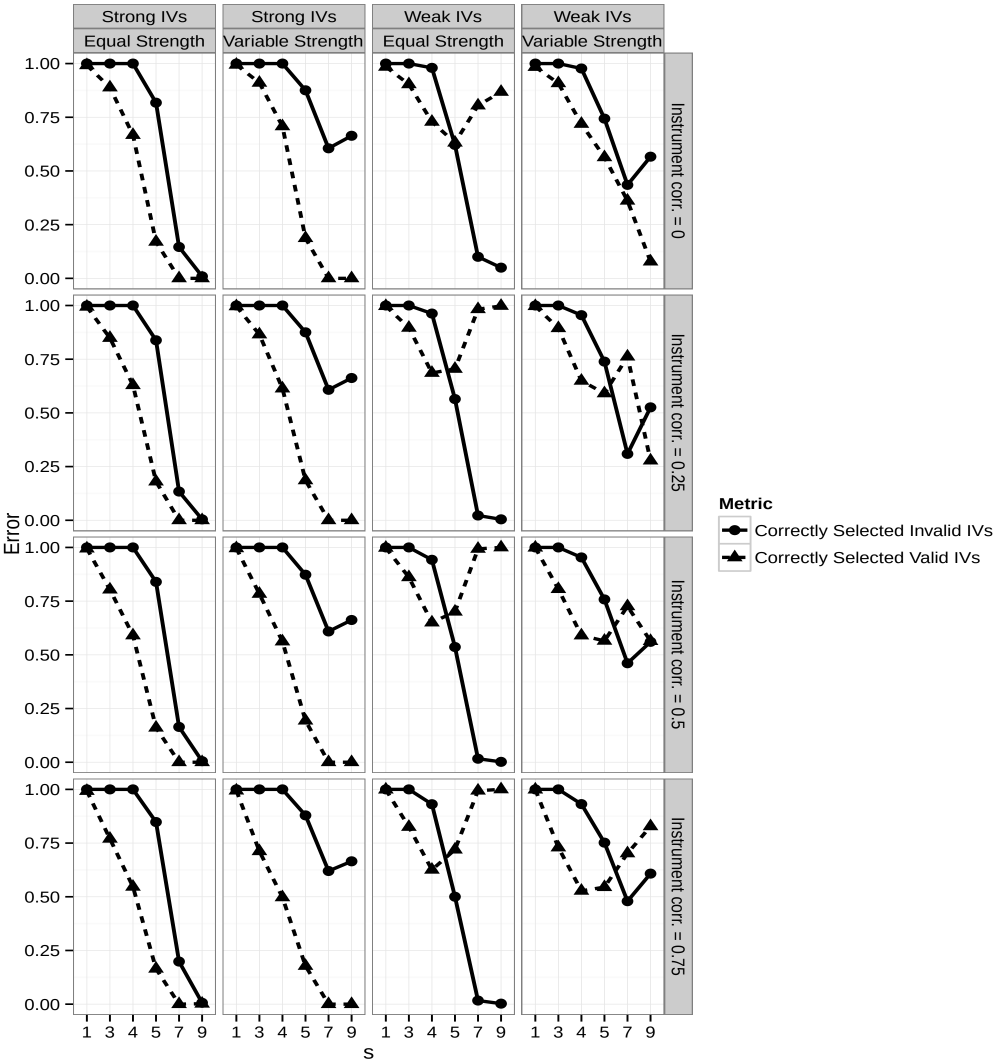

Figure 21: sisVIVE Simulation Study of $\boldsymbol{\alpha}^*$ With Different Number of Invalid IVs and Where Correlation Only Exists Between Valid and and Invalid IVs (Setup 2). There are ten ($L = 10$) instruments. Each line represents the average proportions of correctly selected valid instruments and correctly selected invalid instruments after 500 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 $\boldsymbol{\gamma}^*$ while holding the absolute strength fixed. Each row corresponds to the maximum correlation between instruments, but correlation only exists between valid and invalid instruments.

and invalid instruments, we should choose λ that correctly selects the invalid instruments, even if some valid instruments are selected as invalid. In particular, λ should generally be small so that there is less ℓ_1 penalty on $\|\boldsymbol{\alpha}\|_1$, but not too small so that the penalty has no effect. As a result, few elements of $\hat{\boldsymbol{\alpha}}_\lambda$ will be zero and more instruments will be selected as invalid. We discuss the choice of λ in more detail in Section 2.4.8.

Second, we look at simulation setups in Section 2.4.4. Figures 22 and 23 represent the case where we vary endogeneity and s, respectively. The behavior of the two curves are similar to what we observed before. That is, whenever sisVIVE performs badly in estimating β^*, there is a large decrease in the proportion of correctly selected invalid instruments. Also, there is no relationship between sisVIVE's median absolute bias of $\hat{\beta}_\lambda$ and the proportion of correctly selected valid instruments. For example, when we vary endogeneity (Figure 22), the proportion of correctly selected invalid instruments remain at 1 except when the overall strength of the instruments is weak and the invalid instruments are stronger than the valid instruments (i.e. "Stronger Invalid"). Regardless, in all cases, a smaller median absolute deviation in Figure 8 corresponds with having a high proportion of correctly selected invalid instruments in Figure 22. In contrast, the proportion of correctly selected valid instruments remains below 1 if the invalid instruments are stronger than the valid instruments (i.e. "Stronger Invalid") and close to 1 if the valid instruments are stronger than the invalid instruments (i.e. "Stronger Valid"). Furthermore, there does not seem to be any relationship between the proportion of correctly selected valid instruments and the estimation error of β^* in Figure 8.

Similarly, when we vary s (Figure 23) and we are under the case where the invalid instruments are stronger than the valid instruments (i.e. "Stronger Invalid"), the proportion of correctly selected invalid instruments move away from 1 at $s = 4$ when the overall strength of the instruments is strong and at $s = 3$ when the overall strength of the instruments is weak. When the valid instruments are stronger than the invalid instruments (i.e. "Stronger Valid"), the proportion of correctly selected invalid instruments move away from 1 at $s = 7$

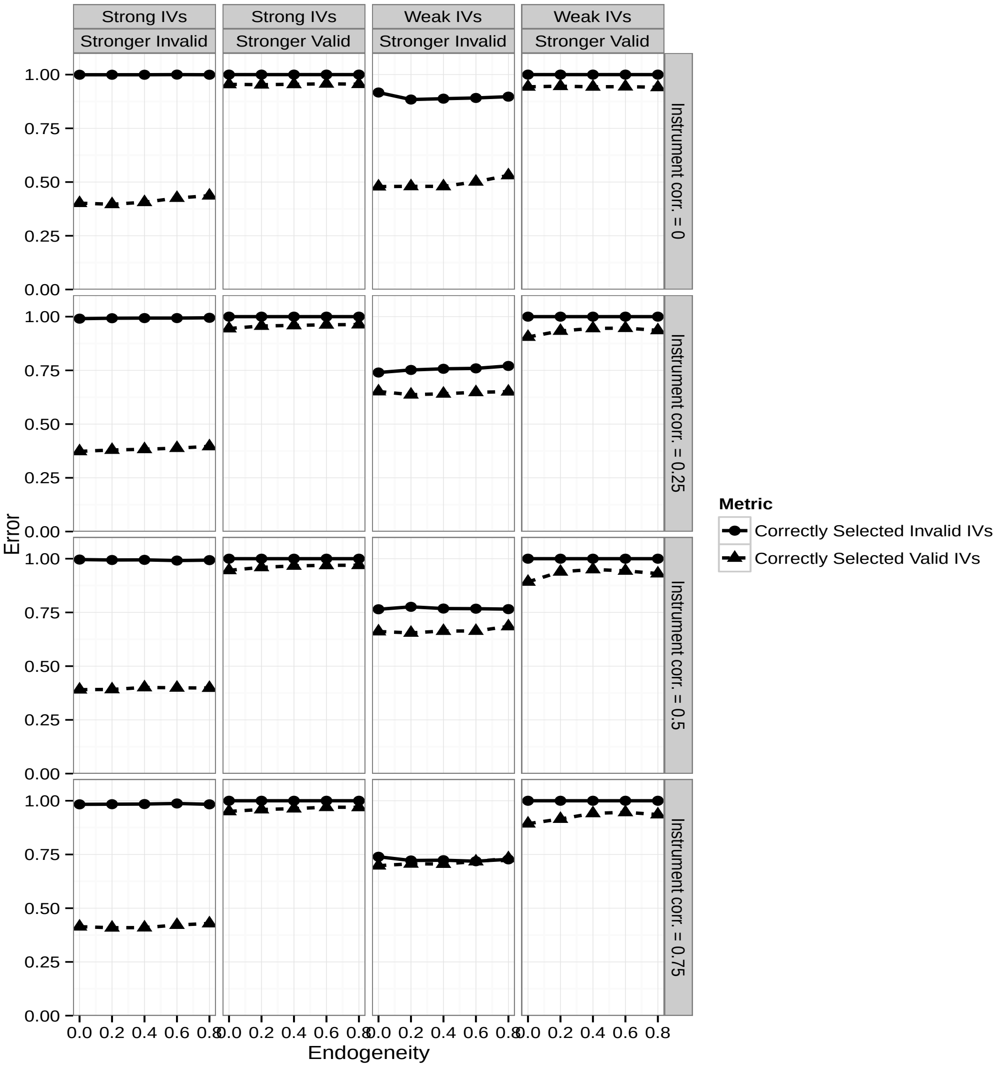

Figure 22: sisVIVE Simulation Study of $\boldsymbol{\alpha}^*$ with Different Endogeneity and Wher Correlation Exists Between All IVs (Setup 3). We also vary the instrument strength of valid and invalid instruments. There are ten ($L = 10$) instruments. Each line represents the average proportions of correctly selected valid instruments and correctly selected invalid instruments after 500 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 strengths for valid and invalid instruments, "Stronger Invalid" and "Stronger Valid", determined by varying $\boldsymbol{\gamma}^*$ while holding the absolute strength fixed. Each row corresponds to the maximum correlation between instruments.

for strong instruments and $s = 6$ for weak instruments. Again, similar to what we observed before, these points of s correspond to sisVIVE's deviation from the oracle in Figure 9. In contrast, the proportion of correctly selected valid instruments vary widely in Figure 23 and there does not seem to be any relationship between it and sisVIVE's deviation from the oracle in Figure 9.

Third, we look at the simulation setup in Section 2.4.5. Similar to our observations in Section 2.4.5, the pattern of simulations when we have different correlation structure between instruments is similar to the pattern of simulation when we have equi-correlation between all instruments in Section 2.4.4 when examining the performance of sisVIVE on $\boldsymbol{\alpha}^*$ (Figures 24, 25, 26, and 27)

Fourth, we look at the simulation setup in Section 2.4.6. Figures 28 and 29 represent cases where we vary endogeneity and the number of invalid instruments, respectively. Similar to what we observed with $L = 10$ and where all IVs have same pairwise correlation, when we vary endogeneity (Figure 28), but fix s to 30, we see that the proportion of correctly selected invalid instruments is 1. When we vary s (Figure 29), we again notice a sharp decrease in the proportion of correctly selected valid invalid instruments around $s = 50$ for all instrument strengths and magnitudes of the correlation.

In summary, measuring sisVIVE's performance on $\boldsymbol{\alpha}^*$ shows that a good estimate of β^* depends strongly on correctly selecting the invalid instruments more than correctly selecting the valid instruments. This observation remains true regardless of instrument correlation structure, types of instrument strength, levels of endogeneity, degree of instrument correlation, or the number of invalid instruments.

2.4.8. Choice of λ

In this section, we look at different ways to select λ. As discussed in Section 2.3.4, the choice of λ impacts the performance of sisVIVE where a high value of λ will push most elements of $\hat{\boldsymbol{\alpha}}_\lambda$ to zero while a low value of λ will do the opposite. In Section 2.3.4, we

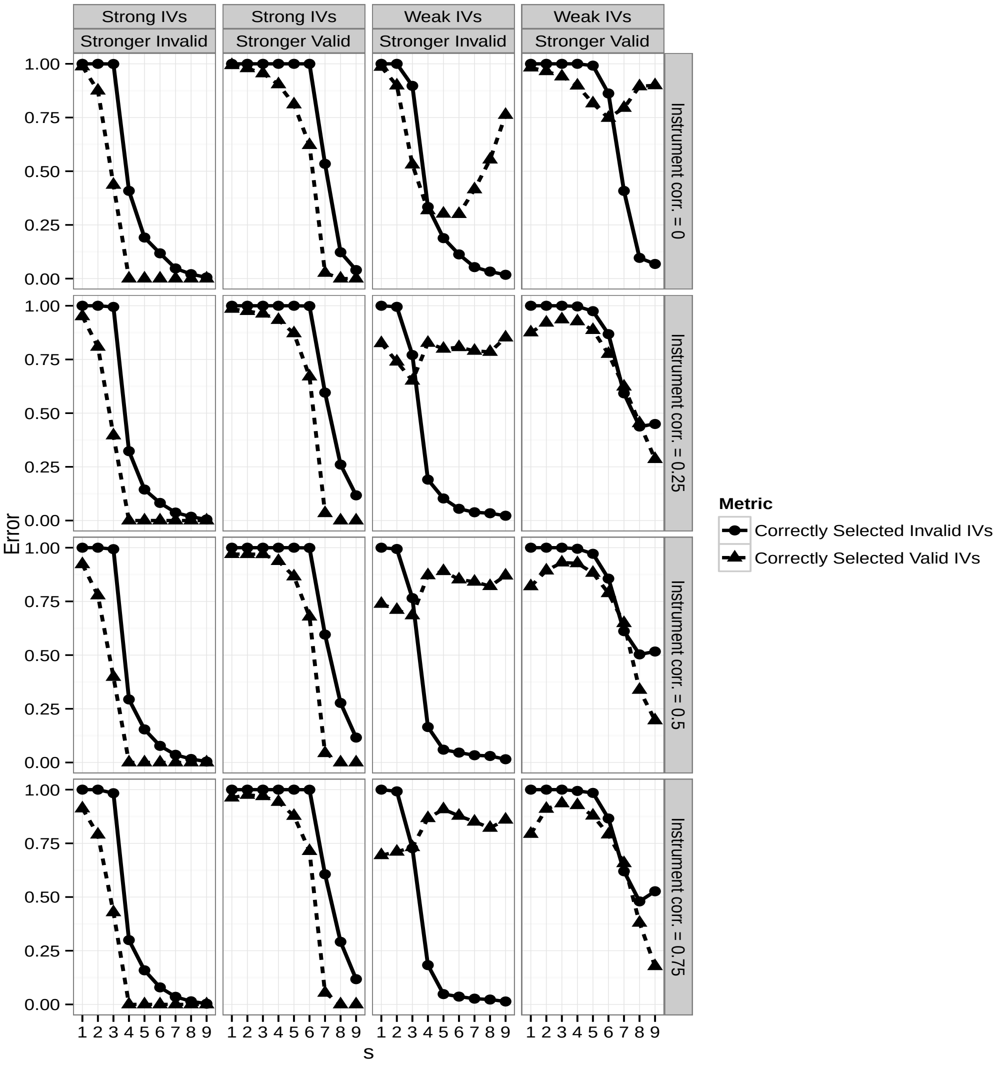

Figure 23: sisVIVE Simulation Study of $\boldsymbol{\alpha}^*$ With Different Number of Invalid IVs and Where Correlation Exists Between All IVs (Setup 3). We also vary the instrument strength of valid and invalid instruments. There are ten ($L = 10$) instruments. Each line represents the average proportions of correctly selected valid instruments and correctly selected invalid instruments after 500 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 strengths for valid and invalid instruments, "Stronger Invalid" and "Stronger Valid", determined by varying $\boldsymbol{\gamma}^*$ while holding the absolute strength fixed. Each row corresponds to the maximum correlation between instruments.

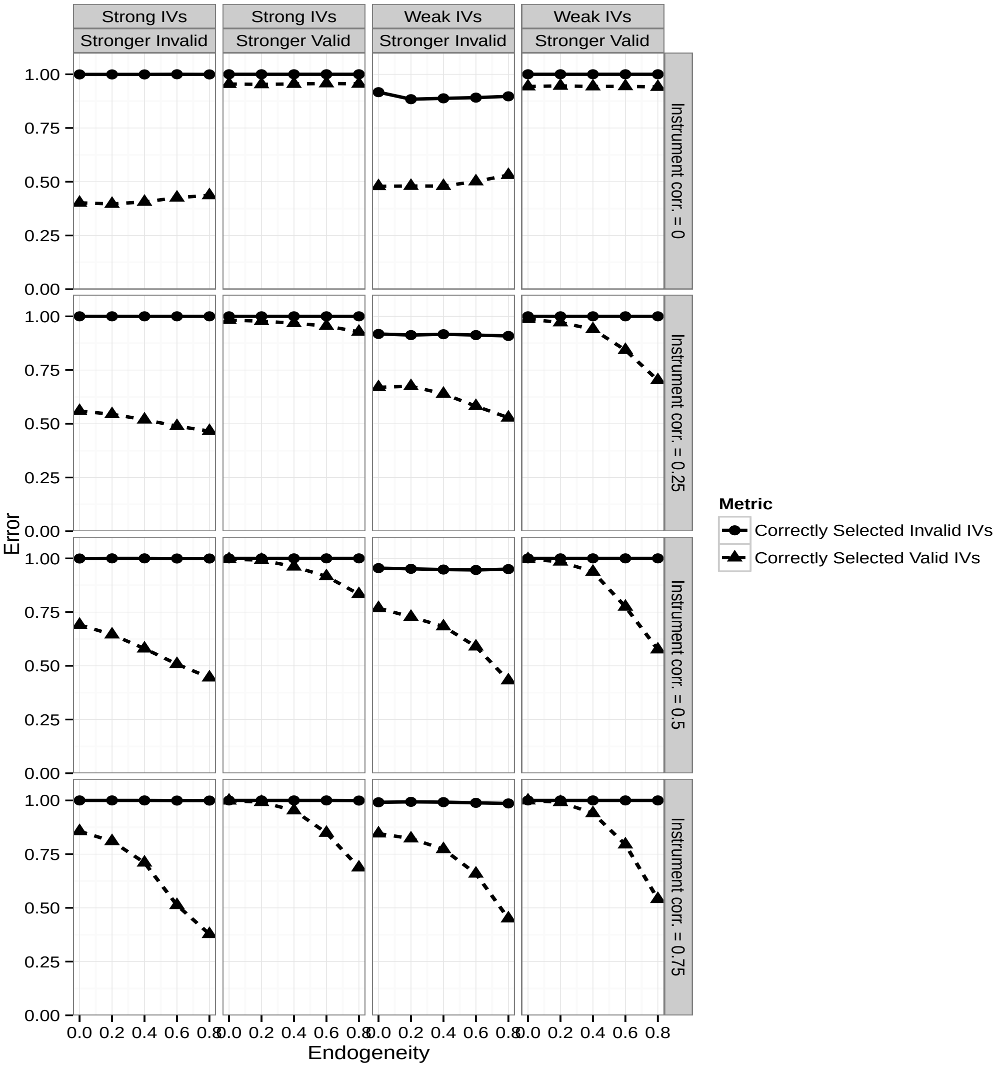

Figure 24: sisVIVE Simulation Study of $\boldsymbol{\alpha}^*$ With Different Endogeneity and Where Correlation Only Exists Within Valid and Invalid IVs (Setup 4). We also vary the instrument strength of valid and invalid instruments. There are ten ($L = 10$) instruments. Each line represents the average proportions of correctly selected valid instruments and correctly selected invalid instruments after 500 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 strengths for valid and invalid instruments, "Stronger Invalid" and "Stronger Valid", determined by varying $\boldsymbol{\gamma}^*$ while holding the absolute strength fixed. Each row corresponds to the maximum correlation between instruments, but correlation only exists within valid and invalid instruments.

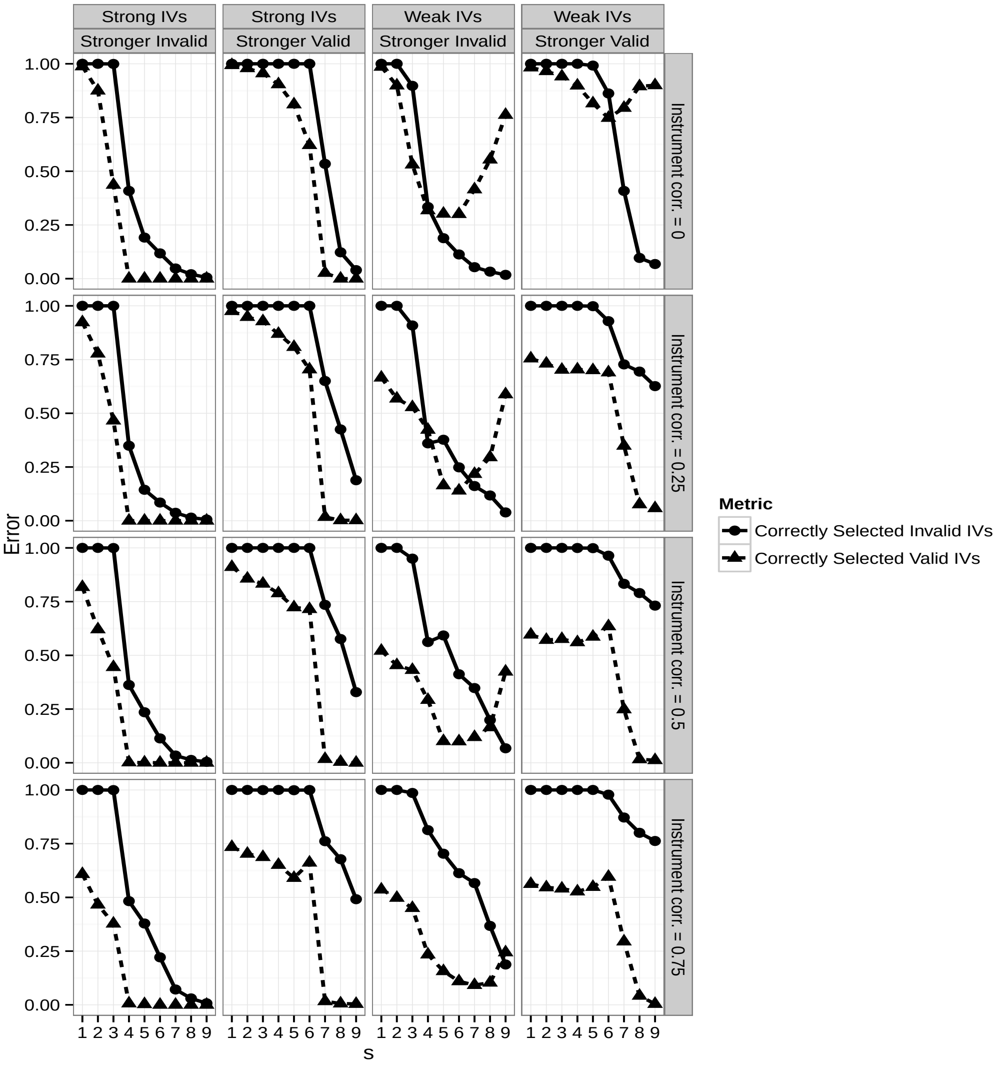

Figure 25: sisVIVE Simulation Study of $\boldsymbol{\alpha}^*$ With Different Number of Invalid IVs and Where Correlation Only Exists Within Valid and and Invalid IVs (Setup 4). We also vary the instrument strength of valid and invalid instruments. There are ten ($L = 10$) instruments. Each line represents the average proportions of correctly selected valid instruments and correctly selected invalid instruments after 500 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 strengths for valid and invalid instruments, "Stronger Invalid" and "Stronger Valid", determined by varying $\boldsymbol{\gamma}^*$ while holding the absolute strength fixed. Each row corresponds to the maximum correlation between instruments, but correlation only exists within valid and invalid instruments.

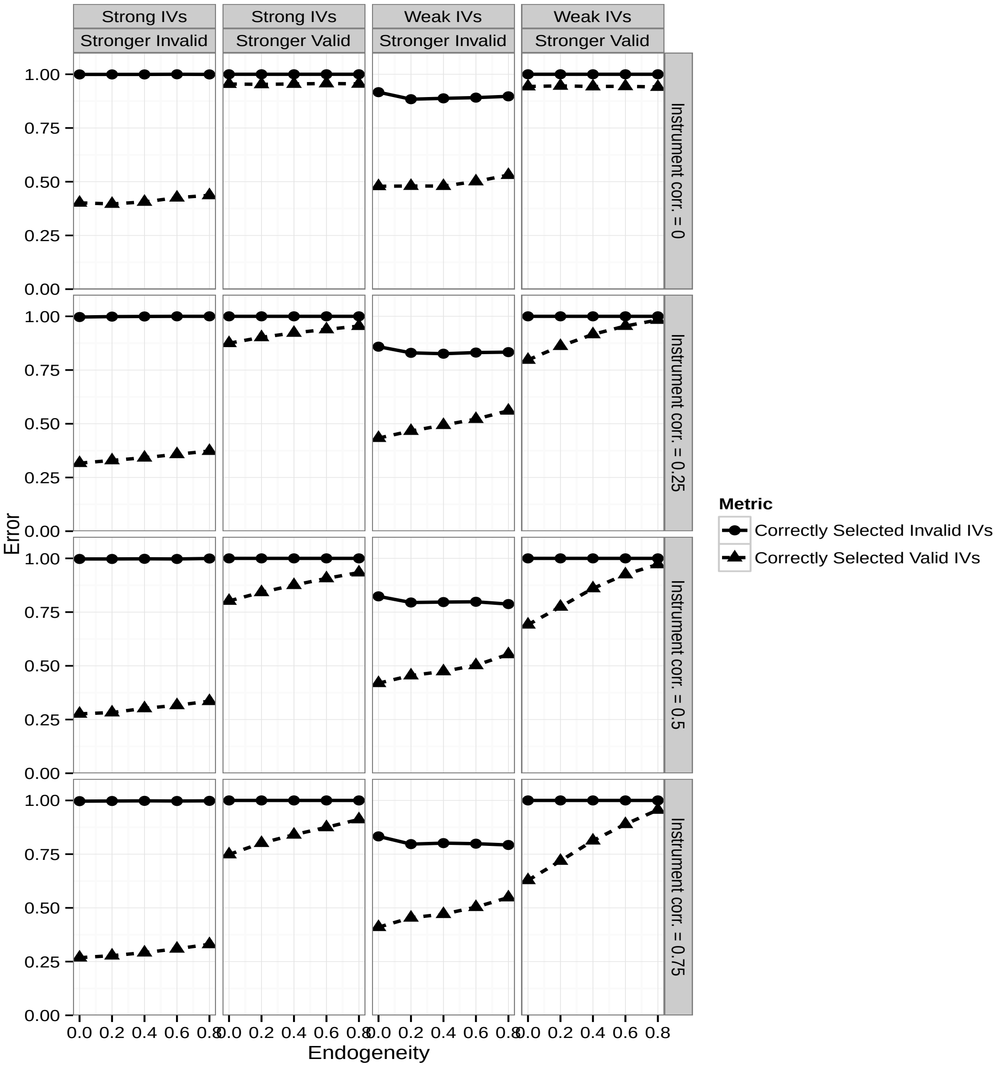

Figure 26: sisVIVE Simulation Study of $\boldsymbol{\alpha}^*$ With Different Endogeneity and Where Correlation Only Exists Between Valid and Invalid IVs (Setup 4). We also vary the instrument strength of valid and invalid instruments. There are ten ($L = 10$) instruments. Each line represents the average proportions of correctly selected valid instruments and correctly selected invalid instruments after 500 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 strengths for valid and invalid instruments, "Stronger Invalid" and "Stronger Valid", determined by varying $\boldsymbol{\gamma}^*$ while holding the absolute strength fixed. Each row corresponds to maximum correlation between instruments, but correlation only exists between valid and invalid instruments.

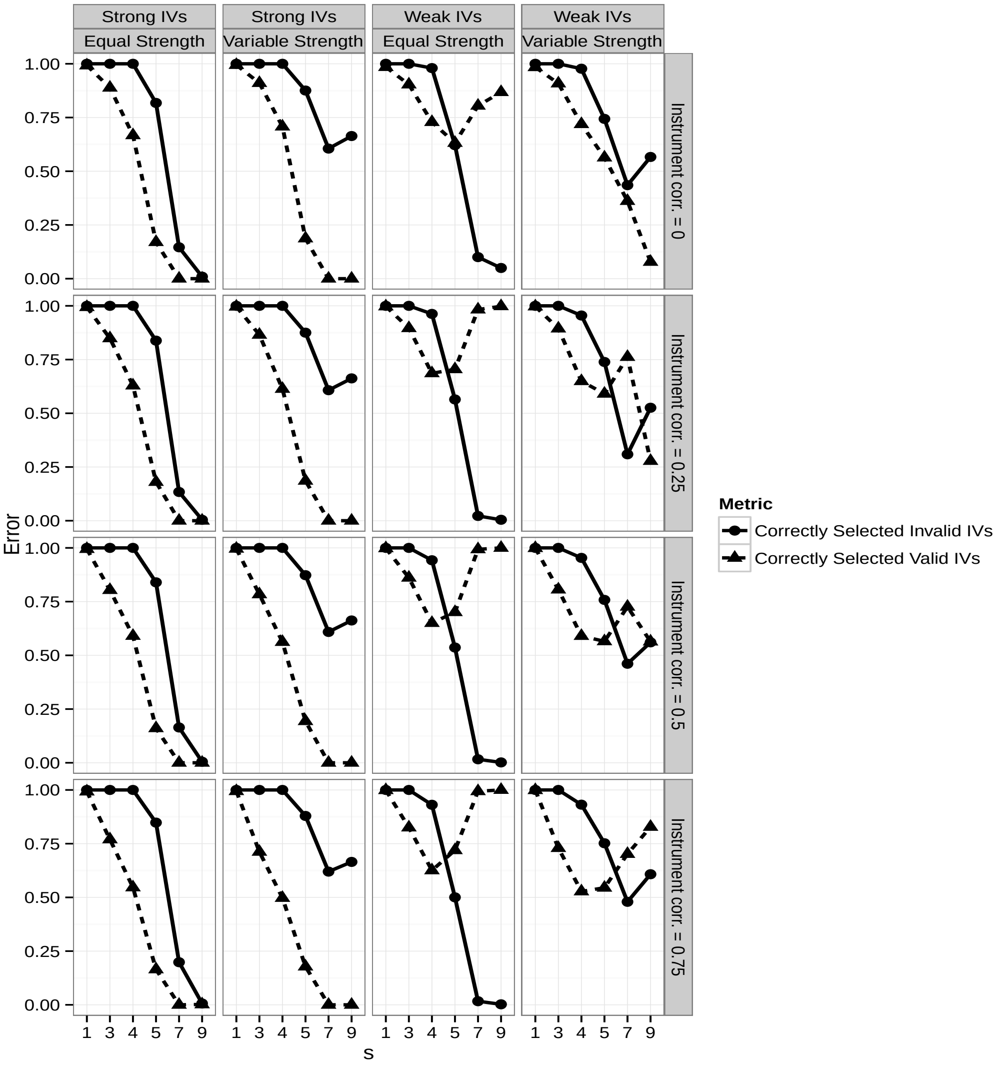

Figure 27: sisVIVE Simulation Study of $\boldsymbol{\alpha}^*$ With Different Number of Invalid IVs and Where Correlation Only Exists Between Valid and and Invalid IVs (Setup 4). We also vary the instrument strength of valid and invalid instruments. There are ten ($L = 10$) instruments. Each line represents the average proportions of correctly selected valid instruments and correctly selected invalid instruments after 500 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 strengths for valid and invalid instruments, "Stronger Invalid" and "Stronger Valid", determined by varying $\boldsymbol{\gamma}^*$ while holding the absolute strength fixed. Each row corresponds to maximum correlation between instruments, but correlation only exists between valid and invalid instruments.

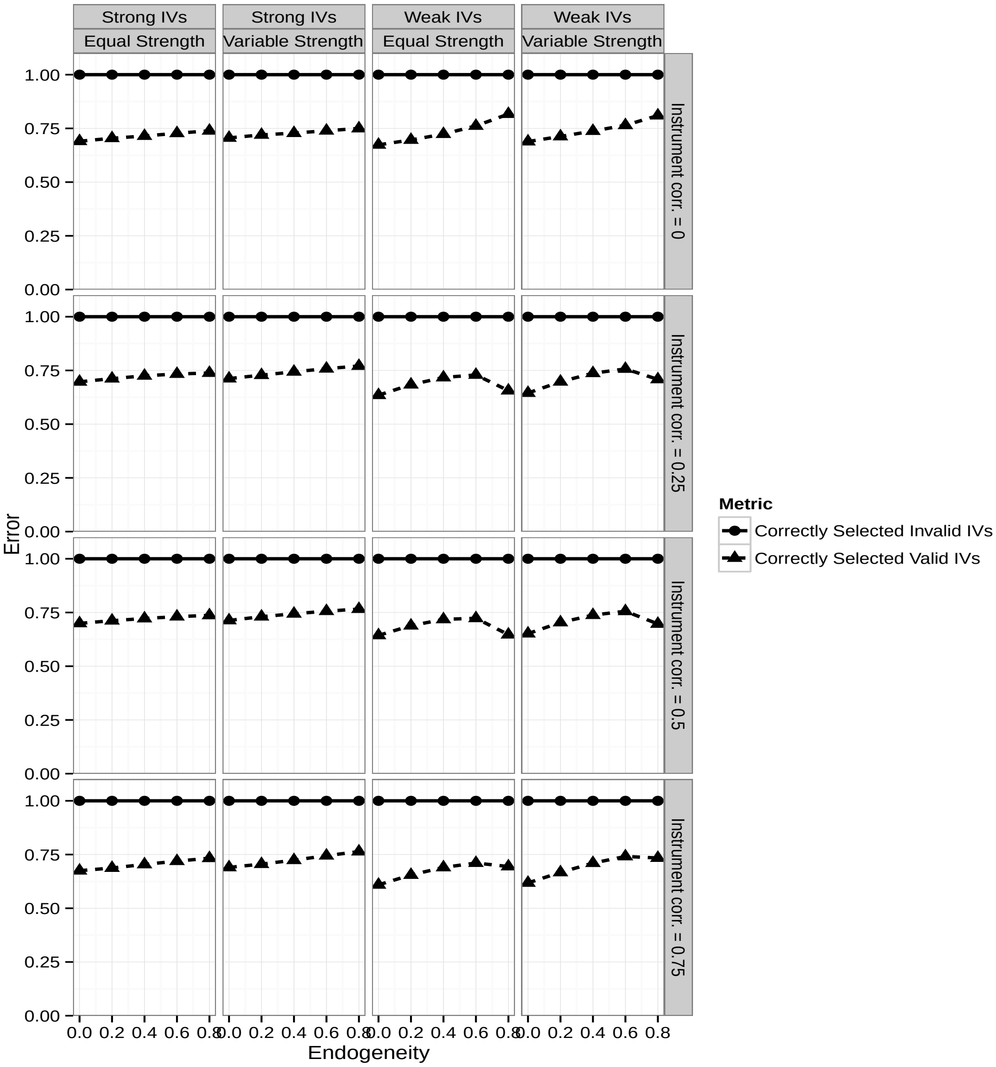

Figure 28: sisVIVE Simulation Study of $\boldsymbol{\alpha}^*$ With Different Endogeneity and Where Correlation Exists Between All IVs (Setup 5). There are ten ($L = 100$) instruments. Each line represents the average proportions of correctly selected valid instruments and correctly selected invalid instruments after 500 simulations. We fix the number of invalid instruments to $s = 30$. 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 $\boldsymbol{\gamma}^*$ while holding the absolute strength (i.e. concentration parameter) fixed. Each row corresponds to the maximum correlation between all instruments.

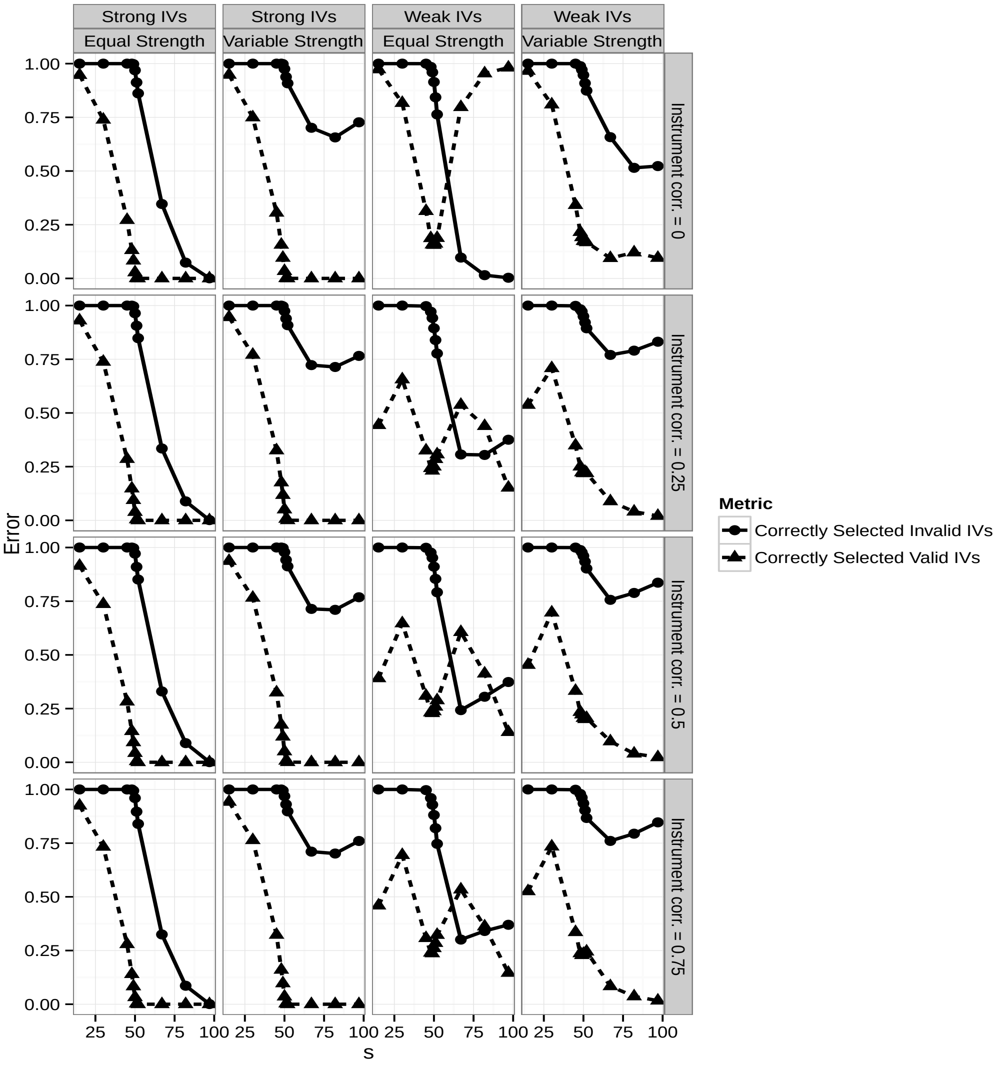

Figure 29: sisVIVE Simulation Study of $\boldsymbol{\alpha}^*$ With Different Number of Invalid IVs and Where Correlation Exists Between All IVs (Setup 5). There are 100 ($L = 100$) instruments. Each line represents the average proportions of correctly selected valid instruments and correctly selected invalid instruments after 500 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 $\boldsymbol{\gamma}^*$ while holding the absolute strength fixed. Each row corresponds to maximum correlation between all instruments.

suggested cross-validation with the “one standard error” rule as a data-driven method to choose the tuning parameter. In addition, in Section 2.3.5, we provided theoretical results which suggested choosing a λ that is greater than $3\|\mathbf{Z}^T \mathbf{P}_{\hat{\mathbf{D}}^{\perp}} \boldsymbol{\epsilon}\|_{\infty}$. We explore these two possible choices of λ and their impact on estimation.

We begin with a simulation study similar to Sections 2.4.2 and 2.4.4. In particular, we have $L = 10$ instruments of which the pairwise correlation between all instruments is 0.75 and the endogeneity is fixed at 0.8. We vary s, the number of invalid instruments and vary instruments' absolute strength, relative strength, and other strengths considered in Section 2.4.4. In short, the simulation setups we consider correspond to the last row of Figures 3 and 9. We do not simulate other correlation structures or different Ls because the simulation results in Sections 2.4.3 and 2.4.6 showed sisVIVE behaves similarly as the cases we consider in this section.

Table 7 shows the different values of λ averaged across 500 simulations where the overall, absolute instrument strength is strong (see Section 2.4.1 for details on the definition of absolute instrument strength). The column labeled “CV” denotes the average λs based on cross validation laid out in Section 2.3.4. The column labeled “Theory” denotes the average λs based on Theorem 2.2, specifically the average of $3\|\mathbf{Z}^T \mathbf{P}_{\hat{\mathbf{D}}^{\perp}} \boldsymbol{\epsilon}\|_{\infty}$ over 500 simulations. We use the same column heading labels in Figures 3 and 9. In almost all cases, cross validation tends to choose a smaller λ than one prescribed by Theorem 2.2, with the exception of $s = 9$ in the “Equal” column and $s = 7, 8,$ and 9 in the “Stronger Valid” column. Except for these cases, cross validation tends to prefer a small λ, thereby preferring $\hat{\boldsymbol{\alpha}}_{\lambda}$ to have more non-zero entries than zero entries and more instruments selected as invalid instruments than valid instruments.

Table 8 shows the estimation performance of sisVIVE, the median of $|\beta^* - \hat{\beta}_{\lambda}|$ over 500 simulations, based on two different λs, one based on cross validation and one based on Theorem 2.2. In most cases, sisVIVE with a cross validated λ performs better than sisVIVE with a theory-based λ. For the “Equal” and ”Variable” case, when $s < 5$, sisVIVE

	Equal		Variable		Stronger Invalid		Stronger Valid	
s	CV	Theory	CV	Theory	CV	Theory	CV	Theory
1	1.88	2.70	2.04	2.71	1.53	2.70	2.06	2.72
2	1.36	2.66	1.39	2.67	0.95	2.65	1.58	2.68
3	1.06	2.64	1.12	2.66	0.84	2.64	1.33	2.68
4	0.84	2.64	0.86	2.65	1.08	2.63	1.16	2.68
5	1.70	2.63	1.33	2.64	0.87	2.62	0.99	2.67
6	1.78	2.62	1.10	2.63	0.85	2.61	0.96	2.67
7	2.02	2.62	0.79	2.64	0.91	2.61	3.40	2.68
8	2.41	2.62	0.86	2.62	1.01	2.61	3.74	2.67
9	3.19	2.62	0.45	2.62	1.31	2.60	6.03	2.67

Table 7: Average λ From Cross Validation (CV) and Theorem 2.2 (Theory) for Strong IVs. Averages are taken after 500 simulations.

with a cross-validated λ performs better than sisVIVE with a theory-based λ. For the "Stronger Invalid" case, when $s < 3$, sisVIVE with a cross validated λ performs better than sisVIVE with a theory-based λ. However, when $s \geq 3$, sisVIVE with a cross validated λ performs worse than sisVIVE with a theory-based λ, although the differences between the two decrease as s increases. For the "Stronger Valid" case, sisVIVE with a cross validated λ always dominates sisVIVE with a theory-based λ, although the differences between the two are slight when $s \geq 7$.

	Equal		Variable		Stronger Invalid		Stronger Valid	
s	CV	Theory	CV	Theory	CV	Theory	CV	Theory
1	0.13	0.17	0.14	0.16	0.13	0.19	0.14	0.16
2	0.16	0.27	0.16	0.27	0.16	0.34	0.16	0.24
3	0.18	0.39	0.18	0.37	0.24	0.54	0.18	0.32
4	0.21	0.53	0.22	0.53	1.57	1.34	0.20	0.41
5	0.71	1.15	0.76	1.43	1.43	1.25	0.23	0.55
6	2.43	2.34	2.05	1.93	1.35	1.23	0.28	0.71
7	2.42	2.37	1.83	1.95	1.28	1.21	3.83	3.95
8	2.35	2.34	1.98	2.05	1.22	1.18	4.24	4.39
9	2.29	3.01	1.23	1.37	1.17	1.16	4.34	4.51

Table 8: sisVIVE Estimation Performance of β^* Between λ by Cross Validation (CV) and from Theorem 2.2 (Theory) for Strong IVs. Averages are taken after 500 simulations.

Table 9 considers the same setup as Table 7, except we now look at instruments where their overall, absolute strength is weak. Under this case, we see drastic differences between cross validation and Theorem 2.2. For example, for the “Equal” and “Variable” cases, when $s < 5$, λ chosen based on cross validation is, on average, smaller than λ chosen based on Theorem 2.2. When $s \geq 5$, λ chosen based on cross validation is, on average, bigger than λ chosen based on Theorem 2.2. For the “Stronger Invalid” case, when $s < 3$, λ based on cross validation is, on average, smaller than λ based on Theorem 2.2. But, when $s \geq 3$, the opposite is the case. Finally, for the “Stronger Valid” case, this transition phenomena occurs at $s = 6$.

	Equal		Variable		Stronger Invalid		Stronger Valid	
s	CV	Theory	CV	Theory	CV	Theory	CV	Theory
1	1.36	3.20	1.56	3.23	1.05	3.13	1.52	3.24
2	1.25	3.00	1.22	3.01	0.93	2.92	1.47	3.07
3	1.12	2.91	1.11	2.94	3.67	2.81	1.26	3.00
4	2.06	2.86	1.83	2.89	9.47	2.75	1.13	2.97
5	6.30	2.80	4.34	2.84	10.52	2.71	1.20	2.92
6	11.99	2.78	7.48	2.80	10.74	2.69	3.36	2.93
7	14.14	2.76	5.92	2.77	10.58	2.67	7.79	2.93
8	14.04	2.75	5.94	2.75	9.92	2.66	9.70	2.93
9	13.16	2.74	2.02	2.68	9.47	2.64	7.09	2.96

Table 9: Average λ From Cross Validation (CV) and Theorem 2.2 (Theory) for Weak IVs. Averages are taken after 500 simulations.

Table 10 considers the same setup as Table 8, except we now look at instruments where their overall, absolute strength is weak. Similar to Table 8, sisVIVE with a cross validated λ performs better than sisVIVE with a theory-based λ, with the only exception at $s = 5$ under the “Equal” column. In fact, sisVIVE with a cross validated λ performs drastically better than sisVIVE with a λ based on Theorem 2.2 in the following cases: $s < 5$ (for “Equal” and “Variable” cases), $s < 3$ (for “Stronger Invalid” case), and $s < 7$ (for “Stronger Valid” case).

Based on these simulations, sisVIVE based on cross validation generally performs better than sisVIVE based on Theorem 2.2, especially when the overall instrument strength is weak. We also note that cross validation tends to choose a smaller λ than the one based on Theorem 2.2, suggesting that for better estimation, it is preferable to set only a few elements

	Equal		Variable		Stronger Invalid		Stronger Valid	
s	CV	Theory	CV	Theory	CV	Theory	CV	Theory
1	0.44	0.63	0.44	0.60	0.43	0.69	0.44	0.61
2	0.51	0.96	0.50	0.94	0.50	1.13	0.52	0.88
3	0.55	1.30	0.55	1.26	0.70	1.86	0.56	1.13
4	0.61	1.74	0.61	1.75	3.19	3.77	0.58	1.43
5	4.10	3.80	3.98	3.93	3.25	3.78	0.62	1.83
6	5.28	6.03	5.28	5.54	3.36	3.79	0.73	2.52
7	5.84	6.55	5.58	5.63	3.47	3.77	7.51	7.68
8	6.29	6.75	6.19	6.19	3.52	3.70	9.69	9.77
9	6.72	6.90	4.18	4.34	3.56	3.64	10.86	10.91

Table 10: sisVIVE Estimation Performance of β^* Between λ by Cross Validation (CV) and from Theorem 2.2 (Theory) for Weak IVs. Averages are taken after 500 simulations.

of $\hat{\boldsymbol{\alpha}}_\lambda$ to zero and declare more instruments to be invalid than valid. This observation was also seen in our simulation in Section 2.4.7 where low median absolute error, $|\beta^* - \hat{\beta}_\lambda|$, was tied to high proportion of correctly chosen invalid instruments. As an aside, this observation is in contrast with estimating sparse vectors in typical high dimensional regression settings where many zeroed elements are desirable in the estimated sparse vector.

Despite the simulation evidence suggesting the use of cross validation to choose λ over Theorem 2.2 to choose λ, unfortunately, there is little theory to justify the use of cross validation in ℓ_1 penalization settings (Hastie et al., 2009; Bühlmann and van der Geer, 2011). However, Section 2.5.1 of Bühlmann and van der Geer (2011) does provide limited theoretical results suggesting that λ based on cross validation tends to set few elements of $\hat{\boldsymbol{\alpha}}_\lambda$ to zero, a desirable property in our setting where we want to select more instruments to be invalid than valid for better estimation performance of $\hat{\beta}_\lambda$.

Besides cross validation and Theorem 2.2, there is another way to choose λ if we assume Corollary 2.1 holds for our data. Specifically, if we are in the always identified region where $s < U \leq L/2$, one possible method of choosing λ would be to find the λ where exactly $U = L/2$, say $\lambda_{L/2}$. From there, we grid the values of potential λs between 0 and $\lambda_{L/2}$ and choose the λ that minimizes the estimating equation $||\mathbf{P_Z}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha} - \mathbf{D}\beta)||_2$. It would be interesting to investigate this method in future research.

2.4.9. Summary of Simulation Studies

The simulation studies above covered different types of instrument strength, correlation structure between instruments, and total number of potential instruments. We also explored different metrics of error, such as the proportion of correctly selected valid instruments and invalid instruments, to analyze the relationship between estimating β^* and $\boldsymbol{\alpha}^*$. In addition, we also computed the conditions for Corollary 2.2, specifically ρ, μ, and λ required to achieve the performance bound. In every setting 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 stronger than the valid instruments and $s = 4$. In addition, we showed 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 Theorem 2.2 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 Theorem 2.2. In fact, in most cases, sisVIVE with a cross-validated λ performs better than sisVIVE with a λ chosen by Theorem 2.2.

Overall, sisVIVE using a cross-validated λ does much better than naive TSLS, the most frequently used estimator in MR and IV. In many cases, sisVIVE beats the naive TSLS and is comparable to oracle TSLS. The promising simulation results suggest that sisVIVE should be used whenever there is concern about invalid instruments.

2.5. Data Analysis: The Effect of Obesity on Quality of Life

2.5.1. Background

We demonstrate the potential benefit of using sisVIVE in MR by analyzing the effect of obesity, the exposure, on health-related quality of life, 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. (2006) have shown that there is a negative association between obesity and health-related quality of life. However, a fundamental difficulty with these studies is that the outcome, health-related quality of life, encompasses various factors about the individual, making it difficult to control for all possible confounders that may affect obesity and health-related quality of life (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 individuals with more than 10% missing genotype data (Roetker et al., 2012). Our analysis of the data set contains $n = 3712$ individuals with 1913 females and 1799 males born mostly between 1938 to 1940.

To measure health-related quality of life, 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 body mass index (BMI) and the US National Institute of Health clinical guidelines (National Institute of Health, 1998) that were also used in Trakas et al. (2001) and Sach et al. (2006) in their analysis. 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 US National Institute of Health clinical guidelines (National Institute of Health, 1998). For instance, exposure value of zero corresponds to non-obese 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 (2.1) will correspond to the effect of moving up one obese class on the HUI-3 index of health-related quality of life. In Section 2.5.3, we explore different definitions to quantify obesity and the resulting estimates from different definitions.

For potential candidate instruments, we use the following single nucleotide polymorphisms (SNPs) in the WLS that have been previously shown to be associated with obesity: rs1421085, rs1501299, and rs2241766 (see Table 11). rs1421085 is in the FTO gene and it has been shown to be strongly associated with obesity (Dina et al., 2007; Price et al., 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.

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)

Table 11: Summary of Instruments in the Data Analysis. MAF stands for minor allele frequency

2.5.2. Analysis

A simple ordinary least squares analysis estimates that an increase in one obese class is associated with a 0.052 (SE: 0.0040) decrease in HUI-3 score, which is consistent with prior literature (Trakas et al., 2001; Sach et al., 2006). If we use TSLS, under the operating assumption that all the instruments are valid, the estimated causal effect is -0.00094 (SE: 0.081), i.e. 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.

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 health-related quality of life 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 for both analysis are summarized in Tables 12 and 13. The reduced form estimates are computed by using ordinary least squares (OLS) where the genetic instruments are the explanatory variables and the dependent variables are BMI and Health Utility Index Mark 3 (HUI-3).

Also, for the data analysis with three SNPs, the Sargan overidentification test (Sargan,

Instruments	BMI (SE)	HUI-3 (SE)
rs1421085	-0.05 (0.02)	0.0003 (0.004)
rs1501299	0.01 (0.02)	0.002 (0.005)
rs2241766	-0.0007 (0.03)	-0.0001 (0.007)

Table 12: Reduced Form Estimates for HUI-3 and BMI for Three SNPs

Instruments	BMI (SE)	HUI-3 (SE)
rs1421085	-0.05 (0.02)	0.0004 (0.004)
rs1501299	0.01 (0.02)	0.002 (0.005)
rs2241766	-0.0006 (0.03)	-0.0004 (0.007)
rs6265	-0.004 (0.02)	-0.008 (0.005)

Table 13: Reduced Form Estimates for HUI-3 and BMI for Four SNPs

1958), which tests assumptions (A2) and (A3) in the presence of multiple instruments, gives a Chi-squared value of 0.12 (p-value: 0.94), retaining the null hypothesis that the instruments are all valid under the 0.05 significance level. For the data analysis with four SNPs, the Sargan overidentification test gives a Chi-squared value of 2.53 (p-value: 0.47). The first stage F statistic with three instruments is 3.16. The first stage F statistic with four instruments is 2.38. Based on the two F statistics, the instruments are generally weak.

We also estimate the implied structural correlation from our model, specifically the correlation between D_i, the exposure, and ϵ_i. We estimate ϵ_i by taking the residual from the estimates of β^* and $\boldsymbol{\alpha}^*$, $\hat{\epsilon}_i = Y_i - D_i\hat{\beta}_\lambda - \mathbf{Z}_{i.}^T\hat{\boldsymbol{\alpha}}_\lambda$ where λ is chosen by cross-validation described in Section 2.3.4. We find that our estimate of this correlation is -0.2, suggesting a mild form of endogeneity.

In both data analyses, sisVIVE operates under the assumption that there may be invalid instruments, which are typical in MR studies, while TSLS operates under the assumption that all instruments are valid. 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.

2.5.3. A Digression: Defining Obesity Using BMI

In this section, we carefully look at different methods to quantify obesity, our exposure in the data set, using BMI. First, we looked at BMI across several categories of obesity. The categories were based on US National Institute of Health clinical guidelines (National Institute of Health, 1998) and were also used in Trakas et al. (2001) and Sach et al. (2006) in their analysis. Table 14 summarizes the relationship between obesity categories and HUI-3 in our data.

		Health Utility Index Mark 3		
Obesity Categories	N	1st quartile	Median	3rd quartile
Not obese (BMI $<$ 30)	2581	0.84	0.92	0.97
Obese class I (30 $\leq$ BMI $<$ 35)	777	0.73	0.91	0.97
Obese class II (35 $\leq$ BMI $<$ 40)	246	0.66	0.85	0.97
Obese class III (40 $\leq$ BMI)	108	0.51	0.72	0.91
All categories	3712	0.78	0.92	0.97

Table 14: Relationship Between Obesity and Health Utility Index Mark 3 (HUI-3)

We notice that among different obese classes, the median HUI-3 scores are different. Hence, simply classifying individuals as obese vs. not obese ignores the variation of HUI-3 scores among different obese classes. This led us to explore different ways of quantifying obesity through BMI as follows.

1. The binary BMI takes a value of one if BMI is greater than or equal to 30 (i.e. obese) and zero otherwise.

2. BMI A is what we use in the main analysis.

3. BMI B is defined to be similar to Trakas et al. (2001), except the magnitude of the BMIs is taken into consideration. Specifically, if an individual's BMI is less than 30, the individual's exposure is assigned a value of zero. If an individual's BMI is between 30 and 35 (i.e. Obese Class I), the individual's exposure is assigned a value of one. If an individual's BMI is between 35 and 40 (i.e. Obese Class II), the individual's exposure is assigned a value of three. If an individual's BMI is above 40 (i.e. Obese

Class III), the individual's exposure is assigned a value of six.

4. The censored BMI takes into account the actual value of BMI at the obese range so that it not only indicates obesity, but also measures its severity. Specifically, the censored BMI is defined as the maximum of (BMI -30) and 0 (i.e. $\max(\text{BMI}-30, 0)$).

For each method of quantifying obesity, we estimate β^* by using ordinary least squares (OLS), two stage least squares (TSLS) under the assumption that all the instruments are valid, and sisVIVE. These results are reported in Tables 15 and 16 for the cases of three and four instruments used in our main analysis. Overall, we notice that the estimates of OLS, TSLS, and sisVIVE tend to be similar across different definitions of obesity. Granted, it is difficult to compare the estimates since each exposure variable measures slightly different aspects about obesity and its impact on HUI-3. We also note that in the case of four instruments where one of the instrument, rs6265, was suspect, sisVIVE correctly picks rs6265 to be an invalid instrument in every method of quantifying obesity.

Exposure	OLS (SE)	TSLS (SE)	sisVIVE, Invalid Instrument
Binary BMI	-0.074 (SE: 0.0070)	-0.012 (SE: 0.18)	-0.012, None
BMI A	-0.052 (SE: 0.0040)	-0.00094 (SE: 0.081)	-0.00094, None
BMI B	-0.031 (SE: 0.0024)	-0.0011 (SE: 0.051)	-0.0011, None
Censored BMI	-0.013 (SE: 0.0010)	-0.00019 (SE: 0.022)	-0.00019, None

Table 15: Different Definitions of Obesity and The Resulting Estimates With Three Instruments

Exposure	OLS (SE)	TSLS (SE)	sisVIVE, Invalid Instrument
Binary BMI	-0.074 (SE: 0.0070)	-0.097 (SE: 0.17)	-0.039, rs6265
BMI A	-0.052 (SE: 0.0040)	-0.0086 (SE: 0.080)	-0.0037, rs6265
BMI B	-0.031 (SE: 0.0024)	-0.0012 (SE: 0.051)	-0.0017, rs6265
Censored BMI	-0.013 (SE: 0.0010)	0.00091 (SE: 0.022)	-0.00011, rs6265

Table 16: Different Definitions of Obesity and The Resulting Estimates With Four Instruments

2.6. Discussion

This paper 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 paper discusses a model in which treatment effects are constant. Angrist et al. (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 2.1 might not be equal to each other for different sets of valid instruments and Theorem 2.1 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 2.2 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.

CHAPTER 3 : Robust Confidence Interval Estimation of Causal Effects With Possibly Invalid Instruments

This is joint work with Tony Cai and Dylan Small

3.1. Introduction

In the previous chapter, we considered violations of (A2) and (A3) and proposed identification results based on imposing an upper bound on the number of invalid instruments among the candidate instruments, without knowing exactly which instruments are valid or knowing the exact number of invalid instruments, or imposing any structure on the instruments. The previous chapter also proposed a point estimator, called sisVIVE, to estimate the causal effect when invalid instruments are present.

This chapter focuses on the same setting, but we develop robust confidence intervals when candidate instruments might violate (A2) and (A3). Like before, we only assume that we know an upper bound on the number of invalid instruments, without knowing exactly which instruments are invalid. In this setting, we propose a simple and general confidence interval procedure that theoretically guarantees the correct coverage rate and is robust to possibly invalid instruments. The confidence interval is based on inverting statistical tests over a range of subsets of instruments that are potentially valid. We also propose various ways to obtain short and informative confidence intervals with our procedure by exploring various tests common in instrumental variables and conducting pretests. The simulation study shows that our method is robust when invalid instruments are present compared to other popular methods in the instrumental variables literature. We also demonstrate that our method can produce valid, short, and informative confidence intervals by analyzing a data set concerning the causal effect of income on food expenditure.

3.2. Robust Confidence Intervals by Inverting Tests

3.2.1. Review of Notation

We use the potential outcomes notation (Rubin, 1974) for instruments laid out in Holland (1988), Small (2007) and Chapter 2. Specifically, let there be L potential candidate instruments and n individuals in the sample. Let $Y_i^{(d,z)}$ be the potential outcome that individual i would have if the individual were to have exposure d, a scalar value, and instruments z, an L dimensional vector. Let D_i^z be the potential exposure if the individual had instruments z. For each individual, only one possible realization of $Y_i^{(d,z)}$ and $D_i^{(z)}$ is observed, denoted as Y_i and D_i, respectively, based on his/her observed instrument values $Z_{i.}$, an L dimensional vector, and observed exposure D_i. In total, we have n observations of $(Y_i, D_i, Z_{i.})$. We denote $Y = (Y_1, \ldots, Y_n)$, $D = (D_1, \ldots, D_n)$ and Z to be the n by L matrix where row i consists of $Z_{i.}$.

For any subset $A \subseteq \{1, \ldots, L\}$ with cardinality $c(A)$, let Z_A be an n by $c(A)$ matrix of instruments where the columns of Z_A are from the set A, $P_{Z_A} = Z_A(Z_A^T Z_A)^{-1} Z_A^T$ be the orthogonal projection matrix onto the column space of Z_A and $R_{Z_A} = I - P_{Z_A}$ be the residual projection matrix where I is an n by n identity matrix. We assume that $Z_A^T Z_A$ has a proper inverse unless otherwise stated. Also, for any L dimensional vector π, let π_A only consist of elements of the vector π determined by the set A.

3.2.2. Review of Model and Definition of Valid Instruments

For two possible values of the exposure d', d and instruments z', z, we assume the following potential outcomes model

$$Y_i^{(d',z')} - Y_i^{(d,z)} = (z' - z)^T \phi^* + (d' - d)\beta^*, \quad E\{Y_i^{(0,0)} \mid Z_{i.}\} = Z_{i.}^T \psi^* \tag{3.1}$$

where ϕ^*, ψ^*, and β^* are unknown parameters. The parameter β^* represents the causal parameter of interest, the causal effect (divided by $d' - d$) of changing the exposure from

d' to d on the outcome. The parameter ϕ^* represents violation of (A2), the direct effect of the instruments on the outcome. If (A2) holds, then $\phi^* = 0$. The parameter ψ^* represents violation of (A3), the presence of unmeasured confounding between the instrument and the outcome. If (A3) holds, then $\psi^* = 0$.

Let $\pi^* = \phi^* + \psi^*$ and $\epsilon_i = Y_i^{(0,0)} - E\{Y_i^{(0,0)} \mid Z_{i.}\}$. When we combine equations (3.1) along with the definition of ϵ_i, the observed data model becomes

$$Y_i = Z_{i.}^T \pi^* + D_i \beta^* + \epsilon_i, \quad E(\epsilon_i \mid Z_{i.}) = 0 \tag{3.2}$$

The observed model is also known as the under-identified single-equation linear model in econometrics (page 83 of Wooldridge (2010)). Note that (3.2) is not a usual regression model because D_i might be correlated with ϵ_i. In particular, the parameter β^* measures the causal effect of changing D on Y rather than an association. As mentioned in Chapter 2, we discuss extensions of the model (3.2) to include heterogeneous causal effects and non-linear effects. Also, the model can incorporate exogenous covariates, say $\mathbf{X}_{i.}$ and we can project them out by using Frisch-Waugh-Lovell Theorem to reduce the model to (3.2) (Davidson and MacKinnon, 1993). The parameter π^* in the observed data model (3.2) combines both the violation of (A2), represented by ϕ^*, and the violation of (A3), represented by ψ^*. If both (A2) and (A3) are satisfied, then $\phi^* = \psi^* = 0$ and $\pi^* = 0$. Hence, the value of π^* captures whether instruments are valid versus invalid. Definition 3.1 formalizes this idea.

Definition 3.1. Suppose we have L candidate instruments along with the models (3.1)–(3.2). We say that instrument $j = 1, \ldots, L$ is valid if $\pi_j^* = 0$ and invalid if $\pi_j^* \neq 0$.

When there is only one instrument, $L = 1$, Definition 3.1 of a valid instrument is identical to the definition of a valid instrument in Holland (1988). Specifically, assumption (A2), the exclusion restriction, which means $Y_i^{(d,z)} = Y_i^{(d,z')}$ for all d, z, z', is equivalent to $\phi^* = 0$ and assumption (A3), no unmeasured confounding, which means $Y_i^{(d,z)}$ and $D_i^{(z)}$ are independent of $Z_{i.}$ for all d and z, is equivalent to $\psi^* = 0$, implying $\pi^* = \phi^* + \psi^* = 0$. Definition 3.1 is also a special case of the definition of a valid instrument in Angrist et al. (1996) where here we assume the model is additive, linear, and has a constant treatment effect β^*. Hence, when multiple instruments, $L > 1$, are present, our models (3.1)–(3.2) and Definition 3.1 can be viewed as a generalization of the definition of valid instruments in Holland (1988).

Let $s = 0, \ldots, L-1$ to be the number of invalid instruments and U be an upper bound on s plus 1, i.e. the number of invalid instruments is assumed to be less than U. We assume that there is at least one valid IV, even if we don't know which among the L IV is valid, since if all L IVs are invalid (i.e. $s = L$), identification would not be possible. This is the setup considered in Chapter 2 as a relaxation to traditional instrumental variables setups where one knows exactly which instruments are valid and invalid. For simplicity, we consider the case where at less than half of the candidate instruments are invalid, $U \leq L/2$, because all the parameters in the model (3.2) are always identified under this setup (see Chapter 2 for details). However, the proposed procedures will work for any upper bound U, exceeding $L/2$.

3.2.3. A General Procedure for Robust Confidence Intervals

Let $I = \{1, \ldots, L\}$ be the L candidate instruments and $B^* \subseteq \{1, \ldots, L\}$ be the true set of valid instruments. Given B^*, consider a test statistic $T(\beta_0, B^*)$ of the null hypothesis $H_0 : \beta^* = \beta_0$ versus the alternative $H_a : \beta^* \neq \beta_0$. It is well known that inverting a test based on $T(\beta_0, B^*)$ that has level α provides a $1-\alpha$ confidence interval for β^*, denoted as $C_{1-\alpha}(Y, D, Z, B^*)$.

$$C_{1-\alpha}(Y, D, Z, B^*) = \{\beta_0 \mid T(\beta_0, B^*) \leq \nu_{1-\alpha}\} \tag{3.3}$$

where $\nu_{1-\alpha}$ is the $1-\alpha$ quantile of the null distribution of $T(\beta_0, B^*)$.

Unfortunately, in our problem, we do not know the true set B^* of valid instruments, so we cannot directly use (3.3). However, in our model description in Section 3.2.2, we have an upper bound on the number of invalid instruments, s, by U where $s < U$ and consequently, a lower bound for the number of valid instruments, $L - s > L - U$ and thus a lower bound on the cardinality of the set B^*, $c(B^*)$, $c(B^*) > L - U$. Using this lower bound, we can take unions of $C_{1-\alpha}(Y, D, Z, B)$ over possible sets of valid instruments $B \subseteq I$ where $c(B) > L - U$; the confidence interval using the true set of instruments $C(Y, D, Z, B^*)$ will be in this union since $c(B^*) > L - U$. Our proposal is exactly this, except that we restrict the subsets B to be of size $c(B) = L - U + 1$.

$$C_{1-\alpha}(Y, D, Z) = \cup_B \{C_{1-\alpha}(Y, D, Z, B) \mid B \subseteq I, c(B) = L - U + 1\} \tag{3.4}$$

The proposed confidence interval $C_{1-\alpha}(Y, D, Z)$ is simple and general; for any test statistic $T(\beta_0, B)$ with a valid size for $B \subseteq B^*$, one simply takes unions of confidence intervals of $T(\beta_0, B)$ over subsets of instruments B where $c(B) = L - U + 1$. In addition, a key feature of our procedure is that it is not necessary to go through all the subsets of possible valid instruments larger than $c(B) > L - U$; simply looking at the smallest possible subsets of valid instruments, i.e. those subsets that are at the lower boundary of $L - U$, $c(B) = L - U + 1$, is sufficient to provide the $1 - \alpha$ coverage.

Theorem 3.1 states that the procedure in (3.4) produces a valid confidence interval since $c(B^*) > L - U$, there is some subset of valid instruments with cardinality $L - U + 1$.

Theorem 3.1. *Suppose model* (3.2) *holds and* $s < U$. *Given* α, *consider any test statistic* $T(\beta_0, B)$ *with the property that for any* $B \subseteq B^*$, $T(\beta_0, B)$ *has size at most* α *under the null hypothesis* $H_0 : \beta^* = \beta_0$. *Then,* $C_{1-\alpha}(Y, D, Z)$ *in* (3.4) *always has at least* $1 - \alpha$ *coverage.*

Proof of Theorem 3.1. By $s < U$, we have $c(B^*) > L - U$. Consequently, there is a subset $\tilde{B} \subseteq B^*$ where $c(\tilde{B}) = L - U + 1$ and $\tilde{B}$ only contains only valid instruments. Since $\tilde{B}$ only contains valid instruments, $\text{pr}\{\beta^* \in C_{1-\alpha}(Y, D, Z, \tilde{B})\} \geq 1 - \alpha$ for all π^*, β^*. Hence, we have

$$\text{pr}\{\beta^* \in C_{1-\alpha}(Y, D, Z)\} \geq \text{pr}\{\beta^* \in C_{1-\alpha}(Y, D, Z, \tilde{B})\} \geq 1 - \alpha$$

for all values of π^*, β^*. □

A potential caveat to our procedure is computational feasibility. Even though we restrict the union to subsets of exactly size $c(B) = L - U + 1$, if the number of candidate instruments, L, grows, $C(Y, D, Z)$ becomes computationally burdensome. However, in many instrumental variables studies, it is difficult to find good candidate instruments and rarely the number of these candidates instruments exceed $L = 20$, which modern computing can handle. Hence, our procedure in (3.4) is computationally tractable for most practical applications.

3.2.4. Choice of Test Statistics

In the instrumental variables literature, there are many tests of causal effects $T(\beta_0, B)$ that can be used with Theorem 3.1 to construct valid $1 - \alpha$ confidence interval $C_{1-\alpha}(Y, D, Z)$ in the presence of invalid instruments. A natural question to ask, then, is among these tests, which test statistic, when used with Theorem 3.1, provides the smallest length confidence interval and thus, from a practical standpoint, provides the most informative confidence interval?

The most popular test is the t-test based on based on the asymptotic normal distribution of the two stage least squares estimator. The two stage least squares estimator of β^* for a given B, denoted as $\hat{\beta}_{B,TSLS}$, is the solution to the minimization problem $\|P_{\tilde{Z}_B} R_{Z_A}(Y - D\beta)\|_2^2$ where A is the complement of the set B, $A = I \setminus B$, and $\tilde{Z}_B = R_{Z_A} Z_B$. If $\hat{u}(B)$ is the residuals from the fitted model, $\hat{u}(B) = R_{Z_A}(Y - D\hat{\beta}_{B,TSLS})$ and $\hat{D}(B)$ is the projection of D on to the column space of $\tilde{Z}_B$, $\hat{D}(B) = P_{\tilde{Z}_B} D$, then the t-test is defined as

$$TSLS(\beta_0, B) = \sqrt{n - c(A) - 1} \left\{ \frac{\hat{\beta}_{B,TSLS} - \beta_0^*}{\sqrt{\|\hat{u}(B)\|_2^2 / \|\hat{D}(B)\|_2^2}} \right\} \tag{3.5}$$

If $B \subseteq B^*$, standard econometrics arguments show that (3.5) converges to an asymptotic Normal distribution (Wooldridge, 2010). In practice, the test (3.5) is approximately valid when all the subset of instruments B among the candidate instruments I are strong, or in other words, strongly associated with the exposure. Unfortunately, instruments can be weak in practice and the nominal size of tests based on two stage least squares can be misleading (Staiger and Stock, 1997).

Stock et al. (2002) presents a survey of tests that are robust to weak instruments. Specifically, for a given B, let $W(B)$ be an n by 2 matrix where the first column contains $R_{Z_A}Y$ and the second column contains $R_{Z_A}D$. Let $a_0 = (\beta_0, 1)$ and $b_0 = (1, -\beta_0)$ to be two-dimensional vectors and $\hat{\Sigma} = W(B)^T M_{\tilde{Z}_B} W(B)/(n-L)$. Let $\hat{S}(B)$ and $\hat{T}(B)$ be two-dimensional vectors

$$\hat{S}(B) = \frac{(\tilde{Z}_B^T \tilde{Z}_B)^{-1/2} \tilde{Z}_B^T W(B) b_0}{\sqrt{b_0^T \hat{\Sigma} b_0}}, \quad \hat{T}(B) = \frac{(\tilde{Z}_B^T \tilde{Z}_B)^{-1/2} \tilde{Z}_B^T W(B) \hat{\Sigma}^{-1} a_0}{\sqrt{a_0^T \hat{\Sigma}^{-1} a_0}}$$

along with the following scalar values

$$\begin{aligned} &\hat{Q}_{11}(B) = \hat{S}(B)^T \hat{S}(B), \quad \hat{Q}_{12}(B) = \hat{S}(B)^T \hat{T}(B) \\ &\hat{Q}_{22}(B) = \hat{T}(B)^T \hat{T}(B) \end{aligned}$$

Based on $\hat{Q}_{11}(B)$, $\hat{Q}_{12}(B)$, and $\hat{Q}_{22}(B)$, we define the following tests, the Anderson-Rubin test (Anderson and Rubin, 1949), the Lagrangian multiplier test (Kleibergen, 2002), and the conditional likelihood test (Moreira, 2003).

$$AR(\beta_0, B) = \hat{Q}_{11}(B)/c(B) \tag{3.6}$$

$$LM(\beta_0, B) = \hat{Q}_{12}^2(B)/\hat{Q}_{22}(B) \tag{3.7}$$

$$\begin{aligned} CLR(\beta_0, B) = &\frac{1}{2}\left\{\hat{Q}_{11}(B) - \hat{Q}_{22}(B)\right\} \\ &+ \frac{1}{2}\sqrt{\{\hat{Q}_{11}(B) + \hat{Q}_{22}(B)\}^2 - 4\{\hat{Q}_{11}(B)\hat{Q}_{22} - \hat{Q}_{12}^2(B)\}} \end{aligned} \tag{3.8}$$

Each of the three tests have their unique robustness characteristics and properties, but all of them have been shown to be robust to weak instruments(Staiger and Stock, 1997; Stock et al., 2002; Kleibergen, 2002; Moreira, 2003; Dufour, 2003; Andrews et al., 2006). There is no uniformly most powerful test among the three tests, but Andrews et al. (2006) and Mikusheva (2010) have suggested using (3.8) due to its generally favorable power compared to (3.6) and (3.7) in most cases when weak instruments are present. However, the Lagrangian multiplier test (3.7) and the Anderson-Rubin test (3.6) have the unique feature where both tests (or derivatives of) can be used as a pretest to check whether the candidate subset of instruments B contain only valid instruments. This feature is particularly useful for our problem where we have possibly invalid instruments (see Section 3.2.6). Also, among the three tests, the Anderson-Rubin test is the simplest in that it can be written as a standard F -test in regression where the outcome is $R_{Z_A}(Y - D\beta_0)$, the regressors are $\tilde{Z}_B$, and we are testing whether the coefficients associated with $\tilde{Z}_B$ are zero or not with the standard F-test. Finally, the Lagrangian multiplier test and the conditional likelihood ratio test require an assumption that the exposure, D, is linearly related to the exposure Z by $D_i = Z_{i.}^T \gamma^* + \xi_i$ where γ^* is an L dimensional vector and ξ_i is a random error term with mean zero, homoscedastic variance, and is independent of Z; the Anderson-Rubin test does not require this linearity assumption.

3.2.5. Empty Confidence Intervals and the Anderson-Rubin test

Our procedure $C_{1-\alpha}(Y, D, Z)$ involves taking the union of confidence intervals, at least one of which is based on valid instruments, but some of which may be based on invalid instruments. For instance, if we have a subset B with $c(B) = L - U + 1$, but it is not a the subset of B^*, B contains at least one true invalid instruments from $A^* = I \setminus B^*$ and we may end up with confidence intervals $C_{1-\alpha}(Y, D, Z, B)$ that are biased. Such a potentially biased interval is included in the interval for $C_{1-\alpha}(Y, D, Z)$. Even though $C_{1-\alpha}(Y, D, Z)$ will have correct coverage regardless of this inclusion, the unnecessary inclusion of the biased interval may elongate the interval $C_{1-\alpha}(Y, D, Z)$ and produce an uninformative interval.

One method to deal with this problem is to choose a test statistic where for $B \not\subseteq B^*$, $C_{1-\alpha}(Y, D, Z, B)$ will usually produce an empty interval. For example, the Anderson-Rubin test statistic in (3.6) has this feature. To illustrate, suppose we assume $D_i = Z_i^T \gamma^* + \xi_i$ where ϵ_i, ξ_i are independent and bivariate Normal with mean 0 and covariance Σ. If we subtract $D\beta_0$ and multiply by R_{Z_A} from both sides of (3.2) and substitute D_i with $D_i = Z_i^T \gamma^* + \xi_i$, we obtain

$$R_{Z_A}(Y - D\beta_0) = \tilde{Z}_B \kappa^* + R_{Z_A}\epsilon, \quad \kappa = \pi_B^* + \gamma_B(\beta^* - \beta_0) \tag{3.9}$$

where γ_B^* and π_B^* are the components of γ^* and α^* vectors for the indices that belong to the subset B. As explained in Section 3.2.4, the Anderson–Rubin test can also be written as an F test where the null is $H_0 : \kappa^* = 0$. This null corresponds to testing both whether the instruments are valid, $\pi_B^* = 0$, and whether the treatment effect is β_0, $\beta^* = \beta_0$. Rejecting $H_0 : \kappa^* = 0$ in favor of the alternative would imply that one is rejecting the null because the treatment effect is not β_0 or because the instruments in set B are not valid. Thus, when a candidate set of instruments B contains an invalid instrument, the Anderson–Rubin test will likely reject when $\beta^* = \beta_0$, and so the inversion of the Anderson–Rubin test will produce an empty confidence interval or a short confidence interval (see Kadane and Anderson (1977) and Small (2007) for the exact circumstances under which the Anderson–Rubin test will have this property).

3.2.6. Pretest for Invalid Instruments

Another method to avoid taking unions of unnecessary intervals is by conducting a preliminary test that checks whether each of the subsets B where $c(B) = L - U + 1$ contains any invalid instruments before proceeding to construct a confidence interval with B. Specifically, for a desired confidence interval $1 - \alpha$, consider the null hypothesis that B contains only valid instruments, $\pi_B^* = 0$, and the corresponding test statistic $S(B)$, which serve as a pretest for the validity of the instruments in B. For any $\alpha_1 < \alpha$, suppose the test based on $S(B)$ has level α_1 under the null hypothesis that B only contains valid instruments with $q_{1-\alpha_1}$ as the $1 - \alpha_1$ quantile of $S(B)$ under the null hypothesis. Then, a $1 - \alpha$ confidence interval can be constructed based on $S(B)$ as follows.

$$P_{1-\alpha}(Y, D, Z) = \cup_B \{C_{1-\alpha_2}(Y, D, Z, B) \mid B \subseteq I, c(B) = L - U + 1, S(B) \leq q_{1-\alpha_1}\} \tag{3.10}$$

where $\alpha = \alpha_1 + \alpha_2$. For example, if the desired confidence level is 95% where $\alpha = 0 \cdot 05$, we can set $\alpha_1 = 0.01$ and $\alpha_2 = 0.04$.

Given α, α_1, and α_2 where $\alpha = \alpha_1 + \alpha_2$, Theorem 3.2 shows that $\tilde{C}_{1-\alpha}$ achieves the desired $1-\alpha$ coverage in the presence of possibly invalid instruments.

Theorem 3.2. *Suppose we have the same assumptions about the model and the test statistic $T(\beta_0, B)$ as in Theorem 3.1. For any pretest $S(B)$ that has the correct size under the null hypothesis that B contains only valid instruments, $P_{1-\alpha}(Y, D, Z)$ always has at least $1-\alpha$ coverage.*

Proof of Theorem 3.2. Consider $\tilde{B} \subseteq B^*$ where $c(\tilde{B}) = L - U + 1$. Since $S(B)$ has the correct size, under the null hypothesis, $\text{pr}\{S(\tilde{B}) \geq q_{1-\alpha_1}\} \leq \alpha_1$. Then, for $S(B)$ and $T(\beta_0, B)$, we can use Bonferroni's inequality to obtain

$$\begin{aligned}\text{pr}\{\beta^* \in P_{1-\alpha}(Y, D, Z)\} &\geq \text{pr}\{\beta^* \in C_{1-\alpha_2}(Y, D, Z, \tilde{B}) \cap S(\tilde{B}) \leq q_{1-\alpha_1}\} \\ &\geq 1 - \text{pr}\{\beta^* \notin C_{1-\alpha_2}(Y, D, Z, \tilde{B})\} - \text{pr}\{S(\tilde{B}) \geq q_{1-\alpha_1}\} \\ &= 1 - \alpha_1 - \alpha_2 = 1 - \alpha\end{aligned}$$

thereby guaranteeing the correct coverage. □

Similar to Theorem 3.1, the procedure in (3.10) is general in the sense that any pretest $S(B)$ with the correct size under the null hypothesis that B contains only valid instruments will guarantee that the pretest confidence interval $P_{1-\alpha}(Y, D, Z)$ will have the desired level of coverage. For example, the test statistic proposed by Kleibergen (2007), which is simply the difference between the Anderson-Rubin test in (3.6) and the Lagrangian multiplier test in (3.7)

$$JLM(\beta_0, B) = AR(\beta_0, B) - LM(\beta_0, B) \tag{3.11}$$

satisfies the size criterion in Theorem 3.2 under some assumptions, most notably the linear modeling assumption between D_i and $Z_{i.}$. Furthermore, Kleibergen (2007) proved that under the null hypothesis for $H_0 : \beta^* = \beta_0$, (3.11) is independent of the Lagrangian multiplier test and converges to a $\chi^2_{c(B)-1}$. Hence, we can use the two tests, JLM(β_0, B) and LM(β_0, B), to construct $P_{1-\alpha}(Y, D, Z)$ in (3.10) by first conducting a pretest with JLM(β_0, B) at α_1 level where $q_{1-\alpha_1}$ would be the $1 - \alpha_1$ quantile of $\chi^2_{c(B)-1}$. If the test fails to reject the null hypothesis that B contains only valid instruments, we can then proceed to construct a $1 - \alpha_2$ confidence interval using this B and the Lagrangian multiplier test of LM(β_0, B).

Another pretest that can be used is the Sargan test for overidentification (Sargan, 1958), which tests, among other things, whether the instruments B contain only valid instruments (Dufour, 2003). The Sargan test is

$$SAR(B) = \frac{\|P_{\tilde{Z}_B} \hat{u}(B)\|_2^2}{\|\hat{u}(B)\|_2^2 / n} \tag{3.12}$$

where the $\hat{u}(B)$ corresponds to the residual from the two stage least squares estimator in (3.5). Under model (3.2) and the null hypothesis that Z_B is independent of ϵ_i, SAR(B) converges to a $\chi^2_{c(B)-1}$ distribution. In other words, as long as B contains a set of valid instruments, $S(B)$ converges to a $\chi^2_{c(B)-1}$. Thus, if we use the Sargan test as a pretest for $P_{1-\alpha}(Y, D, Z)$, then $q_{1-\alpha_1}$ in (3.10) would be the $1 - \alpha_1$ quantile of a $\chi^2_{c(B)-1}$ distribution and we would only proceed to construct a confidence interval with the test statistic $T(\beta_0, B)$ at $1 - \alpha_2$ if the null hypothesis is retained.

3.2.7. Prior Information About s and U

Throughout our discussion, we used the $U = L/2$ upper bound, that is given L candidates, less than 50% are invalid, out of simplicity along with the fact that at $U \leq L/2$, the parameters in our model (3.2) are always identifiable (see Chapter 2). However, in practice, practitioners may be able to use their subject matter knowledge to assume a smaller upper bound on the number of invalid instruments and we want to be able to incorporate this information into our confidence interval procedures. By having a tighter upper bound on s by U than $U = L/2$, our methods in (3.4) and (3.10) are only left with smaller number of subsets of possibly valid instruments to go through. Specifically, in (3.4), we take less unions over possibly unnecessary intervals and this provides more informative intervals. In (3.10), having a tighter bound on s translates to doing fewer pretests and having less subsets to take unions of, leading to more informative intervals. In Section 3.3.2, we examine the effect of having more prior information about s via U on our methods producing more informative intervals through a simulation study.

3.3. Simulation

3.3.1. Robustness With Invalid Instruments

We first compare in the simulation study the robustness of our method compared to popular methods for confidence intervals in the instrumental variables literature when there are concerns for invalid instruments.

The simulation setup is similar to the traditional single-equation linear models. We have $n = 5000$ individuals with $L = 10$ candidate instruments where each pair of instruments are correlated with correlation 0.6. For the data generating model, we assume the model in (3.2) and a linear model between D_i and $Z_{i.}$, specifically $D_i = Z_i^T \gamma^* + \xi_i$ where ϵ_i, ξ_i are either (i) independent and bivariate Normal with mean 0, marginal variance 1, and correlation 0.99, (ii) bivariate t with 3 degrees of freedom with the same moments as (i) and (iii) where the log of the error terms is bivariate Normal with the same moments as (i) so that the error distributions are skewed. The individuals $i = 1, \ldots, n$ are independent. We vary the number of invalid instruments, s, from 0 to 5. We consider the setting where less than 50% of the instruments are invalid since β^* is always identified under this case (see Chapter 2). We set γ^* based on the concentration parameter, which is the expected value of the F statistic for the coefficients Z_{B^*} in the regression of D and Z and is a measure of instrument strength (Stock et al., 2002). Specifically, γ^* is set so that either (i) the instruments are strong with a concentration parameter above 1000 or (ii) the instruments are weak with a concentration parameter below 10.

We compare our methods in (3.4) and (3.10) to "naive" and "oracles" methods. Naive methods are methods that assume all candidate instruments are valid, which is typically done in practice; we use the four tests described in Section 3.2.4, specifically the two-stage least squares test in (3.5), the Anderson-Rubin test in (3.6), the Lagrange multiplier test in (3.7), and the conditional likelihood ratio test in (3.8), all with $B = \{1, \ldots, L\}$ (Murray, 2006). Oracles correspond to knowing exactly which instruments are valid and invalid, specifically using the four procedures with $B = B^*$; these methods typically cannot be used in practice because of the incomplete knowledge about exactly which instruments are invalid versus valid. Also, for our methods involving pretests in (3.10), we use the Sargan test as the pretests for the two stage least squares test and the conditional likelihood ratio test, both at level $\alpha_1 = 0.01$ for the pretest, and $\alpha_2 = 0.04$ for the subsequent tests. For the Lagrange multiplier test, we use the pretest in (3.11) at level $\alpha_1 = 0.01$ and construct the confidence interval with $\alpha_2 = 0.04$. We do not use the pretesting method for the Anderson-Rubin test since the test produces informative intervals by encouraging empty intervals for subsets B that contain invalid instruments (see Section 3.2.5). We repeat the simulation 1000 times for each setting. For interpretability, among all our methods, we take the convex hull of the union of confidence intervals to obtain non-disjoint intervals.

Tables 17, 18, and 19 show the coverage proportion of the four procedures when we vary s and assume that at most 50% of the instruments are invalid, $U = L/2 = 5$, for the bivariate Normal, the bivariate t, and the skewed errors, respectively. When there are no invalid instruments, $s = 0$ and the instruments are strong, the naive procedures have the desired 95% coverage. Our methods have higher than 95% coverage because they need to overcompensate to allow for the possibility that not all candidate instruments are valid. When the instruments are weak and there are no invalid instruments, $s = 0$, any procedure using two stage least squares undercovers, which is to be expected from the literature on two stage least squares' poor performance in the presence of weak instruments (see references in Section 3.2.4). As the number of invalid instruments, s, increases, regardless of the strength of the instruments, the naive methods fail to have any coverage. The oracle

Strength	Case	Test	$s = 0$	$s = 1$	$s = 2$	$s = 3$	$s = 4$
Strong	Naive	TSLS	94	0	0	0	0
		AR	95	0	0	0	0
		LM	98	0	0	0	0
		CLR	95	0	0	0	0
	Our method	TSLS	100	100	100	100	96
		AR	100	100	100	100	95
		LM	100	100	100	100	97
		CLR	100	100	100	100	97
		SAR + TSLS	100	100	100	100	94
		JLM + LM	100	100	100	100	92
		SAR + CLR	100	100	100	100	95
	Oracle	TSLS	94	95	94	95	94
		AR	95	96	95	95	95
		LM	98	98	97	97	97
		CLR	95	95	94	95	94
Weak	Naive	TSLS	5	0	0	0	0
		AR	96	0	0	0	0
		LM	98	0	0	0	0
		CLR	98	0	0	0	0
	Our method	TSLS	30	43	39	30	17
		AR	100	100	100	100	96
		LM	100	100	100	100	97
		CLR	100	100	100	100	97
		SAR + TSLS	31	44	41	32	18
		JLM + LM	99	96	92	77	42
		SAR + CLR	100	100	98	91	56
	Oracle	TSLS	5	7	10	13	17
		AR	96	96	96	96	96
		LM	98	97	97	97	97
		CLR	98	97	97	97	97

Table 17: Comparison of Coverage Between 95% IV Confidence Intervals Under Normal Errors. TSLS, two stage least squares; AR, Anderson–Rubin test ; LM, Lagrange multiplier test; CLR, conditional likelihood ratio test; SAR, Sargan test; JLM, pretest in (3.11). There are $L = 10$ candidate instruments and U is set to $L/2 = 5$. Strong instruments correspond to concentration parameter exceeding 100. Weak instruments correspond to concentration parameter value around 2. The standard error for all the coverage proportions do not exceed 2%.

methods have proper coverage, except two stage least squares when the instruments are weak. Our methods have the desired level of coverage, with the coverage level reaching nominal levels when s is at the boundary of $s < U$, i.e. $s = 4$. The only notable exceptions to our methods having correct coverage are in the presence of weak instruments when the two stage least squares t-test is used as test statistics or when pretests are used. This is not surprising because the two stage least squares t-test and Sargan test are known to have actual Type I error rate that can differ greatly from the nominal Type I error rate in the presence of weak instruments (Staiger and Stock, 1997). The simulations suggest that methods with pretests are only useful when the instruments are sufficiently strong. By contrast, our method using the Anderson-Rubin's test is valid regardless of the strength of the instruments.

In short, in the presence of possibly invalid instruments, the naive, popular approach of simply assuming all the instruments are valid would lead to misleading inference. In contrast, our methods, especially the method in (3.4), provide honest coverage regardless of whether instruments are invalid or valid (as long as the number of invalid instruments is less than the assumed upper bound U) and should be used whenever there is concern for possibly invalid instruments. In particular, (3.4) works regardless of the strength of the instruments while our method in (3.10) provides a desired level of coverage so long as the instruments are strong.

3.3.2. Informative Intervals and Median Length

While our methods provide the desired level of coverage, both theoretically and in simulation, it is unclear whether the resulting robust intervals would be informative in terms of not being too long. It is expected that our methods will produce longer confidence intervals than the oracles since the oracles know more about instrument validity than our methods assumes. In this section, we quantify this difference through a simulation study.

The first simulation setup is identical to Section 3.3.1 and we look at the median length

Strength	Case	Test	$s=0$	$s=1$	$s=2$	$s=3$	$s=4$
Strong	Naive	TSLS	95	0	0	0	0
		AR	95	0	0	0	0
		LM	98	0	0	0	0
		CLR	95	0	0	0	0
	Our method	TSLS	100	100	100	100	97
		AR	100	100	100	100	96
		LM	100	100	100	100	98
		CLR	100	100	100	100	98
		SAR + TSLS	100	100	100	100	96
		JLM + LM	100	100	100	100	94
		SAR + CLR	100	100	100	100	96
	Oracle	TSLS	95	95	95	96	96
		AR	95	96	95	96	96
		LM	98	98	97	98	98
		CLR	95	95	96	96	95
Weak	Naive	TSLS	5	0	0	0	0
		AR	96	0	0	0	0
		LM	98	0	0	0	0
		CLR	98	0	0	0	0
	Our method	TSLS	30	45	41	32	16
		AR	100	100	100	100	97
		LM	100	100	100	100	98
		CLR	100	100	100	100	98
		SAR + TSLS	31	47	45	34	17
		JLM + LM	99	95	94	79	45
		SAR + CLR	100	100	98	90	58
	Oracle	TSLS	5	6	8	12	15
		AR	96	96	96	96	96
		LM	98	97	98	97	98
		CLR	98	97	98	97	98

Table 18: Comparison of Coverage Between 95% IV Confidence Intervals Under Bivariate t Errors. TSLS, two stage least squares; AR, Anderson–Rubin test ; LM, Lagrange multiplier test; CLR, conditional likelihood ratio test; SAR, Sargan test; JLM, pretest in (3.11). There are $L = 10$ candidate instruments and U is set to $L/2 = 5$. Strong instruments correspond to concentration parameter exceeding 100. Weak instruments correspond to concentration parameter value around 2. The standard error for all the coverage proportions do not exceed 1%.

Strength	Case	Test	$s = 0$	$s = 1$	$s = 2$	$s = 3$	$s = 4$
Strong	Naive	TSLS	94	0	0	0	0
		AR	95	0	0	0	0
		LM	98	0	0	0	0
		CLR	95	0	0	0	0
	Our method	TSLS	100	100	100	100	95
		AR	100	100	100	100	95
		LM	100	100	100	100	97
		CLR	100	100	100	100	97
		SAR + TSLS	100	100	100	100	94
		JLM + LM	100	100	100	100	92
		SAR + CLR	100	100	100	100	94
	Oracle	TSLS	94	94	94	93	94
		AR	95	95	94	94	95
		LM	98	97	97	97	97
		CLR	95	94	94	94	94
Weak	Naive	TSLS	0	0	0	0	0
		AR	96	45	1	0	0
		LM	98	15	0	0	0
		CLR	97	15	0	0	0
	Our method	TSLS	17	60	60	48	26
		AR	100	100	100	100	99
		LM	100	100	100	100	100
		CLR	100	100	100	100	100
		SAR + TSLS	18	55	56	48	25
		JLM + LM	100	100	99	97	87
		SAR + CLR	100	100	100	100	89
	Oracle	TSLS	0	0	0	1	2
		AR	96	96	96	95	96
		LM	98	97	96	96	96
		CLR	97	97	96	96	96

Table 19: Comparison of Coverage Between 95% IV Confidence Intervals Under Skewed Errors. TSLS, two stage least squares; AR, Anderson–Rubin test ; LM, Lagrange multiplier test; CLR, conditional likelihood ratio test; SAR, Sargan test; JLM, pretest in (3.11). There are $L = 10$ candidate instruments and U is set to $L/2 = 5$. Strong instruments correspond to concentration parameter exceeding 100. Weak instruments correspond to concentration parameter value around 2. The standard error for all the coverage proportions do not exceed 2%.

Strength	Case	Test	$s = 0$	$s = 1$	$s = 2$	$s = 3$	$s = 4$
Strong	Our method	TSLS	$0 \cdot 28$	$0 \cdot 73$	$0 \cdot 59$	$0 \cdot 51$	$0 \cdot 44$
		AR	$0 \cdot 38$	$0 \cdot 22$	$0 \cdot 15$	$0 \cdot 11$	$0 \cdot 07$
		LM	$1 \cdot 18$	$1 \cdot 13$	$1 \cdot 09$	$1 \cdot 07$	$1 \cdot 05$
		CLR	$0 \cdot 29$	$0 \cdot 67$	$0 \cdot 58$	$0 \cdot 50$	$0 \cdot 44$
		SAR + TSLS	$0 \cdot 29$	$0 \cdot 17$	$0 \cdot 12$	$0 \cdot 08$	$0 \cdot 05$
		JLM + LM	$0 \cdot 28$	$0 \cdot 16$	$0 \cdot 11$	$0 \cdot 08$	$0 \cdot 05$
		SAR + CLR	$0 \cdot 29$	$0 \cdot 17$	$0 \cdot 12$	$0 \cdot 08$	$0 \cdot 05$
	Oracle	TSLS	$0 \cdot 04$	$0 \cdot 04$	$0 \cdot 04$	$0 \cdot 05$	$0 \cdot 05$
		AR	$0 \cdot 06$	$0 \cdot 06$	$0 \cdot 06$	$0 \cdot 07$	$0 \cdot 07$
		LM	$1 \cdot 03$	$1 \cdot 03$	$1 \cdot 03$	$1 \cdot 03$	$1 \cdot 04$
		CLR	$0 \cdot 04$	$0 \cdot 04$	$0 \cdot 04$	$0 \cdot 05$	$0 \cdot 05$
Weak	Our method	AR	∞	∞	∞	∞	∞
		LM	∞	∞	∞	∞	∞
		CLR	∞	∞	∞	∞	∞
		JLM + LM	∞	$300 \cdot 12$	$160 \cdot 80$	$115 \cdot 12$	$101 \cdot 89$
		SAR + CLR	∞	∞	∞	∞	$46 \cdot 12$
	Oracle	AR	∞	∞	∞	∞	∞
		LM	$10 \cdot 22$	$18 \cdot 79$	∞	∞	∞
		CLR	$9 \cdot 45$	$17 \cdot 97$	∞	∞	∞

Table 20: Comparison of Median Lengths Between Different 95% IV Confidence Intervals Under Normal Errors. TSLS, two stage least squares; AR, Anderson–Rubin test ; LM, Lagrange multiplier test; CLR, conditional likelihood ratio test; SAR, Sargan test; JLM, pretest in (3.11). There are $L = 10$ candidate instruments and U is set to $L/2 = 5$. Strong instruments correspond to concentration parameter exceeding 100. Weak instruments correspond to concentration parameter value around 2. The interquartile range of our intervals and strong oracle intervals do not exceed $0 \cdot 05$ and $0 \cdot 02$, respectively. The interquartile range of all weak intervals are infinite except for JLM + LM, which range from $1774 \cdot 62$ ($s = 0$) to $55 \cdot 73$ ($s = 4$).

Strength	Case	Test	$s=0$	$s=1$	$s=2$	$s=3$	$s=4$
Strong	Our method	TSLS	$0\cdot28$	$0\cdot73$	$0\cdot58$	$0\cdot50$	$0\cdot44$
		AR	$0\cdot37$	$0\cdot22$	$0\cdot15$	$0\cdot11$	$0\cdot07$
		LM	$1\cdot17$	$1\cdot13$	$1\cdot09$	$1\cdot07$	$1\cdot05$
		CLR	$0\cdot28$	$0\cdot67$	$0\cdot58$	$0\cdot50$	$0\cdot44$
		SAR + TSLS	$0\cdot28$	$0\cdot17$	$0\cdot12$	$0\cdot08$	$0\cdot05$
		JLM + LM	$0\cdot28$	$0\cdot16$	$0\cdot11$	$0\cdot08$	$0\cdot05$
		SAR + CLR	$0\cdot29$	$0\cdot17$	$0\cdot12$	$0\cdot08$	$0\cdot05$
	Oracle	TSLS	$0\cdot04$	$0\cdot04$	$0\cdot04$	$0\cdot05$	$0\cdot05$
		AR	$0\cdot06$	$0\cdot06$	$0\cdot07$	$0\cdot07$	$0\cdot07$
		LM	$1\cdot03$	$1\cdot03$	$1\cdot03$	$1\cdot03$	$1\cdot04$
		CLR	$0\cdot04$	$0\cdot04$	$0\cdot04$	$0\cdot05$	$0\cdot05$
Weak	Our method	AR	∞	∞	∞	∞	∞
		LM	∞	∞	∞	∞	∞
		CLR	∞	∞	∞	∞	∞
		JLM + LM	∞	$282\cdot38$	$163\cdot28$	$113\cdot81$	$101\cdot88$
		SAR + CLR	∞	∞	∞	∞	$46\cdot53$
	Oracle	AR	∞	∞	∞	∞	∞
		LM	$9\cdot40$	$15\cdot34$	$130\cdot38$	∞	∞
		CLR	$8\cdot98$	$14\cdot11$	$167\cdot52$	∞	∞

Table 21: Comparison of Median Lengths Between Different 95% IV Confidence Intervals Under Bivarate t errors. TSLS, two stage least squares; AR, Anderson–Rubin test ; LM, Lagrange multiplier test; CLR, conditional likelihood ratio test; SAR, Sargan test; JLM, pretest in (3.11). There are $L = 10$ candidate instruments and U is set to $L/2 = 5$. Strong instruments correspond to concentration parameter exceeding 100. Weak instruments correspond to concentration parameter value around 2. The interquartile range of our intervals and strong oracle intervals do not exceed $0\cdot05$ and $0\cdot02$, respectively. The interquartile range of all weak intervals are infinite except for JLM + LM, which range from $2684\cdot6$ ($s = 1$) to $58\cdot48$ ($s = 4$).

of the confidence intervals in Table 17. We exclude the naive methods since they do not provide the desired level of coverage. Also, for weak instruments, we exclude two stage least squares since it is not robust to weak instruments and does not provide correct coverage.

In Tables 20 and 21, for both bivariate Normal errors and bivariate t errors, we see that the discrepancy between our method and the oracles shrinks as s grows for strong instruments, especially when $s = 3$ and $s = 4$. The one notable exception is our method using two stage least squares, which still has wide intervals as s increases. We also find that our method using pretests tends to provide the shortest intervals among the various versions of our method under the strong instrument case. This is to be expected since the motivation for the pretesting was to remove taking unnecessary unions of intervals in (3.10). For weak instruments, our method and the oracles are generally in agreement by providing infinite length intervals, with our method almost always producing infinite length intervals. This agreement is to be expected since using tests that are robust to weak instruments must produce infinite intervals (Dufour, 1997).

Table 22 presents the same simulation results as Tables 20 and 21, except the errors are skewed. While the patterns of simulations are mostly the same as the two preceding tables, one notable exception is when the instruments are strong and $s = 0$. In this case, two stage least squares dominates our pretesting method as well as the Anderson and Rubin confidence intervals. Otherwise, the patterns of the simulations are similar across the three tables.

The second simulation study examines the strategy in Section 3.2.7 where prior information on s and U are available and whether the prior information provides informative intervals. The simulation setup is, again, identical as above, except we fix $s = 2$ and vary U from $3, 4$ and 5; if U were to be less than s where $U \leq s$, our methods cannot produce the right coverage since U was mis-specified. We compare our methods to the oracle intervals in Table 20, specifically the column corresponding to $s = 2$.

Strength	Case	Test	$s = 0$	$s = 1$	$s = 2$	$s = 3$	$s = 4$
Strong	Our method	TSLS	$0 \cdot 62$	$0 \cdot 84$	$0 \cdot 66$	$0 \cdot 56$	$0 \cdot 47$
		AR	$0 \cdot 94$	$0 \cdot 50$	$0 \cdot 34$	$0 \cdot 24$	$0 \cdot 16$
		LM	$1 \cdot 46$	$1 \cdot 28$	$1 \cdot 19$	$1 \cdot 14$	$1 \cdot 10$
		CLR	$0 \cdot 67$	$0 \cdot 81$	$0 \cdot 66$	$0 \cdot 56$	$0 \cdot 48$
		SAR + TSLS	$0 \cdot 64$	$0 \cdot 37$	$0 \cdot 25$	$0 \cdot 18$	$0 \cdot 11$
		JLM + LM	$0 \cdot 66$	$0 \cdot 36$	$0 \cdot 24$	$0 \cdot 17$	$0 \cdot 10$
		SAR + CLR	$0 \cdot 69$	$0 \cdot 38$	$0 \cdot 26$	$0 \cdot 18$	$0 \cdot 11$
	Oracle	TSLS	$0 \cdot 08$	$0 \cdot 09$	$0 \cdot 09$	$0 \cdot 10$	$0 \cdot 11$
		AR	$0 \cdot 14$	$0 \cdot 14$	$0 \cdot 15$	$0 \cdot 15$	$0 \cdot 16$
		LM	$1 \cdot 06$	$1 \cdot 06$	$1 \cdot 07$	$1 \cdot 07$	$1 \cdot 07$
		CLR	$0 \cdot 08$	$0 \cdot 09$	$0 \cdot 09$	$0 \cdot 10$	$0 \cdot 11$
Weak	Our method	AR	∞	∞	∞	∞	∞
		LM	∞	∞	∞	∞	∞
		CLR	∞	∞	∞	∞	∞
		JLM + LM	∞	∞	∞	∞	$2220 \cdot 55$
		SAR + CLR	∞	∞	∞	∞	∞
	Oracle	AR	∞	∞	∞	∞	∞
		LM	∞	∞	∞	∞	∞
		CLR	∞	∞	∞	∞	∞

Table 22: Comparison of Median Lengths Between Different 95% IV Confidence Intervals Under Skewed Errors. TSLS, two stage least squares; AR, Anderson–Rubin test ; LM, Lagrange multiplier test; CLR, conditional likelihood ratio test; SAR, Sargan test; JLM, pretest in (3.11). There are $L = 10$ candidate instruments and U is set to $L/2 = 5$. Strong instruments correspond to concentration parameter exceeding 100. Weak instruments correspond to concentration parameter value around 2. The interquartile range of our intervals and strong oracle intervals do not exceed $0 \cdot 20$ and $0 \cdot 05$, respectively. The interquartile range of all weak intervals are infinite except for JLM + LM when $s = 4$ which is $9349 \cdot 87$.

Strength	Case	Test	$U = 3$	$U = 4$	$U = 5$
Strong	Our method	TSLS	$0 \cdot 51$	$0 \cdot 55$	$0 \cdot 59$
		AR	$0 \cdot 07$	$0 \cdot 11$	$0 \cdot 15$
		LM	$1 \cdot 05$	$1 \cdot 07$	$1 \cdot 09$
		CLR	$0 \cdot 50$	$0 \cdot 54$	$0 \cdot 58$
		SAR + TSLS	$0 \cdot 05$	$0 \cdot 08$	$0 \cdot 12$
		JLM + LM	$0 \cdot 04$	$0 \cdot 08$	$0 \cdot 11$
		SAR + CLR	$0 \cdot 04$	$0 \cdot 08$	$0 \cdot 12$
Weak	Our method	AR	∞	∞	∞
		LM	∞	∞	∞
		CLR	∞	∞	∞
		JLM + LM	$102 \cdot 22$	$124 \cdot 24$	$160 \cdot 80$
		SAR + CLR	$59 \cdot 73$	∞	∞

Table 23: Comparison of Median Lengths Between Different 95% IV Confidence Intervals With Prior Information on s and U. TSLS, two stage least squares; AR, Anderson–Rubin test ; LM, Lagrange multiplier test; CLR, conditional likelihood ratio test; SAR, Sargan test; JLM, pretest in (3.11). There are $L = 10$ candidate instruments and U is set to $L/2 = 5$. Strong instruments correspond to concentration parameter exceeding 100. Weak instruments correspond to concentration parameter value around 2. The interquartile range of our intervals do not exceed $0 \cdot 02$. The interquartile range of all weak intervals are infinite except for JLM + LM (for all U) and SAR + CLR ($U = 3$), which range from $160 \cdot 55$ ($U = 3$) to $42 \cdot 75$ ($U = 5$).

Table 23 shows the result from the simulation. We see that if U is close to the true $s = 2$, our interval lengths are very close to the oracle intervals in Table 20 for strong instruments. Again, the notable exception is our method using two stage least squares which produces wide intervals. As U increases, our methods tend to produce longer intervals, which is expected since our prior information about s at $U = 5$ is not as accurate as when $U = 3$. Also, similar to Table 20, our method with pretesting seems to produce the most informative interval compared to our method without pretesting. For weak instruments, our intervals produce the same type of non-informative intervals as the oracle intervals in Table 20. Prior information does not help, perhaps because the instruments are already weak and no extra information can be gained by having more accurate ideas about s.

3.4. Data Analysis

We reanalyze the instrumental variables analysis done in Bouis and Haddad (1990), Bouis and Haddad (1992), and Small (2007) to demonstrate our method in a practical setting. The goal is to study the causal effect of income on food expenditures among Philippine farm households from a survey of $n = 406$ Philippine farm households. The exposure is the household's log income, D_i and the outcome is the household's food expenditures, Y_i. We have four candidate instruments, cultivated area per capita, Z_{i1}, worth of assets, Z_{i2}, a binary dummy variable on presence of electricity at the household, Z_{i3}, and quality of flooring at the house, Z_{i4}. Page 82 of Bouis and Haddad (1990) states that the reasoning behind proposing these variables as instrumental variables is that "land availability is assumed to be a constraint in the short run, and therefore exogenous to the household decision making process". We also control for the measured covariates, which are mother's education, father's education, mother's age, father's age, mother's nutritional knowledge, price of corn, price of rice, population density of the municipality, and number of household members in adult equivalents; see Bouis and Haddad (1990) and Bouis and Haddad (1992) for further details on the data.

The F-statistic for instrument strength is 103·77, indicating reasonably strong instruments.

Case	Test	95% Confidence Interval
Naive	TSLS	($0 \cdot 043, 0 \cdot 053$)
	AR	($0 \cdot 044, 0 \cdot 054$)
	LM	($-0 \cdot 031, 0 \cdot 055$)
	CLR	($0 \cdot 043, 0 \cdot 055$)
Our Method	TSLS	($0 \cdot 031, 0 \cdot 059$)
	AR	($0 \cdot 037, 0 \cdot 058$)
	LM	($-0 \cdot 037, 0 \cdot 067$)
	CLR	($0 \cdot 034, 0 \cdot 066$)
	SAR + TSLS	($0 \cdot 031, 0 \cdot 058$)
	JLM + LM	($0 \cdot 034, 0 \cdot 067$)
	SAR + CLR	($0 \cdot 034, 0 \cdot 066$)

Table 24: Comparison of Median Lengths Between Different 95% IV Confidence Intervals for the Agricultural Data. TSLS, two stage least squares; AR, Anderson–Rubin test ; LM, Lagrange multiplier test; CLR, conditional likelihood ratio test; SAR, Sargan test; JLM, pretest in (3.11). There are four candidate instruments and we assume that at most one is invalid.

The Sargan test for overidentification, which tests assumptions (A2) and (A3), produces a p-value of 0·079. Even though the p-value is low, usually practitioners of the instrumental variables method would naively assume (A2) and (A3) are true since the p-value is above 0·0.5, the typical threshold for significance level and use one of the four procedures in (3.5)–(3.8) to obtain confidence intervals. In contrast, our methods do not take for granted that the four instruments are valid. Instead, we assume there may be invalid instruments, specifically we consider that there may be at most one invalid instruments, which corresponds to $U = L/2 = 2$. The results from both the naive method and our methods are in Table 23. For tests that produced multiple, disjoint intervals, we took the lowermost and uppermost values of all the confidence intervals (i.e. the convex hull) to obtain a non-disjoint confidence interval. Also, for procedures with pretests, we used the same α_1 and α_2 thresholds as we did in Section 3.3.1.

As long as the modeling assumption is true and that no more than one instruments is invalid, we have a theoretical guarantee that our methods provide the correct 95% confidence interval, which cannot be said for the four naive intervals in Table 23. Also, even though our confidence interval is longer than the the naive intervals, it is still informative in the sense that most of our intervals do not contain $\beta^* = 0$ and therefore, the null hypothesis of no causal effect can be rejected at the usual 5% significance level. The notable exception is the confidence interval based on the Lagrange multiplier test without any pretests. For this test, both the naive method and the method based on (3.4) contain zero. Among the intervals that are theoretically guaranteed to have $1 - \alpha$ coverage, our method in (3.4) using the Anderson–Rubin provides the shortest interval.

The data example illustrates the usefulness of our procedure whenever there is a concern for invalid instruments in practice. Our procedures yield confidence intervals that are honest with respect to coverage and can be informative.

3.5. Discussion

This paper proposes a simple and general method to construct robust confidence intervals for causal effects using instrumental variables estimates when the instruments are possibly invalid, with theoretical guarantees with respect to coverage. We propose two methods in (3.4) and (3.10), with the latter using pretests tending to produce informative intervals when the instruments are strong. Our data analysis example illustrates that our method can be a robust alternative to confidence interval estimation that has the proper coverage whenever there is concern for possibly invalid instruments.

CHAPTER 4 : A Nonparametric, Full Matching Approach to Instrumental Variables Estimation

This is joint work with Benno Kreuels, Jürgen May, and Dylan Small.

4.1. Instrumental Variables With Measured Covariates

4.1.1. Two Stage Least Squares (TSLS)

As mentioned before, instrumental variables (IVs) is a popular method to estimate the causal effect of an exposure on the outcome when there is unmeasured confounding, provided that a valid instrument is available (Angrist, Imbens, and Rubin, 1996; Hernán and Robins, 2006; Brookhart and Schneeweiss, 2007; Cheng, Qin, and Zhang, 2009; Swanson and Hernán, 2013; Baiocchi, Cheng, and Small, 2014). The core assumptions for a variable to be a valid instrumental variable are that the variable (A1) is associated with the exposure, (A2) has no direct pathways to the outcome, and (A3) is not associated with any unmeasured confounders. If measured covariates are available, which is frequently the case in many IV studies, the plausibility of the instrument satisfying the three core assumptions can be improved, especially (A3), by conditioning on the covariates.

The most popular and well-studied method that use an IV and measured covariates to estimate causal effects is two stage least squares (TSLS) (Angrist and Krueger, 1991; Card, 1995; Wooldridge, 2010). For example, in Card (1995), which studied the effect of education on wages, TSLS with proximity to a 4-year college as an IV was used to control for measured covariates such as race and parents' education. Specifically, TSLS first estimates, via least squares, the predicted exposure (education) given the instrument, (proximity to 4-year college) and the measured covariates, and second, regresses the outcome (earnings) on this predicted exposure and the measured covariates; the TSLS estimate of the causal effect is the coefficient on the predicted exposure in the second regression. Standard results in econometrics show TSLS estimator is consistent and efficient under linear single-variable structural equation models with a constant treatment effect (Wooldridge, 2010). When treatment effects are not constant, Angrist and Imbens (1995) showed that under certain monotonicity assumptions, TSLS converges to a weighted average of the covariate-specific treatment effects with the weights proportional to the average conditional variance of the expected value of the treatment given the covariates and the instrument. Other IV methods to estimate causal effects in the presence of measured covariates include Bayesian methods (Imbens and Rubin, 1997), semiparametric methods (Abadie, 2003; Tan, 2006; Ogburn et al., 2015), and nonparametric methods (Frölich, 2007).

Despite its attractive estimation properties, TSLS has some drawbacks, specifically in (i) lack of transparency of the population to which the estimate applies, (ii) lack of blinding of the analyst/researcher and (iii) dependence on parametric assumptions. First, with regards to transparency, suppose that there are some values of the covariates for which the instrument is almost always low, some values for which the instrument is almost always high and some values of the covariates for which the instrument takes on both low and high values. Then, the TSLS estimate will put most of its weight on the causal effect for subjects with the values of the covariates for which the instrument takes on both low and high values, and little weight on subjects with the values of the covariates for which the instrument usually takes on low (or high) values. For example, in the case of education on earnings, this would mean that there might be some states (a measured covariate) that are receiving little weight in the TSLS estimate; consequently, the TSLS estimate might not be helpful for understanding the effect of education on earnings in some states even though these states might have contributed many subjects to the analysis. Although the weighting function in TSLS can be studied, there is nothing in the TSLS estimation procedure itself that warns us when some values of the covariates are receiving little weight and it is rare to see discussion of the weighting function for TSLS in empirical papers.

Second, TSLS lacks blinding with respect to the outcome data when adjusting for covariates. Cochran (1965), Rubin (2007) and Rosenbaum (2010) argue that the best observational studies resemble randomized experiments. An important feature of the design of randomized experiments is that when designing the study and planning the analysis, the researcher is blinded to the outcome data. However, in regression based procedures for adjusting for covariates like TSLS, there is often judgment that needs to be exercised in choosing covariate adjustment models and this requires one to look at the outcome data along with estimates of causal effects. It is difficult even for the most honest researcher to be completely objective in comparing models when the researcher has an a priori hypothesis or expectation about the direction of the causal effect (Rubin and Waterman, 2006).

Third, TSLS relies on proper specification of how the measured covariates affect the outcome. Often, parametric modeling assumptions are made for how the measured confounders affect the outcome. In particular, TSLS, as usually implemented, relies on the measured confounders having a linear effect on the expected outcome.

4.1.2. Instrumental Variables With Full Matching

Matching is an alternative method to adjust for measured covariates. A matching algorithm groups individuals in the data with different values of the instrument but similar values of the observed covariates, so that within each group, the only difference between the individuals is their values of the instrument (Haviland, Nagin, and Rosenbaum, 2007; Rosenbaum, 2010; Stuart, 2010). We can then compare the outcome between individuals with high and low values of the instrument within a matched set to assess the causal effect of the exposure on the outcome (Baiocchi et al., 2010).

Matching addresses the drawbacks of TSLS discussed in the previous section. First, if there are values of covariates for which almost all subjects have a high (or low) value of the IV, then the matching algorithm and associated diagnostics will tell us that matched sets cannot be formed when subjects in the matched sets have certain values of the covariates but different levels of the IV; thus, it will be transparent that for these values of the covariates, the causal effect cannot be estimated without extrapolation. Relatedly, matching allows us to control the weighting of subjects with different values of the covariates to make the weighting transparent, such as weighting the covariates in proportion to their population frequency. Second, matching is blind to the outcome data; a matching algorithm only requires the measured covariates and the instrument values for each individual in the data. Diagnostics can be done and the matching can be adjusted until it is adequate, all without looking at the outcome data. Finally, when estimating the causal effect, matching makes non-parametric inference; it does not use any parametric modeling assumptions such as linearity.

Previous work using matching in studying causality is abundant in non-IV settings; see Stuart (2010) for a complete overview. In contrast, work on using matching methods on IV estimation is limited to pair matching (Baiocchi et al., 2010) and fixed control matching, i.e. each unit with level 1 of the IV is matched to a fixed number of units with level 0 of the IV (Kang et al. (2013)). A drawback to these matching methods is that they do not use the full data (Keele and Morgan, 2013; Zubizarreta et al., 2013). In particular, the method in Kang et al. (2013) was limited to matching with fixed controls and the method had to drop roughly 25% of individuals in the final statistical inference from a total of 884 individuals.

In this paper, we develop an IV full matching approach that uses the full data. Full matching is the most general, flexible, and optimal type of matching (Rosenbaum, 1991; Hansen, 2004; Rosenbaum, 2010). Specifically, full matching is the generalization of any type of matching, such as pair matching, matching with fixed controls, or matching with variable controls. Full matching is also flexible in that it can incorporate constraints on matched set structures, such as limiting the number of individuals in each matched set, to improve statistical efficiency. Finally, full matching is optimal in the sense that it produces matched sets where within each set, measured covariates between individuals with different instrument values are most similar (Rosenbaum, 1991).

Under IV estimation with full matching, we derive a randomization-based testing procedure and sensitivity analysis based on the proposed test statistic. We conduct simulation studies to study the performance of TSLS versus full matching IV estimation, specifically analyzing the robustness of both methods to non-linearity (see Section 4.3.1).

4.2. Method

4.2.1. Notation

To introduce the idea of matching in IV estimation, we introduce the following notation. Let $i = 1, \dots, I$ index the I total matched sets that individuals are matched into. Each matched set i contains $n_i \geq 2$ subjects who are indexed by $j = 1, \dots, n_i$ and there are a total of $N = \sum_{i=1}^{I} n_i$ individuals in the data. Let Z_{ij} denote a binary instrument for subject j in matched set i. In each matched set i, there are m_i subjects with $Z_{ij} = 1$ and $n_i - m_i$ subjects with $Z_{ij} = 0$. Let $\mathbf{Z}$ be a random variable that consists of the collection of Z_{ij}'s, $\mathbf{Z} = (Z_{11}, Z_{12}, ..., Z_{I,n_I})$. Define Ω as the set that contains all possible values $\mathbf{z}$ of $\mathbf{Z}$, so $\mathbf{z} \in \Omega$ if z_{ij} is binary and $\sum_{j=1}^{n_i} z_{ij} = m_i$ for all I matched sets. Thus, the cardinality of Ω, denoted as $|\Omega|$, is $|\Omega| = \prod_{i=1}^{I} \binom{n_i}{m_i}$. Denote $\mathcal{Z}$ to be the event that $\mathbf{Z} \in \Omega$. Also define $\chi(\cdot)$ to be an indicator function.

For individual j in matched set i, define d_{1ij} and d_{0ij} to be the potential exposure values under $Z_{ij} = 1$ or $Z_{ij} = 0$, respectively. Also, define $r_{1ij}^{(k)}$ to be the outcome individual j would have if she were assigned instrument value 1 and level k of the exposure, and $r_{0ij}^{(k)}$ to be the outcome individual j would have if she were assigned instrumental value 0 and level k of the exposure. Then, $r_{1ij}^{(d_{1ij})}$ and $r_{0ij}^{(d_{0ij})}$ are the potential outcomes if the individual were assigned levels 1 and 0 of the instrument, respectively, and the exposure took its natural levels given the instrument resulting in d_{1ij} and d_{0ij}, respectively. The potential outcome notations assume the Stable Unit Treatment Value Assumption that an individual's outcome and exposure depend only on her own value of the instrument and not on other people's instrument values (Rubin, 1980).

For individual j in matched set i, let R_{ij} be the binary observed outcome and D_{ij} be the observed exposure. The potential outcomes $r_{1ij}^{(d_{1ij})}, r_{0ij}^{(d_{0ij})}, d_{1ij}$, and d_{0ij} and the observed values R_{ij}, D_{ij}, and Z_{ij} are related by the following equation:

$$R_{ij} = r_{1ij}^{(d_{1ij})} Z_{ij} + r_{0ij}^{(d_{0ij})}(1 - Z_{ij}) \qquad D_{ij} = d_{1ij} Z_{ij} + d_{0ij}(1 - Z_{ij}) \tag{4.1}$$

For individual j in matched set i, let $\mathbf{X}_{ij}$ be a vector of observed covariates and u_{ij} be the unobserved covariates. We define the set $\mathcal{F} = \{(r_{1ij}^{(d_{1ij})}, r_{0ij}^{(d_{0ij})}, d_{1ij}, d_{0ij}, \mathbf{X}_{ij}, u_{ij}), i = 1, ..., I, j = 1, ..., n_i\}$ to be the collection of potential outcomes and all covariates/confounders, observed and unobserved.

4.2.2. Full Matching Algorithm

A matching algorithm controls the bias resulting from different observed covariates by creating I matched sets indexed by i, $i = 1, \ldots, I$ such that individuals within each matched set have similar covariate values $\mathbf{x}_{ij}$ and the only difference between individuals in each matched set is their instrument values, Z_{ij}. In a full matching algorithm, each matched set i either contains $m_i = 1$ individual with $Z_{ij} = 1$ and $n_i - 1$ individuals with $Z_{ij} = 0$ or $m_i = n_i - 1$ individuals with $Z_{ij} = 1$ and 1 individual with $Z_{ij} = 0$.

Rosenbaum (2002, 2010), Hansen (2004), and Stuart (2010) provide an overview of matching and a discussion on various distance metrics and tools to measure similarity for observed and missing covariates. Once we have obtained the distance matrix, we use an R package available on CRAN called *optmatch* developed by Hansen and Klopfer (2006) to find the optimal full matching.

4.2.3. Definition of a Valid Instrument

Using the notation in 4.2.1, we formalize the core assumptions of an instrumental variable below (Holland, 1988; Angrist et al., 1996; Yang et al., 2014).

(A1) The instrument must be associated with the exposure, or in $\mathcal{F}$, $\sum_{i=1}^{I} \sum_{j=1}^{n_i} (d_{1ij} - d_{0ij}) \neq 0$

(A2) The instrument can only affect the outcome if it affects the exposure, or in $\mathcal{F}$, $r_{1ij}^{(k)} = r_{0ij}^{(k)} \equiv r_{ij}^{(k)}$ for all k, where the last equality drops the r's dependence on Z_{ij} (exclusion restriction)

(A3) The instrument is effectively randomly assigned within a matched set, $P(Z_{ij} = 1|\mathcal{F}, \mathcal{Z}) = m_i/n_i$ for each i.

One assumption worth mentioning within the context of observed covariates is assumption (A3). Assumption (A3) is more plausible if we control for observed variables. Specifically, within the framework of full matching, for each matched set i, if the observed variables $\mathbf{x}_{ij}$ are similar among all n_i individuals, it may be more plausible that the unobserved variable u_{ij} plays no role in the distribution of Z_{ij} among the n_i individuals. If (A3) exactly holds and subjects are exactly matched for X_{ij}, then within each matched set i, Z_{ij} is simply a result of random assignment where $Z_{ij} = 1$ with probability m_i/n_i and $Z_{ij} = 0$ with probability $(n_i - m_i)/n_i$ when we condition on the number of units in the matched sets with $Z_{ij} = 1$ being m_i. In Section 4.2.7, we discuss a sensitivity analysis that allows for the possibility that even after matching for observed variables, an unobserved variable u_{ij} may still influence the assignment of Z_{ij} in each matched set i, meaning that assumption (A3) is violated.

There are also other assumptions associated with instrumental variables, most notably the Stable Unit Treatment Value Assumption (SUTVA) in Section 4.2.1 and the monotonicity assumption in Angrist et al. (1996). SUTVA, within the framework of MR, states that one's individual potential outcomes are not affected by the genotype assignment of another individual. This is fairly reasonable in MR since the instrument was determined at the conception of the child and a child's genotype only affects his exposure and outcome, and not the exposures and outcomes of other children.

Monotonicity, within the framework of MR, states that there are no individuals who would have an adverse effect on the exposure from inheriting the genotype which is purported to bring positive effect on the exposure. In MR where the chosen genetic instruments usually bring about a positive effect on the exposure, monotonicity is reasonable (see Chapter 5 for an example with malaria and stunted growth in children).

4.2.4. Effect Ratio and the Local Average Treatment Effect

We define the parameter of interest, called the *effect ratio*, which is a parameter of the finite population of $N = \sum_{i=1}^{I} n_i$ individuals characterized by $\mathcal{F}$.

$$\lambda = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} r_{1ij}^{d_{1ij}} - r_{0ij}^{d_{0ij}}}{\sum_{i=1}^{I} \sum_{j=1}^{n_i} d_{1ij} - d_{0ij}} \tag{4.2}$$

The effect ratio is the change in the outcome caused by the instrument divided by the change in the exposure caused by the instrument. The effect ratio can be identified by taking the ratio of the differences in expected values.

$$\lambda = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} E(R_{ij} | Z_{ij} = 1, \mathcal{F}, \mathcal{Z}) - E(R_{ij} | Z_{ij} = 0, \mathcal{F}, \mathcal{Z})}{\sum_{i=1}^{I} \sum_{j=1}^{n_i} E(D_{ij} | Z_{ij} = 1, \mathcal{F}, \mathcal{Z}) - E(D_{ij} | Z_{ij} = 0, \mathcal{F}, \mathcal{Z})} \tag{4.3}$$

The effect ratio also admits a well-known interpretation in IV literature if all the IV assumptions, (A1)-(A3), and the monotonicity assumption whereby $d_{1ij} \geq d_{0ij}$ for every i, j in $\mathcal{F}$, are satisfied. Specifically, suppose d_{1ij} and d_{0ij} are discrete values from 0 to M. Then Theorem 4.1 shows that we can identify the effect ratio and interpret it as the weighted average of the unit causal effect of the exposure on the treatment among individuals whose exposure was affected by the instrument

Theorem 4.1. *Suppose the IV assumptions, (A1)-(A3), in Section 4.2.4 holds and the exposure ranges from* $0, 1, 2, \ldots, M$ *where* M *is an integer. Further suppose that the mono-*

tonicity assumption where $d_{1ij} \geq d_{0ij}$ holds for all i, j. Then,

$$\begin{aligned}
\lambda =& \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i} E(R_{ij}|Z_{ij}=1,\mathcal{F},\mathcal{Z}) - E(R_{ij}|Z_{ij}=0,\mathcal{F},\mathcal{Z})}{\sum_{i=1}^{I}\sum_{j=1}^{n_i} E(D_{ij}|Z_{ij}=1,\mathcal{F},\mathcal{Z}) - E(D_{ij}|Z_{ij}=0,\mathcal{F},\mathcal{Z})} \\
=& \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i}\sum_{k=1}^{M}(r_{ij}^{(k)} - r_{ij}^{(k-1)})\chi(d_{1ij} \geq k > d_{0ij})}{\sum_{i=1}^{I}\sum_{j=1}^{n_i}\sum_{k=1}^{M}\chi(d_{1ij} \geq k > d_{0ij})} \\
=& \sum_{i=1}^{I}\sum_{j=1}^{n_i}\sum_{k=1}^{M}(r_{ij}^{(k)} - r_{ij}^{(k-1)})w_{ijk}
\end{aligned}$$

where

$$w_{ijk} = \frac{\chi(d_{1ij} \geq k > d_{0ij})}{\sum_{i=1}^{I}\sum_{j=1}^{n_i}\sum_{l=1}^{M}\chi(d_{1ij} \geq l > d_{0ij})}$$

In words, with the IV assumptions and the monotonicity assumption, Theorem 4.1 states that the effect ratio can be interpreted as the weighted average of the causal effect of a one unit change in the exposure among the individuals in the study population whose exposure would be affected by a change in the instrument. Each weight w_{ijk} represents whether an ijth individual exposure would be moved from below k to at or above k by the instrument, relative to the number of people in the study population whose exposure would be changed by the instrument. The interpretation of λ is akin to Theorem 1 in Angrist and Imbens (1995), except that our result is for the finite-sample case and is specific to matching.

Also, with regards to identification, technically speaking, only assumptions (A1) and (A3) are necessary to identify the 'bare-bone' interpretation of λ in (4.2), the ratio of causal effects of the instrument on the outcome (numerator) and on the exposure (denominator) since the numerator and the denominator can both be identified by the differences in expectations in (4.3). However, without (A2), i.e. the exclusion restriction, and the monotonicity assumption, this ratio of differences in expectations in (4.3) cannot identify the weighted average of effects of the exposure described in the above paragraph.

When full matching is used so that all subjects are used in the matching, the effect ratio (4.2) and its equivalent expression in Theorem 4.1 are defined for the whole study population.

Additionally, the effect ratio is invariant to the particular full match it used, e.g. if a different distance between pairs of subjects were used that resulted in a different full match, the effect ratio would remain the same. In fact, one of the advantages of using full matching compared to other matching algorithms that discard some data, such as pair matching, matching with fixed controls, and matching with variable controls, is that full matching estimates the effect ratio (4.2) (or equivalently in Theorem 4.1) for the whole study population whereas for the matching methods that discard data, these methods only estimate (4.2) for the data that was not discarded, making the estimated parameter dependent on the individuals that were discarded from the matching algorithm. In contrast, the full matching algorithm incorporates all the individuals in the data and the effect ratio parameter, specifically the subscripts i, j are meant to count all the individuals in the data. On a related note, the effect ratio (4.2) generalizes previous expressions for the effect ratio with pair matching, $n_i = 2$, by Baiocchi et al. (2010) or matching with fixed controls, $n_i = k$, by Kang et al. (2013) to accommodate full matching.

4.2.5. Inference for Effect Ratio

We would like to conduct the following hypothesis test for the effect ratio λ.

$$H_0 : \lambda = \lambda_0, \quad H_a : \lambda \neq \lambda_0 \tag{4.4}$$

To test the hypothesis in (4.4), we propose the following test statistic

$$T(\lambda_0) = \frac{1}{I} \sum_{i=1}^{I} V_i(\lambda_0) \tag{4.5}$$

where

$$V_i(\lambda_0) = \frac{n_i}{m_i} \sum_{j=1}^{n_i} Z_{ij}(R_{ij} - \lambda_0 D_{ij}) - \frac{n_i}{n_i - m_i} \sum_{j=1}^{n_i} (1 - Z_{ij})(R_{ij} - \lambda_0 D_{ij})$$

and $S^2(\lambda_0)$, the estimator for the variance of the test statistic, $Var\{T(\lambda_0)|\mathcal{F},\mathcal{Z}\}$

$$S^2(\lambda_0) = \frac{1}{I(I-1)}\sum_{i=1}^{I}\{V_i(\lambda_0) - T(\lambda_0)\}^2 \tag{4.6}$$

Each variable $V_i(\lambda_0)$ is the difference in adjusted responses, $R_{ij} - \lambda_0 D_{ij}$, of those individuals with $Z_{ij} = 1$ and those with $Z_{ij} = 0$. Under the null hypothesis in (4.4), these adjusted responses have the same expected value for $Z_{ij} = 1$ and $Z_{ij} = 0$ and thus, deviation of $T(\lambda_0)$ from zero suggests H_0 is not true.

Theorem 4.2 states that under regularity conditions, the asymptotic null distribution of $T(\lambda_0)/S(\lambda_0)$ is standard Normal.

Theorem 4.2. *Assume that for every I, (i) n_i remains bounded and (ii) $\frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{n_i} r_{1ij}^{(d_{1ij})} - r_{0ij}^{(d_{0ij})}$ and $\frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{n_i} d_{1ij} - d_{0ij}$ remains fixed at $\bar{r}$ and $\bar{d} \neq 0$, respectively, so that $\bar{\lambda} = \bar{r}/\bar{d}$. In addition, we assume the following moment conditions*

$$\sum_{i=1}^{I} E\{V_i^4(\bar{\lambda})|\mathcal{F},\mathcal{Z}\} = o(I^2), \quad \limsup_{I\to\infty} \frac{\sum_{i=1}^{I} E|V_i(\bar{\lambda}) - \mu_{i,\bar{\lambda}}|^3}{\left[\sum_{i=1}^{I} Var\{V_i(\bar{\lambda})\}\right]^{3/2}} = 0 \tag{4.7}$$

Then, under the null hypothesis $H_0 : \lambda = \bar{\lambda}$, for all $t > 0$,

$$\limsup_{I\to\infty} P\left\{\frac{T(\bar{\lambda})}{S(\bar{\lambda})} \leq -t|\mathcal{F},\mathcal{Z}\right\} \leq \Phi(-t), \quad \limsup_{I\to\infty} P\left\{\frac{T(\bar{\lambda})}{S(\bar{\lambda})} \geq t|\mathcal{F},\mathcal{Z}\right\} \leq \Phi(-t)$$

where $\Phi(\cdot)$ is the standard normal distribution.

Theorem 4.2 provides a point estimate as well as a confidence interval for the effect ratio. For the point estimate, in the spirit of Hodges and Lehmann (1963), we find the value of λ that maximizes the p-value, Specifically, setting $T(\lambda)/S(\lambda) = 0$ and solving for λ gives an estimate for the effect ratio, $\hat{\lambda}$

$$\hat{\lambda} = \frac{\sum_{i=1}^{I}\frac{n_i^2}{m_i(n_i-m_i)}\sum_{j=1}^{n_i}(Z_{ij} - \bar{Z}_{i.})(R_{ij} - \bar{R}_{i.})}{\sum_{i=1}^{I}\frac{n_i^2}{m_i(n_i-m_i)}\sum_{j=1}^{n_i}(Z_{ij} - \bar{Z}_{i.})(D_{ij} - \bar{D}_{i.})} \tag{4.8}$$

where $\bar{Z}_{i.}$, $\bar{R}_{i.}$, and $\bar{D}_{i.}$ are averages of the instrument, response, and exposure, respectively, within each matched set. For confidence interval estimation, say 95% confidence interval, we can solve the equation $T(\lambda)/S(\lambda) = \pm 1.96$ for λ to get the confidence interval for the effect ratio. A closed form solution for the confidence interval is provided by Corollary 4.1.

Corollary 4.1. *For any value q, the solution to $T(\lambda)/S(\lambda) = q$ is a solution to the quadratic equation $A_2\lambda^2 + A_1\lambda + A_0 = 0$ where*

$$\begin{aligned}
A_2 &= \bar{H}_.^2 - \frac{q^2}{I(I-1)}\sum_{i=1}^{I}(H_i - \bar{H}_.)^2 \\
A_1 &= -2\bar{G}_.\bar{H}_. + \frac{2q^2}{I(I-1)}\left\{\sum_{i=1}^{I}(G_i - \bar{G}_.)(H_i - \bar{H}_.)\right\} \\
A_0 &= \bar{G}_.^2 - \frac{q^2}{I(I-1)}\sum_{i=1}^{I}(G_i - \bar{G}_.)^2
\end{aligned}$$

where

$$\begin{aligned}
G_i &= \frac{n_i^2}{m_i(n_i - m_i)}\sum_{j=1}^{n_i}(Z_{ij} - \bar{Z}_{i.})(R_{ij} - \bar{R}_{i.}) \\
H_i &= \frac{n_i^2}{m_i(n_i - m_i)}\sum_{j=1}^{n_i}(Z_{ij} - \bar{Z}_{i.})(D_{ij} - \bar{D}_{i.}) \\
\bar{Z}_{i.} &= \frac{1}{n_i}\sum_{j=1}^{n_i} Z_{ij}, \quad \bar{D}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} D_{ij}, \quad \bar{R}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} R_{ij} \\
\bar{H}_. &= \frac{1}{I}\sum_{i=1}^{I} H_i, \quad \bar{G}_. = \frac{1}{I}\sum_{i=1}^{I} G_i
\end{aligned}$$

4.2.6. Formula for Efficiency in Instrumental Variables With Full Matching

One of the advantages of full matching is its flexibility to accommodate various sizes of matched sets. All things being equal in terms of covariate balance, we would like an estimator of the effect ratio λ that is as efficient as possible. This is particularly the case with full matching where an unconstrained full matching can create large matched sets which reduces efficiency (Hansen, 2004). However, we can constrain full matching to increase efficiency by restricting matched sets to have a maximum number of controls and/or treated units per matched set (Hansen, 2004). This section studies statistical efficiency of the estimator for λ under different constraints on full matching.

To study the efficiency of the effect ratio estimator for different n_i and m_i, we study a simple version of the structural equations model popular in econometrics and widely used to study the properties of TSLS, the most popular IV estimator (Wooldridge, 2010). Let (R_{ij}, D_{ij}, Z_{ij}) be i.i.d. observations from an infinite population following this model.

$$R_{ij} = \alpha_i + \beta D_{ij} + \epsilon_{ij}, \quad E(\epsilon_{ij}|Z_{ij}) = 0 \tag{4.9}$$

$$D_{ij} = \tau_i + \gamma Z_{ij} + \xi_{ij}, \quad E(\xi_{ij}|Z_{ij}) = 0 \tag{4.10}$$

with the following moment conditions.

$$Var(\epsilon_{ij}|Z_{ij}) = \sigma_{i,R}^2, \quad Var(\xi_{ij}|Z_{ij}) = \sigma_{i,D}^2, \quad E(\epsilon_{ij}\xi_{ij}|Z_{ij}) = \sigma_{i,RD}$$

The parameters $\alpha_i, i = 1, \ldots, I$ measure the effect on the outcome from being in matched set i. The parameter β is the effect of interest, the effect of the exposure on the outcome. Note that the treatment effect in (4.9) is assumed to be homogeneous for everyone, which is not necessary for the analysis of the effect ratio in general. The parameters $\tau_i, i = 1, \ldots, I$ measure the effect on the exposure from being in matched set i. The parameter γ is the effect of the instrument on the exposure. By including α_i and τ_i, the models (4.9) and (4.10) incorporate the matching aspect of IV estimation since each matched set i has effects on R_{ij} and D_{ij} via α_i and τ_i, respectively, that are unique to that matched set.

The effect ratio, λ, is related to parameters found in standard structural equation models in (4.9) and (4.10). To illustrate this, note that the potential outcomes notation can be rewritten under the models (4.9) and (4.10) as follows.

$$R_{ij} = \begin{cases} r_{1ij}^{(d_{1ij})} = \alpha_i + \beta\tau_i + \beta\gamma + \beta\xi_{ij} + \epsilon_{ij} & \text{if } Z_{ij} = 1 \\ r_{0ij}^{(d_{0ij})} = \alpha_i + \beta\tau_i + \beta\xi_{ij} + \epsilon_{ij} & \text{if } Z_{ij} = 0 \end{cases}$$

$$D_{ij} = \begin{cases} d_{1ij} = \tau_i + \gamma + \xi_{ij} & \text{if } Z_{ij} = 1 \\ d_{0ij} = \tau_i + \xi_{ij} & \text{if } Z_{ij} = 0 \end{cases}$$

Then, the effect ratio in (4.2) turns out to be

$$\lambda = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} r_{1ij}^{(d_{1ij})} - r_{0ij}^{(d_{0ij})}}{\sum_{i=1}^{I} \sum_{j=1}^{n_i} d_{1ij} - d_{0ij}} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} \beta\gamma}{\sum_{i=1}^{I} \sum_{j=1}^{n_i} \gamma} = \frac{\beta\gamma}{\gamma} = \beta$$

Hence, $\lambda = \beta$ and because of this equivalence, inferences for the effect ratio is equivalent to inference for β. For example, the parameter β can be estimated by the effect ratio estimator discussed in Section 4.2.5, specifically equation (4.8)

$$\hat{\beta} = \frac{\sum_{i=1}^{I} \frac{n_i^2}{m_i(n_i - m_i)} \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_{i.})(R_{ij} - \bar{R}_{i.})}{\sum_{i=1}^{I} \frac{n_i^2}{m_i(n_i - m_i)} \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_{i.})(D_{ij} - \bar{D}_{i.})}$$

Theorem 4.3 computes the asymptotic variance of $\hat{\beta}$ to study the efficiency of the effect ratio estimator under the models (4.9) and (4.10).

Theorem 4.3. *Suppose we have models* (4.9) *and* (4.10) *with* $\gamma \neq 0$ *and the third moment of* ϵ_{ij} *is bounded for all* i, j*. Define the following variables*

$$J_i = \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_{i.})(\epsilon_{ij} - \bar{\epsilon}_{i.}), \quad H_i = \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_{i.})(D_{ij} - \bar{D}_{i.}), \quad \bar{\epsilon}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} \epsilon_{ij}$$
$$s_I^2 = \sum_{i=1}^{I} \frac{n_i^3}{m_i(n_i - m_i)} \sigma_{i,R}^2$$

Assume that (i) Z_{ij} *are fixed, (ii)* n_i *remain bounded for all* i*, and the following moment* *conditions are met for J_i and H_i*

$$\limsup_{I \to \infty} \frac{1}{s_I^3} \sum_{i=1}^{I} \frac{n_i^6}{m_i^3 (n_i - m_i)^3} E(|J_i|^3) = 0, \qquad \sum_{i=1}^{I} Var\left(\frac{n_i^2}{m_i(n_i - m_i)} H_i^2 \right) = o(I^2)$$

Then, the asymptotic variance of the effect ratio estimator in (4.8) *is*

$$\sqrt{I}(\hat{\beta} - \beta) \to N\left\{ 0, \frac{\left(\lim_{I \to \infty} \frac{s_I}{\sqrt{I}}\right)^2}{\gamma^2 \left(\lim_{I \to \infty} \frac{1}{I} \sum_{i=1}^{I} n_i\right)^2} \right\}$$

Theorem 4.3 provides an easy way to compare between different types of full matching methods and their effect on the estimation of the effect ratio. For example, in the simple case of homoscedastic variance, the approximate variance of $\hat{\lambda}$ is

$$Var(\hat{\lambda}) \approx K \frac{\sum_{i=1}^{I} \frac{n_i^3}{n_i - 1}}{\left(\sum_{i=1}^{I} n_i\right)^2}$$

where K is some constant that depends on the variance of R_{ij} and the strength of the instrument. Since K will be identical for all full matched designs, we can simply look at the quantities to the right of K to tweak our full matching algorithm to produce the most efficient estimator. In Section 4.3.3, we examine this strategy more closely with a simulation study.

4.2.7. Sensitivity Analysis

Sensitivity analysis attempts to measure the influence of unobserved confounders on the inference on λ. In the case of instrumental variables, a sensitivity analysis quantifies how a violation of assumption (A3) in Section 4.2.3 would impact the inference on λ (Rosenbaum, 2002). Specifically, under assumption (A3), the instrument is assumed to be free from unmeasured confounders or free after conditioning on observed confounders via matching. The latter implies that the instruments are assigned randomly, $P(\mathbf{Z} = z | \mathcal{F}, \mathcal{Z}) = (|\Omega|)^{-1}$, i.e. that within each matched set i, $P(Z_{ij} = 1 | \mathcal{F}, \mathcal{Z}) = m_i / n_i$.

However, as discussed in Section 4.2.3, even after matching for observed confounders, unmeasured confounders may influence the viability of assumption (A3). For example, within a matched set i , two individuals, j and k, may have identical covariates ($\mathbf{x}_{ij} = \mathbf{x}_{ik}$), but have different probabilities for instrument assignment, $P(Z_{ij} = 1|\mathcal{F}) \neq P(Z_{ik} = 1|\mathcal{F})$ due to unmeasured confounders, denoted as u_{ij} and u_{ik} for the jth and kth individuals, respectively. Despite our best efforts to minimize the observed differences in covariates and to adhere to assumption (A3) after conditioning on the matched sets, unmeasured confounders could still be different between the jth and kth child, and this difference could make the instrument Z_{ij} depart from randomized assignment, violating assumption (A3).

To model this deviation from randomized assignment due to unmeasured confounders, let $\pi_{ij} = P(Z_{ij} = 1|\mathcal{F})$ and $\pi_{ik} = P(Z_{ik} = 1|\mathcal{F})$ for each unit j and k in the ith matched set. The odds that unit j will receive $Z_{ij} = 1$ instead of $Z_{ij} = 0$ is $\pi_{ij}/(1 - \pi_{ij})$. Similarly, the odds for unit k is $\pi_{ik}/(1 - \pi_{ik})$. Suppose the ratio of these odds is bounded by $\Gamma \geq 1$

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij}(1 - \pi_{ik})}{\pi_{ik}(1 - \pi_{ij})} \leq \Gamma \tag{4.11}$$

If unmeasured confounders play no role in the assignment of Z_{ij}, then $\pi_{ij} = \pi_{ik}$ and $\Gamma = 1$. If there are unmeasured confounders that affect the distribution of Z_{ij}, then $\pi_{ij} \neq \pi_{ik}$ and $\Gamma > 1$. For a fixed $\Gamma > 1$, we can obtain lower and upper bounds on π_{ij}, which can be used to derive the null distribution of $T(0)/S(0)$ under $H_0 : \lambda = 0$ in the presence of unmeasured confounding and be used to compute a range of possible p-values for the hypothesis $H_0 : \lambda = 0$ (Rosenbaum, 2002). The range of p-values indicates the effect of unmeasured confounders on the conclusions reached by the inference on λ. If the range contains α, the significance value, then we cannot reject the null hypothesis at the α level when there is an unmeasured confounder with an effect quantified by Γ.

Specifically, consider Fisher's sharp null hypothesis, $H_0 : r_{1ij}^{(d_{1ij})} = r_{0ij}^{(d_{0ij})}$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$. Note that this hypothesis implies the hypothesis $H_0 : \lambda = 0$. Furthermore, the test statistic in (4.5) simplifies to

$$\begin{aligned} T(0) &= \frac{1}{I}\sum_{i=1}^{I}\left\{\frac{n_i}{m_i}\sum_{j=1}^{n_i} Z_{ij}R_{ij} - \frac{n_i}{n_i - m_i}\sum_{j=1}^{n_i}(1 - Z_{ij})R_{ij}\right\} \\ &= \frac{1}{I}\sum_{i=1}^{I}\frac{n_i^2}{m_i(n_i - m_i)}\sum_{j=1}^{n_i} Z_{ij}R_{ij} - \frac{1}{I}\sum_{i=1}^{I}\frac{n_i}{n_i - m_i}\sum_{j=1}^{n_i} R_{ij} \end{aligned}$$

Regardless of the distribution of $P(\mathbf{Z}|\mathcal{F}, \mathcal{Z})$, $\frac{1}{I}\sum_{i=1}^{I} n_i/(n_i - m_i)\sum_{j=1}^{n_i} R_{ij}$ is constant since $r_{1ij}^{(d_{1ij})} = r_{0ij}^{(d_{0ij})}$ under Fisher's sharp null hypothesis. Hence, we can use the simpler statistic, $\tilde{T}(0)$,

$$\tilde{T}(0) = \frac{1}{I}\sum_{i=1}^{I}\frac{n_i}{m_i(n_i - m_i)}\sum_{j=1}^{n_i} Z_{ij}R_{ij} \tag{4.12}$$

to test the Fisher's sharp null hypothesis. If the responses are binary, equation (4.12) is the sign-score test statistic for which exact bounds on p-values exist (Rosenbaum, 2002). If the responses are continuous, Gastwirth et al. (2000) and Small et al. (2009) provide an approximate bound on p-values.

In addition, we can amplify the interpretation of Γ using Rosenbaum and Silber (2009) to get a better understanding of the impact of the unmeasured confounding on the outcome and the instrument. To do this , consider a binary unmeasured confounder with two values Δ and Λ where Δ and Λ have the following property

$$\Gamma = \frac{\Delta\Lambda + 1}{\Delta + \Lambda}, \quad \Delta > 0, \Lambda > 0 \tag{4.13}$$

The parameter Λ refers to the odds of having one instrument value over another. The parameter Δ refers to the odds of having one outcome over another. For each Γ, we can use equation (4.13) and translate the interpretation of Γ as the combined effect an unmeasured confounder must have on the instrument, Λ, and on the outcome, Δ, to change the inference.

4.3. Simulation

4.3.1. Comparison to TSLS

One of the advantages of matching based IV estimation versus traditional IV estimation, such as conventional TSLS without matching, is its robustness to parametric assumptions between the outcome and the covariates. Specifically, for conventional TSLS, in order for the estimate to be consistent, the covariates must have a linear effect on the expected outcome. In contrast, matching-based IV estimation puts no constraints on the structure of the relationship between the outcome and the covariates. In this section, we study this phenomena in detail through a simulation study.

Let the outcome R_{ij}, the exposure D_{ij}, the observed covariates $\mathbf{X}_{ij}$, and the instrument Z_{ij} be generated based on the following model known as the structural equations model in econometrics (Wooldridge, 2010).

$$\begin{aligned} R_{ij} &= \alpha + \beta D_{ij} + f(\mathbf{X}_{ij}) + \epsilon_{ij} \\ D_{ij} &= \kappa + \pi Z_{ij} + \boldsymbol{\rho}^T \mathbf{X}_{ij} + \xi_{ij} \end{aligned} \quad , \quad \begin{pmatrix} \epsilon_{ij} \\ \xi_{ij} \end{pmatrix} \overset{\text{iid}}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix} \right)$$

where the parameters α, β, κ and $\boldsymbol{\rho}$ are all fixed throughout the simulation. The parameters α and κ are intercepts. The parameter β is the quantity of interest, the effect of the exposure on the outcome, and is also equal to the effect ratio (see Section 4.2.6). The parameter π quantifies the strength of the instrument. The function $f(\cdot)$ is a pre-defined function that takes in a vector of observed covariates $\mathbf{X}_{ij}$ and produces a scalar value that affects the outcome, R_{ij}. In the simulation, $\mathbf{X}_{ij}$, are five-dimensional vectors or $\mathbf{X_{ij}} = (X_{ij1}, \dots, X_{ij5})$. Also, we consider the following list of functions parametrized by $\boldsymbol{\gamma} \in \mathbb{R}^5$

(a) Linear function: $f(\mathbf{X}_{ij}) = \sum_{k=1}^5 \gamma_k X_{ijk}$

(b) Quadratic function: $f(\mathbf{X}_{ij}) = \sum_{k=1}^5 \gamma_k X_{ijk}^2$

(c) Cubic function: $f(\mathbf{X}_{ij}) = \sum_{k=1}^5 \gamma_k X_{ijk}^3$

(d) Exponential function: $f(\mathbf{X}_{ij}) = \sum_{k=1}^{5} \gamma_k \exp(X_{ijk})$

(e) Log function: $f(\mathbf{X}_{ij}) = \sum_{k=1}^{5} \gamma_k \log(|X_{ijk}|)$

(f) Logistic function: $f(\mathbf{X}_{ij}) = \frac{1}{1+\exp(-\sum_{k=1}^{5} X_{ijk}\gamma_k)}$

(g) Truncated function: $f(\mathbf{X}_{ij}) = \sum_{k=1}^{5} \gamma_k \chi(X_{ijk} \geq 0)$ where $\chi(\cdot)$ is an indicator function.

(h) Square root function: $f(\mathbf{X}_{ij}) = \sum_{k=1}^{5} \gamma_k \sqrt{|X_{ijk}|}$

To generate $\mathbf{X}_{ij}$, we adopt the following scheme. For individuals with $Z_{ij} = 0$, $\mathbf{X}_{ij}$ comes from a five-dimensional multivariate Normal distribution with mean $(0, \ldots, 0)$ and an identity covariance matrix. For individuals with $Z_{ij} = 1$, $\mathbf{X}_{ij}$ comes from a five-dimensional multivariate Normal with mean $(1, 0, \ldots, 0)$ and an identity covariance matrix. The instruments, Z_{ij}, are generated randomly with $P(Z_{ij} = 1) = 1/8$ and $P(Z_{ij} = 0) = 7/8$, similar to the ratio observed in our malaria data (see Chapter 5). For each generated data set, we compute the estimate of β using TSLS and our procedure. TSLS is based on (i) regressing D_{ij} on Z_{ij} and X_{ij} to obtain the predicted value of D_{ij}, say $\hat{D}_{ij}$, and (ii) regressing R_{ij} on $\hat{D}_{ij}$ and $\mathbf{X}_{ij}$. We simulate this process 5000 times and compute the estimates of β produced by the two procedures. We measure the performance of the two procedures by computing the median absolute deviation, the absolute bias of the median (i.e. the absolute value of the bias of the median estimate with respect to β), and the Type 1 error rate over 5000 simulations. For each simulation study, we vary the function $f(\cdot)$ and π.

Figures 30, 31 and 32 compare performances between TSLS and our method when we fix the sample size, but vary the strength of the instrument (i.e. the strength of the effect of the instrument on the treatment) via π. Specifically, we evaluate the strength of the instrument using a popular measure known as the concentration parameter (Bound et al., 1995). High values of the concentration parameter indicate a strong instrument while low values of it indicate a weak instrument. The concentration parameter is the population value of the first stage partial F statistic for the instrument when the treatment is regressed

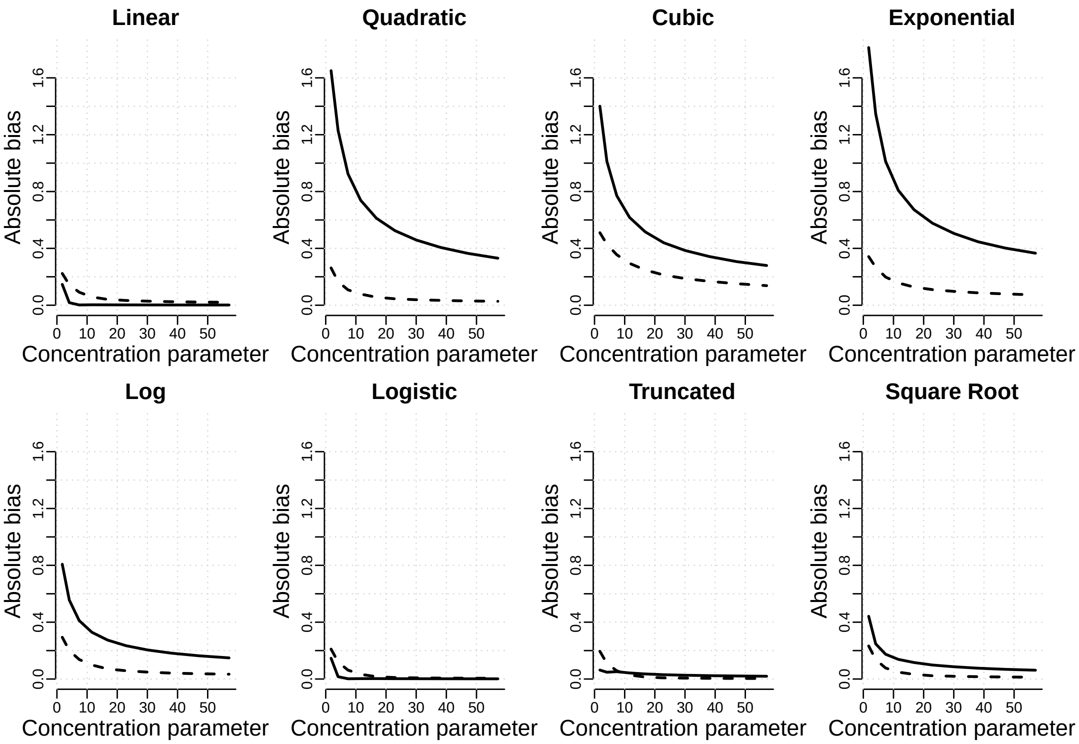

Figure 30: Absolute Bias of the Median Between Our Full Matching Method and TSLS for Different Concentration Parameters. The solid line indicates TSLS and the dashed line indicates our method.

on the instrument and the measured covariates $\mathbf{X}_{ij}$; this first stage F statistic is often used to check instrument strength where an F below 10 suggests that the instruments are weak (Stock et al., 2002). The sample size is fixed at 800 where 100 individuals have $Z_{ij} = 1$ and 700 individuals have $Z_{ij} = 0$. We also vary $f(\cdot)$ based on the functions listed in the previous paragraph.

Figure 30 measures the absolute bias of the median for TSLS and our method. When $f(\cdot)$ is a linear function of the observed covariates $\mathbf{x}_{ij}$, TSLS does slightly better than our method. TSLS doing well for the linear function is to be expected since TSLS is consistent when the model is linear. However, if $f(\cdot)$ is non-linear, our matching estimator does better than TSLS and is never substantially worse for all instrument strengths. For example, for quadratic, cubic, exponential, log, and square root functions, our method has lower bias than TSLS for all strengths of the instrument. For logistic and truncated functions, our method is similar in performance to TSLS for all strengths of the instrument.

Figure 31 measures the median absolute deviation (MAD) of TSLS and our method. Our method tends to have a slightly higher MAD than TSLS. This higher variability of our method is to be expected since our method uses a nonparametric approach whereas TSLS is a parametric approach. However, as the instrument gets stronger (i.e. high concentration parameter), the gap between the two MADs shrinks quickly.

Finally, Figure 32 measures the Type I error rate of TSLS and our method. Regardless of the function type and the instrument strength, our method retains the nominal 0.05 rate. In fact, even for the linear case where TSLS is designed to excel, our estimator has the correct Type I error rate for all instrument strengths while TSLS has higher Type I error for weak instruments. For all the non-linear functions, the Type I error rate for TSLS remains above the 0.05 line while our estimator maintains the nominal Type I error rate. This provides evidence that our estimator will have the correct 95% coverage for confidence intervals regardless of non-linearity or instrument strength.

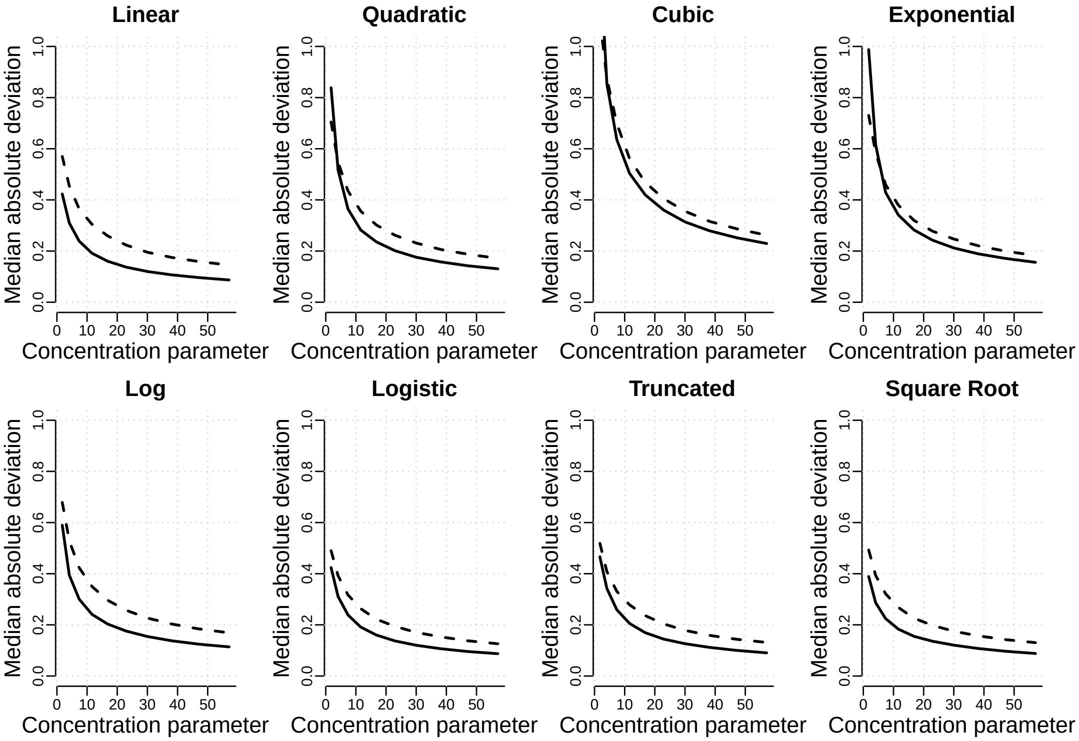

Figure 31: Median Absolute Deviation Between Our Full Matching Method and TSLS for Different Concentration Parameters. The solid line indicates TSLS and the dotted line indicates our method.

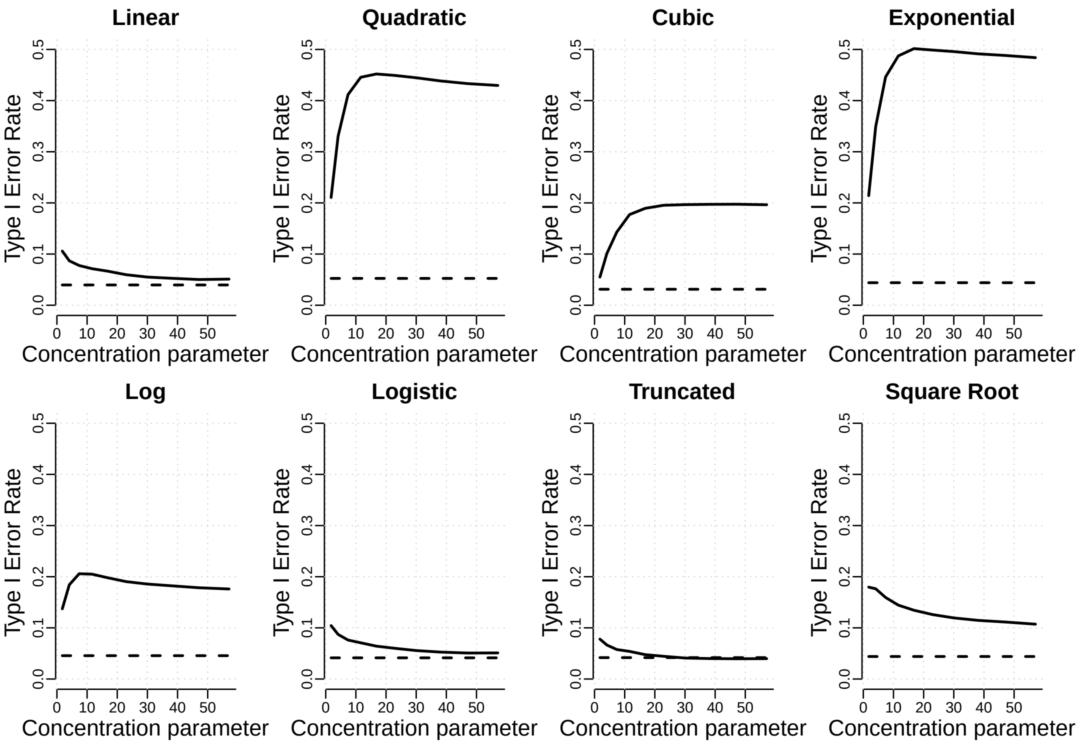

Figure 32: Type I Error Rate Between Our Full Matching Method and TSLS for Different Concentration Parameters. The solid line indicates TSLS and the dashed line indicates our method.

In summary, the simulation study shows promise that our method is generally more robust to assumptions about instrument strength and linearity between the outcome and the covariates than TSLS at the expense of a small increase in dispersion.

4.3.2. Comparison to Frölich (2007)

In addition to comparing our method against the most popular IV estimator, TSLS, we also compare our method to the non-parametric IV method of Frölich (2007) implemented by Frölich and Melly (2010). The simulation setup is identical to Section 4.3.1, except that we discretize the exposure value D_i so that we can compare our method to the method in Frölich (2007). Specifically, let D_{ij}^* be defined as D_{ij} in Section 4.3.1, i.e. $D_{ij}^* = \kappa + \pi Z_{ij} + \rho^T \mathbf{X}_{ij} + \xi_{ij}$. Then, we define

$$D_{ij} = \chi(D_{ij}^* < -1) + 2\chi(-1 \leq D_{ij}^* < 1) + 3\chi(1 \leq D_{ij}^*)$$

The response R_{ij} is generated from the same model as in Section 4.3.1, except with a discretetized D_{ij}. The rest of the data generating process is identical to Section 4.3.1.

For each simulated data, we use the code provided by Frölich and Melly (2010) to generate an estimate for β^*, the local average treatment effect, with the default settings for the tuning parameters. We also use our method to estimate β^*. Finally, for comparison, we run TSLS on the simulate data. As before, we measure the absolute bias of the median and the median absolute deviation (MAD). For each simulation study, we vary the function $f(\cdot)$ and π, the strength of the instrument.

Figures 33 and 34 show the absolute bias and median absolute deviation, respectively, between the three methods. Generally speaking, both our method and method by Frölich (2007) do better than TSLS when $f(\cdot)$ is non-linear. Between our method and one by Frölich (2007), in most cases, our method is better or similar to the method of Frölich (2007) when it comes to bias. With regards to variability, our method and the method of Frölich (2007) are very similar to each other. For the quadratic, cubic, and exponential functions, our

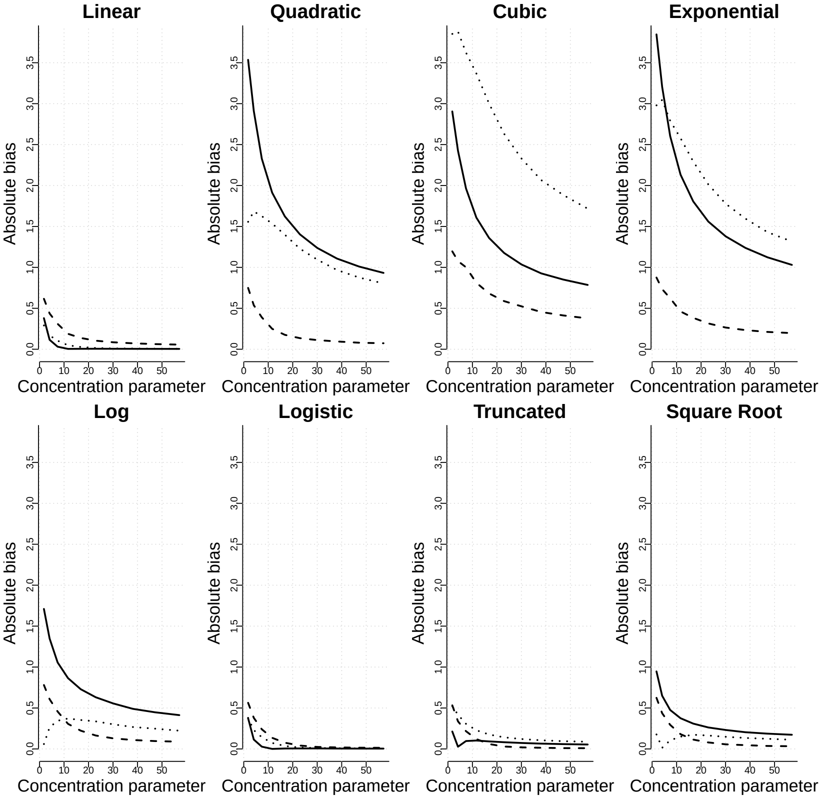

Figure 33: Absolute Bias of the Median Between Our Full Matching Method, TSLS, and Frölich's Method for Different Concentration Parameters. The solid line indicates 2SLS, the dashed line indicates our method, and the dotted line indicates Frölich's method.

simulations show that our method dominates both in bias and variance compared to Frölich (2007).

Unfortunately, we were not able to produce Type I error results for the method of Frölich (2007) because of a coding error in the code provided by Frölich and Melly (2010) which provided negative standard errors on the estimates produced by it. Frölich (personal communication) is aware of the issue and will be releasing a new version in the future.

4.3.3. Approximations of Efficiency

In this section, we assess the accuracy of the efficiency formula provided in Section 4.2.6 by the following simulation study. The variables R_{ij}, D_{ij} and Z_{ij} are generated via the

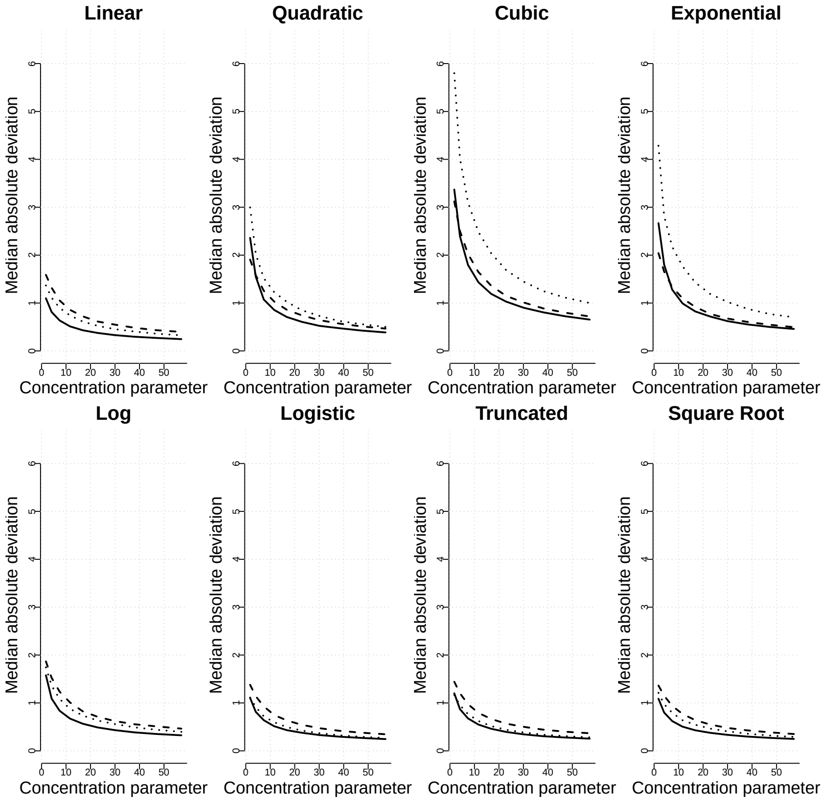

Figure 34: Median Absolute Deviation Between Our Full Matching Method, TSLS, and Frölich's Method for Different Concentration Parameters. The solid line indicates 2SLS, the dashed line indicates our method, and the dotted line indicates Frölich's method.

I	Theoretical Variance		Simulated Variance	
	Strong	Weak	Strong	Weak
50	0.024	0.59	0.028	3224.30
100	0.012	0.30	0.012	181.06
110	0.011	0.27	0.012	2506.92
500	0.0024	0.060	0.0025	2.05
1000	0.0012	0.030	0.0012	0.037
5000	0.00024	0.0060	0.00024	0.0063
10000	0.00012	0.0030	0.00012	0.0030

Table 25: Comparison of Simulated Variance and Theoretical Variance for Different Strength of Instruments and Matched Sets

model in (4.9) and (4.10) with Z_{ij} assumed to be fixed. We randomly pick α_i, τ_i, and β. We pick γ to be 1 for the strong instrument case and -0.2 for the weak instrument case. We assume a homoscedastic variance for the error terms where all the $\sigma_{i,R}^2, \sigma_{i,D}^2$, and $\sigma_{i,RD}$ are the same for every i. We compute the effect ratio estimator, repeat this process 1000 times, and compute the simulated variance. The theoretical variance is calculated based on the formula provided in Theorem 4.3. Table 25 shows the results.

Table 25 shows us that for strong instruments, the agreement between theoretical formula in Theorem 4.3 and simulation is quite good for all values of I. On the other hand, for weak instruments, there is substantial deviation between the theoretical variance and the simulated variance until I is above 5000.

4.4. Discussion

Overall, in contrast to regression-based IV estimation procedures like TSLS, our full matching IV method provide a clear way to assess the balance of observed covariates and design the study without looking at the outcome data along with a way to quantify the effect of unmeasured confounders on our inference of the causal effect. Our method made it explicitly clear how these covariates were adjusted by stratifying individuals based on similar covariate values. Finally, like in a randomized experiment, our method only looks at the outcome data once the balance was acceptable. If the balance was unacceptable, then comparing the outcomes between the two groups would not provide reliable causal inference since any differences in the outcome can be attributed to the differences in the covariates. In contrast, conventional TSLS can only analyze the causal relationship in the presence of outcome data, making the outcome data necessary throughout the entire analysis. Finally, our method is robust to parametric modeling assumptions between the outcome and the covariates with respect to Type I error and point estimate, which cannot be said about TSLS.

At the expense of these benefits, especially blinding and transparency with regards to covariate balance, unfortunately matching estimators tend to be less efficient than TSLS or some of the semiparametric methods mentioned in Section 4.1.1 when the semiparametric methods' assumptions hold. In practice, our estimator's blinding and transparency can be a powerful design and visual tool for applied researchers to assess the validity of the causal conclusions. However, a more careful exploration of the trade-offs between the efficiency of our estimator and the efficiency of some of the semiparametric and non-parametric methods is an interesting direction for future research.

CHAPTER 5 : An Application: The Causal Effect of Malaria on Stunting in Children from Ghana

This is joint work with Benno Kreuels, Ohene Adjei, Ralf Krumkamp, Jürgen May, and Dylan Small.

5.1. Background: Malaria and Childhood Development in sub-Saharan Africa

In 2013 alone, there were 128 million estimated cases of malaria in sub-Saharan Africa, with most cases occurring in children under the age of 5 (World Health Organization, 2014). In addition to being one of the major causes of death in early childhood, repeated malaria episodes are a major cause of chronic anemia and may impair child development (Korenromp et al., 2004). Consequently, it is important to study the impact of malaria on child development to prioritize public health resources.

Previous epidemiological studies on the association between malaria and child growth have produced inconsistent results, which is partly rooted in different methodological approaches. Several studies assessed growth using the mean height-for-age Z-score, while other studies used the prevalence of stunting (height-for-age Z-score < -2) as an indicator of insufficient growth. Stunting is a common condition in African children and is one of the main determinants of childhood morbidity and mortality (Rice et al., 2000). In 1956 a study from the Gambia first showed a tendency to higher mean Z-scores in infants who received malaria prophylaxis compared to children who did not (McGregor et al., 1956). Later an association between malaria and child growth or risk of stunting was also seen in Nigeria (Bradley-Moore et al., 1985), Kenya (ter Kuile et al., 2003), The Gambia (Deen et al., 2002), Ghana Ehrhardt et al. (2006), and Uganda Arinaitwe et al. (2012). Other studies, however, found no association (Snow et al., 1991; Fillol et al., 2009; Deribew et al., 2010; Crookston et al., 2010) or even a higher risk of malaria in children with better z-scores (Genton et al., 1998). Finally, one study demonstrated that the association between stunting and malaria might be strongest in young children (Nyakeriga et al., 2004).

A major limitation common to all previous studies is the inability to fully adjust for confounding. Specifically, nutritional deficiencies are important potential confounders because they are an important determinant of stunting and they also compromise immune function, which could result in a higher risk of infection (Fillol et al., 2009). Further potential confounders are socioeconomic status, living conditions, and other infections. In addition, reverse causality in the association of stunting and malaria seems possible. Randomized trials recruiting children at birth could account for potential confounders and reverse causality but are impractical in this context.

In this paper, we seek to control for confounders in estimating the causal effect of malaria on stunting by using a combination of Mendelian randomization (MR) and matching (in Chapter 4). The basic idea of MR is to extract variation in an exposure (i.e. malaria) that is due to a Mendelian gene, which is independent of confounders, and use this confounder-free variation to estimate the effect of the exposure on the outcome (i.e. stunting) (Davey Smith and Ebrahim, 2003, 2004; Lawlor et al., 2008). The hemoglobin variant HbS, which is caused by a point-mutation at the 6th position of the β-Globin gene (β6Glu Val), serves as the paradigm for balanced polymorphisms; while people homozygote for HbS (HbSS) have sickle cell disease with an increased mortality, heterozygote carriers (HbAS, sickle cell trait) are asymptomatic and protected from malaria (May et al., 2007; Kreuels et al., 2010). A previous analysis of the current data showed a negative association between the HbAS genotype and stunting in an area of high malaria endemicity and computed the magnitude of the association (Kreuels et al., 2009). However, the study did not analyze the effect of malaria on stunting and the magnitude of such an effect. In this analysis, we use HbAS as a Mendelian gene to expand on this finding and estimate the effect of malaria on stunting. To control for measured confounders (e.g. birth weight, ethnic group, mosquito protection), we will use matching laid out in Chapter 4.

5.2. Methods

5.2.1. Study Population and Design

The study was conducted in the Ashanti region in Ghana. A cohort of 1070 infants was recruited as part of a clinical trial on intermittent preventative treatment with Sulphadoxine-Pyrimethamine (SP) (Kobbe et al., 2007). Infants were recruited at three months of age and followed-up monthly until age two with comprehensive examinations including a standardized medical history, a measurement of body temperature, and a thick-and-thin smear for microscopic malaria diagnostics. Passive case detection was performed between scheduled visits. A child was diagnosed with malaria if he/she had a parasite-density of more than 500 parasites/μl and a body temperature greater than 38°C or the mother reported a fever within the last 48 hours. In three monthly intervals, standardized anthropometric measurements, including height and weight, were performed. A child was deemed stunted if her/his length/height-for-age z-score was less than -2 (i.e. moderate or severe stunting) (WHO Multicentre Growth Reference Study Group, 2006). Further details of the study population are published in a previous paper (Kobbe et al., 2007).

5.2.2. Definition of Instrument, Exposure, and Outcome

For this analysis only infants with heterozygote HbAS or wildtype HbAA were considered. Children with homozygote mutation (HbSS) or a different mutation on the same gene leading to hemoglobin C (HbAC, HbCC, HbSC) were excluded. The instrument was a binary variable indicating the HbAS or HbAA genotype. The exposure of interest was the malarial history defined as the total number of malaria episodes during the study. A malaria episode, as stated before, was defined as having a parasite density of more than 500 parasites/μl and a body temperature greater than 38°C or the mother reported a fever within the last 48 hours. The outcome of interest was whether the child was stunted at the last recorded visit, which took place when the child was approximately two years old.

5.2.3. Assumptions for Instrumental Variables with the Sickle Cell Trait

We formalize the core assumptions of an instrumental variable below (Holland, 1988; Angrist et al., 1996; Yang et al., 2014) (see Figure 1)

(A1) The sickle cell is associated with malaria episodes

(A2) All directed pathways from the sickle cell trait to stunting passes through malaria episodes (i.e. there is no pathway that goes directly from the sickle cell genotype to stunted growth)

(A3) There are no unmeasured confounders that are associated with the sickle cell trait and stunted growth

We now assess the validity of (A1)-(A3) for the sickle cell trait, the instrument for our analysis on the effect of malaria on stunting. For assumption (A1), there is substantial evidence that the sickle cell trait does provide protection against malaria as compared to people with two normal copies of the HBB gene (HbAA) (Aidoo et al., 2002; Williams et al., 2005; May et al., 2007; Cholera et al., 2008; Kreuels et al., 2010). Also, with this data, when we characterize the effect of the sickle cell trait on malaria based on a Poisson regression, the difference in episodes of malaria between children with HbAS and HbAA is significant (Risk ratio: 0.82, p-value: 0.02, 95% CI: (0.70, 0.97)), indicating that the sickle cell trait instrument satisfies (A1) of being associated with the exposure. This is also in alignment with previous literature on the relationship between sickle cell genotype and malaria for this data (Kreuels et al., 2010).

For assumption (A2), this could be violated if the sickle cell trait had effects on stunting other than through causing malaria, for instance, if the sickle cell trait was pleiotropic (Davey Smith and Ebrahim, 2003). We can partially test this assumption by examining individuals who carry the sickle cell trait, but who grew up in a region where malaria is not present. That is, if assumption (A2) were violated, heights between individuals with HbAS and HbAA in such a region would be different since there would be a direct arrow between the sickle cell trait and height. Studies among African American children and children from the Dominican Republic and Jamaica for whom the sickle cell trait is common, but there is no malaria in the area, found no evidence that the sickle cell trait affected a child's physical development (Ashcroft et al., 1976; Kramer et al., 1978; Ashcroft et al., 1978; Rehan, 1981). This supports the validity of assumption (A2). Note, however, that although the results of this test support the validity of (A2), (A2) could still be violated. For example, the sickle cell trait could have a direct effect that interacts with the environment in such a way that the direct effect is only present in Africa, but not in the Dominican Republic or Jamaica.

For assumption (A3), this assumption would be questionable in our data if we did not control for any population stratification covariates. Population stratification is a condition where there are subpopulations, some of which are more likely to have the sickle cell trait, and some of which are more likely to be stunted through mechanisms other than malaria (Davey Smith and Ebrahim, 2003). For example, in Table 26 which provides the baseline characteristics for our data, we observed that the village Tano-Odumasi had more children with HbAA than HbAS. It is possible that there are other variables besides HbAA that differ between the village Tano-Odumasi and other villages and affect stunting. Hence, assumption (A3) is more plausible if we control for observed variables, like village of birth, and we use full matching in Chapter 4 to achieve this. Specifically, within the framework of matching, for each matched set, if the observed confounders in Table 26 are similar among all individuals in that matched set, it may be more plausible that the unobserved variable, say u, plays no role in the distribution of the sickle cell genotype among all the individuals in the matched set. If (A3) exactly holds and subjects are exactly matched for their observed confounders, then within each matched set, sickle cell is simply assigned by a random mechanism. In Section 4.2.7, we discuss a sensitivity analysis that allows for the possibility that even after matching for observed variables, the unobserved variable u may still influence the assignment of the sickle cell trait in each matched set, meaning that assumption (A3) is violated.

Other notable IV assumptions, such as Stable Unit Treatment Value Assumption (SUTVA)and monotonicity, are fairly reasonable in this data. SUTVA states that one's individual potential outcomes are not affected by the genotype assignment of another individual (Angrist et al., 1996). Our instrument, the sickle cell genotype, was determined at the conception of the child and hence, a child's genotype only affects his exposure and outcome, and not the exposures and outcomes of other children. Monotonicity, within the framework of MR, states that there are no individuals who would have an adverse effect on the exposure from inheriting the genotype which is purported to bring positive effect on the exposure. In MR where the chosen genetic instruments usually bring about a positive effect on the exposure, monotonicity is reasonable, especially with our instrument, the sickle cell genotype, where it is widely believe that inheriting the trait provides individuals protection from malarial infection compared to not inheriting the trait.

5.2.4. Full Matching on Malaria Data and Efficiency Simulation

We conduct full matching on all observed covariates. In particular, we group children with HbAS and HbAA based on all the observed characteristics in Table 26 as well as match for patterns of missingness. To measure similarity of the observed and missing covariates, we use the rank-based Mahalanobis distance as the distance metric for covariate similarity (Rosenbaum, 2010). In addition, we compute propensity scores by logistic regression. Here, the propensity score is an instrumental propensity score, which is the probability of having the sickle cell trait given the measured confounders (Cheng, 2011). In addition, children with missing values in their covariates were matched to other children with similar patterns of missing data (Rosenbaum, 2010). Once covariate similarity was calculated, the matching algorithm optmatch in R (Hansen and Klopfer, 2006) matched children carrying HbAS with children carrying HbAA in a way that within each matched set, their covariates are similar.

Hansen (2004) discusses how the size of matched sets in full matching can be restricted to gain efficiency and Section 4.2.6 provides a method to compute efficiency. Unfortunately, for us to use the formula in that section, it requires, among other things, a linear model between the outcome and the exposure. In our study where stunting, the outcome, is a binary variable and malaria, the exposure, is a whole number, it is unreasonable to assume that the binary outcome is a linear function of malaria exposures.

To tackle this, we propose a simulation study to analyze efficiency for different full matching schemes. For our malaria data, we fix the instruments and the measured covariates, which, in turn, fixes the matched sets. We assume a Poisson relationship between the number of malaria episodes and the instrument and a logistic relationship between the number of episodes and the stunting outcome. In particular, we use the following model

$$P(R_{ij} = 1) = \frac{1}{1 + e^{-(\alpha_i + \beta D_{ij} + u_{ij})}}, \quad E(D_{ij}) = e^{\tau_i + \gamma Z_{ij}}$$

where R_{ij} is the outcome, D_{ij} is the exposure, and Z_{ij} is the instrument. We fix β, the effect of malaria on stunting, to be 0.32 and γ, the strength of the instrument, to be -0.20 based on the estimates in Kang et al. (2013); the estimate of γ was based on the risk ratio estimate. We also randomly choose α_i and τ_i, the intercepts, from Normal distributions with means -1.67 and -0.19, respectively, and variances 0.12 and 0.027, respectively. The mean and the variance for α_i are from the estimated intercept term and its corresponding standard error of the logistic regression between R_{ij} and D_{ij}. Similarly, the mean and the variance for τ_i are from the estimated intercept term and its corresponding standard error of the Poisson regression between D_{ij} and Z_{ij}. Once all the parameters are set, we sample 884 observations of (R_{ij}, D_{ij}) (i.e. the sample size of the malaria data set) and compute the effect ratio estimator based on the sample of 884. Note that the effect ratio estimator should be able to estimate β since it doesn't rely on the functional form between stunting (i.e. outcome) and malaria episodes (i.e. exposure). We repeat the simulation 5000 times and compute the median absolute deviation as a robust proxy for variance of the effect ratio estimator.

5.2.5. Estimator of the Effect Ratio

After matching, we estimate the effect ratio, as described in Section 4.2.4. In the malaria data, the effect ratio parameter can be interpreted as the weighted average reduction in stunting from a one-unit reduction in malaria episodes among individuals who were protected from malaria by the sickle cell trait. Similarly, each weight represents each individual's protection from at least k malaria episodes by carrying the sickle cell trait compared to the overall number of individuals who are protected from varying degrees of malaria episodes by carrying the sickle cell trait.

We use the test statistic described in Section 4.2.5 to estimate the effect ratio and obtain inferential quantities like p-values and 95% confidence intervals. We note that the regularity conditions, specifically the moment conditions in Theorem 4.2 of Section 4.2.5 (i.e. $V_i^4(\bar{\lambda})$ is uniformly bounded), are automatically met because the responses are binary (i.e. stunted or not stunted) and the malaria episodes are bounded whole numbers. Hence, Theorem 4.2 and the subsequent Corollary 4.1 are used to compute the point estimate, the p-value, and the confidence intervals for the casual effect of malaria on stunting.

Also, for comparison, we computed the multiple regression estimate of the effect ratio, an estimate that only adjusts for measured confounding, but not unmeasured confounding. This estimate is derived from a multiple linear regression with stunting as the dependent variable and all measured confounders and the number of malaria episodes as independent variables. From the regression, we take the estimated slope coefficient for malaria episodes, which is the reduction in the risk of stunting per malaria episode.

5.2.6. Sensitivity Analysis

Despite our best efforts to minimize the observed differences in covariates and to adhere to assumption (A3) after conditioning on the matched sets, unmeasured confounders such as a child's family's ancestry could still be different between the jth and kth child, and this difference could make the inheritance of the sickle cell trait depart from randomized assignment, violating assumption (A3). To quantify the effect of unmeasured confounders on the obtained inference, a sensitivity analysis outlined in Section 4.2.7 was performed. Specifically, we consider a binary unmeasured confounder that has a specified effect on the odds of inheriting HbAS over HbAA and specified effect on the odds of stunting (conditional on measured confounders), and evaluate the effect such an unmeasured confounder would have on the inference we make. Also following Section 4.2.7, we amplify our sensitivity analysis to increase interpretability.

5.3. Results

5.3.1. Basic Data

The analysis was conducted on 884 children with HbAA or HbAS genotype. 774 children were HbAA homozygotes while 110 children were HbAS heterozygotes. 35 children (4.0%) were already stunted at the beginning of the trial and by the end, 168 children (19.0%) were stunted. The t-statistic to test the difference in the time of the last recorded visit amongst HbAA and HbAS did not indicate any variation (p=0.21, 95% CI: (-3.70,16.68)).

Table 26 shows the baseline characteristics of the HbAS and HbAA subjects before matching. Before matching, most characteristics at recruitment were similar between children with HbAS and HbAA. The notable exception is birth weight. There was evidence that birth weight of children with HbAA was lower than of children with HbAS (p=0.006, 95% CI: (-228.27,-39.14)).

5.3.2. Matching and Efficiency

Figure 35 shows covariate balance before and after full matching using absolute standardized differences. Absolute standardized differences before matching are computed by taking the difference of the means between children with HbAS and HbAA for each covariate, taking the absolute value of it, and normalizing it by the within group standard deviation before matching (the square root of the average of the variances within the groups). Absolute

	HbAS ($n = 110$)	HbAA ($n = 774$)
Birth weight (Mean,(SD))	3112.44 (381.9) (32 missing)	2978.7 (467.9) (239 missing) ***
Sex (Male/Female)	46.4% Male	51.0% Male
Birth season (Dry/Rainy)	56.4% Dry	55.3% Dry
Ethnic group (Akan/Northerner)	86.4% Akan	88.8 % Akan (4 missing)
α-globin genotype (Norm/Hetero/Homo)	75.7% / 21.5% / 2.8% (3 missing)	74.4% / 23.1% / 2.6% (29 missing)
Village of residence:		
Afamanso	4.6 %	4.8%
Agona	10.0%	13.6%
Asamang	13.6%	11.1%
Bedomase	5.5%	4.5%
Bipoa	14.5 %	10.7%
Jamasi	15.5 %	13.8%
Kona	16.4 %	12.8%
Tano-Odumasi	4.5 %	12.3%**
Wiamoase	15.5 %	16.4%
Mother's occupation (Nonfarmer/Farmer)	79.0% Nonfarmer	78.0% Nonfarmer (11 missing)
Mother's education (Literate/Illiterate)	91.7% Literate (2 missing)	90.5% Literate (8 missing)
Family's financial status (Good/Poor)	69.1% Good (13 missing)	70.1% Good (84 missing)
Mosquito protection (None/Net/Screen)	55.7% / 32.0% / 12.4% (13 missing)	45.4%* / 35.1% / 19.5% (76 missing)
Sulphadoxine pyrimethamine (Placebo/SP)	49.1% Placebo	50.1% Placebo

Table 26: Characteristics of Study Participants at Recruitment for the Malaria Study. P-values were obtained by doing a Pearson's chi-squared test for categorical covariates and two-sample t tests for numerical variables. *** corresponds to a p-value of less than 0.01, ** corresponds to a p-value between 0.01 and 0.05, and * corresponds to a p-value between 0.05 and 0.1.

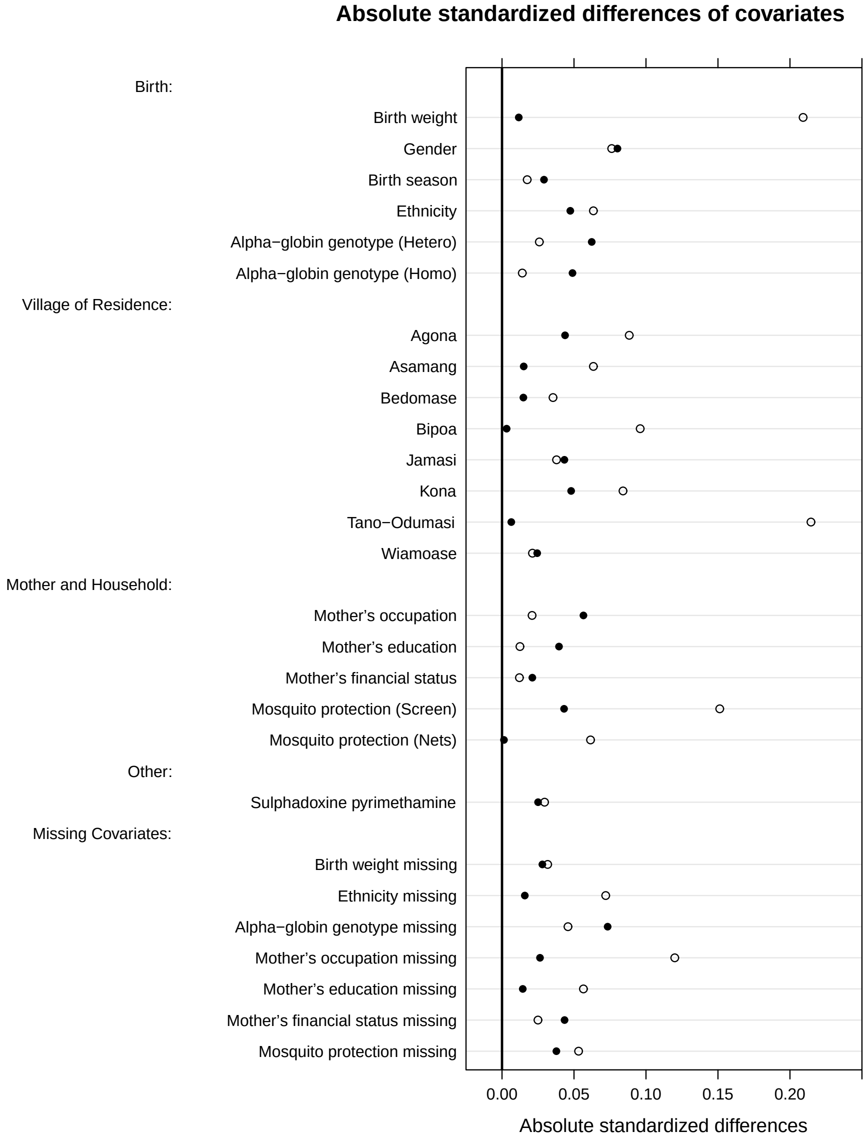

Figure 35: Absolute Standardized Differences Before and After Full Matching for the Malaria Data. Unfilled circles indicate differences before matching and filled circles indicate differences after matching.

Matching	Median absolute deviation	Standardized bias
Full matching (max strata size is 9)	0.90	0.23
Full matching (max strata size is 10)	0.96	0.19
Full matching (max strata size is 15)	0.97	0.10
Full matching (unrestricted)	0.98	0.055

Table 27: Trade-off Between Efficiency and Balance for Different Full Matching Schemes in the Malaria Data

standardized differences after matching are computed by taking the differences of the means between children with HbAS and HbAA within each strata, averaging this difference across strata, taking the absolute value of it, and normalizing it by the same within group standard deviation before matching as before. Before matching, there are differences in birth weight, mosquito protection, and village of residence between children with HbAS and HbAA. After matching, these covariates are balanced. Specifically, the standardized differences for birth weight, village of residence, and mosquito protection, are under 0.1 indicating balance (Normand et al., 2001). In fact, all the covariates are balanced after matching and the p-values used to test the differences between HbAS and HbAA in Table 26 are no longer significant after matching.

Table 27 shows the trade-off between efficiency and covariate balance for different full matching schemes that use all 884 samples of the malaria data. In particular, we restrict the matched set sizes to different values to see the impact on efficiency and standardized bias. The standardized bias is the instrumental propensity score (Cheng, 2011) and is calculated as the difference in propensity scores before and after matching normalized by the within group standard deviation before matching (the square root of the average of the variances within the group). We see that unrestricted full matching has the lowest bias among all other full matching schemes. However, full matching with restricted strata size of 9 has the lowest median absolute deviation, albeit by a little in comparison to other matching schemes. Given the large bias reduction by using unrestricted full matching with a small gain in median absolute deviation, we use unrestricted full matching for our analysis.

Methods	Estimate	P-value	95% confidence interval
Our method	0.22	0.011	(0.044, 1)
Two stage least squares	0.21	0.14	(-0.065, 0.47)
Multiple regression	0.018	0.016	(0.0034, 0.033)

Table 28: Estimate of the Effect Ratio in the Malaria Data.

5.3.3. Effect Ratio

Table 28 shows the estimates of the causal effect of malaria on stunting from different methods, specifically our method, conventional two stage least squares (TSLS), and multiple regression. Our method computed the estimate by the procedure outlined in Section 4.2.5. TSLS computed the estimate by regressing all the measured covariates and the instrument on the exposure and using the prediction from that regression and the measured covariates to obtain the estimated effect. Inference for TSLS was derived using standard asymptotic Normality arguments (Wooldridge, 2010). Finally, the multiple regression estimate was derived by regressing the outcome on the exposure and the covariates and the inference on the estimate was based on a standard t test.

We see that the full matching method estimates λ to be 0.22. That is, the risk of stunting among children with the sickle cell trait is estimated to decrease by 0.22 for every malaria episode prevented by the sickle cell trait. Furthermore, we reject the hypothesis $H_0 : \lambda = 0$, that malaria does not cause stunting, at the 0.05 significance level. The confidence interval for λ is $(0.044, 1.0)$. Even the lower limit of this confidence interval of 0.044 means that malaria has a substantial effect on stunting; it would mean that the risk of stunting among children with the sickle cell trait decreases by 0.044 for every malaria episode prevented by the sickle cell trait.

The estimate based on TSLS is 0.21, similar to our method. However, our method achieves statistical significance but TSLS does not. Also, multiple regression, which does not control for unmeasured confounders, estimates a much smaller effect of 0.018.

We also compute the strength of the instrument for our matching method by regressing the

Γ	Range of significance
1.1	(0.0082, 0.041)
1.2	(0.0034, 0.074)
1.3	(0.0015, 0.12)

Table 29: Sensitivity Analysis for the Malaria Data. The range of significance is the range of p-values over the different possible distributions of the unmeasured confounder given a particular value of Γ, which represents the effect of unobserved confounders on the inference of λ.

exposure (malaria episodes) onto the sickle cell trait and dummy variables that indicate which matched group a child belongs to and evaluating the F statistics from this regression. For instrument strength for full matching, the F statistic is 4.15 and its R^2 is 0.21. For instrument strength for TSLS, the F statistic is 4.36 and its R^2 is 0.22.

5.3.4. Sensitivity Analysis

Table 29 shows the sensitivity analysis due to unmeasured confounders. Specifically, we measure how sensitive our method in Table 28 is to violation of assumption (A3) in Section 4.2.3, even after matching. We see that our results are somewhat sensitive to unmeasured confounders at the 0.05 significance level. If there is an unmeasured confounder that increases the odds of inheriting HbAS over HbAA by 10%, i.e. $\Gamma = 1.1$, then we would still have strong evidence that malaria causes stunting. But, if an unmeasured confounder increases the odds of inheriting HbAS over HbAA in a child by 20% (i.e. $\Gamma = 1.2$), the range of possible p-values includes 0.05, the significance level, meaning that we would not reject the null hypothesis of $H_0 : \lambda = 0$, that malaria does not cause stunting.

Figure 36 shows the result of applying the amplification of Γ by looking at the effect by unmeasured confounders on the odds of stunting and odds of inheriting HbAS over HbAA. Specifically, the different values of Γ in the sensitivity analysis provides us with range of possible p-values. Also, each Γ is associated with two other sensitivity parameters Δ, odds of stunting, and Λ, odds of inheriting HbAS over HbAA, and can be presented as a two-dimensional plot with each axis representing Δ and Λ. For example, the point $(\Delta = 1.5, \Lambda = 1.5)$ on Figure 36 represents an unmeasured confounder that increases the odds of stunting

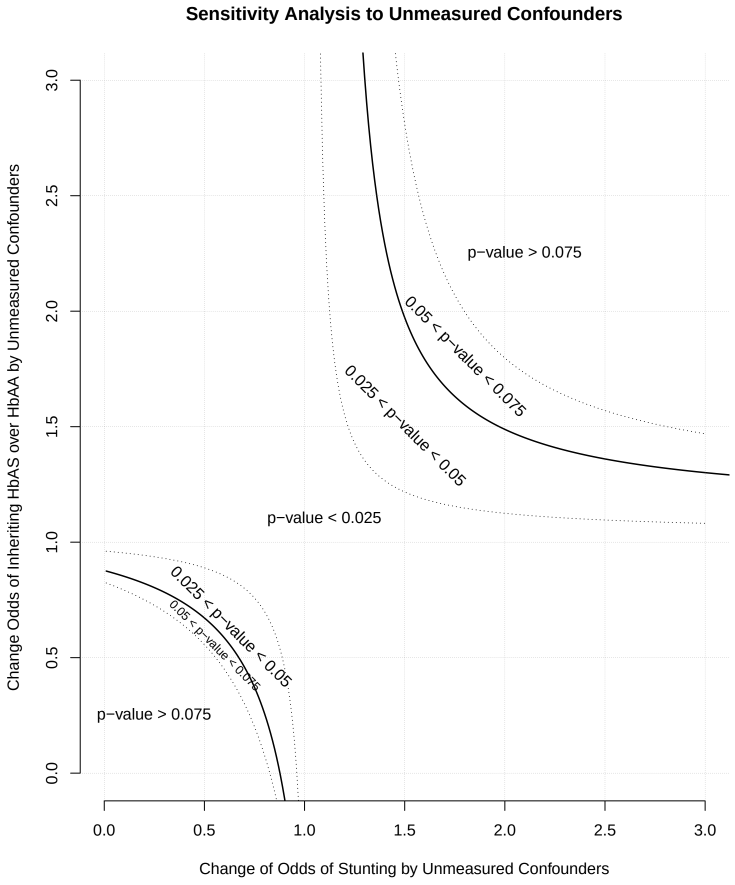

Figure 36: Amplification of Sensitivity Analysis in the Malaria Data. Each point on the graph represents an effect by an unmeasured confounder on the instrument (HbAS) and on the outcome (stunting) to change the inference, specifically the p-value. Points within the two bold curves correspond to effects by unmeasured confounders that will give us p-values < 0.05 and points outside the two bold curves correspond to effects that will give us p-values > 0.05, thereby retaining our null hypothesis.

and inheriting HbAS over HbAA by a factor of 1.5 and produces a p-value in between 0.025 and 0.05, which does not contain the significance level of 0.05. Hence, the null hypothesis would still be rejected despite having such an unmeasured confounder. In contrast, if the unmeasured confounder had an effect of $(2.0, 2.0)$ specified on the plot, the null hypothesis would be retained since the p-value contains the significance level of 0.05.

5.4. Discussion

By using Mendelian randomization with sickle cell trait as the instrument and matching techniques to account for potential confounders, we found evidence of a causal effect of malarial episodes on stunting. Roughly speaking, each increase by one malaria episode increased the risk of stunting by 0.22 (95% CI: $(0.044, 1)$), indicating that the effect of malaria on stunting is substantial in our cohort of infants under two years of age.

Our results confirm findings about an association between malaria and stunting from previous studies (Deen et al., 2002; Ehrhardt et al., 2006; Arinaitwe et al., 2012) as well as findings from earlier studies on an association between mean height-for-age z-scores and malaria (McGregor et al., 1956; Bradley-Moore et al., 1985; ter Kuile et al., 2003). Previous studies were unable to fully adjust for confounding; a large number of personal characteristics, such as nutritional deficiencies, low socioeconomic status, and poor living conditions, are likely to be predictors for both malaria and stunting. Differing levels of confounding in previous studies may have led to findings of no association between malaria and stunting or mean Z-scores (Snow et al., 1991; Deribew et al., 2010; Crookston et al., 2010) or a negative correlation (Genton et al., 1998) or false conclusions about associations. In our study the large difference between the estimate for the effect ratio from the multiple regression and the estimate derived after matching (0.018 vs. 0.22) indicates a substantial level of confounding in multiple regression. MR takes into account unmeasured confounders that are frequently present in observational studies and are not controlled for in standard regression. Under the assumptions stated in the methods section, MR will control for both unmeasured and measured confounding and provide an unbiased estimate (Lawlor et al., 2008; Sheehan et al., 2008; Glymour et al., 2012). The necessity of these assumptions is a potential limitation that is inherent to our approach. However, we are convinced that the assumptions of an association between HbAS and malaria (May et al., 2007; Allison, 1964; Willcox et al., 1983; Hill et al., 1991; Aidoo et al., 2002) and no association between HbAS and stunting other than through malaria (Kramer et al., 1978; Rehan, 1981; Ashcroft et al., 1976, 1978) are valid. Ghansah et al. (2012) have described the HbAS haplotype in a Ghanaian population as an extended haplotype of 1.5 Mb containing 25 additional genes. Their analysis shows that this genomic region has a considerable degree of linkage disequilibrium, which potentially could violate our assumption that HbAS is independent of unmeasured confounders. To identify a potential violation, we searched PubMed for reports on associations between stunting or malnutrition and any of the other 25 genes on the extended haplotype, including possible alternative gene names, allelic variants and resulting phenotypes, based on searches in the National Center for Biotechnology Information (NCBI) gene database and the Online Mendelian Inheritance in Man database (OMIM) (see Kang et al. (2013) for details). These searches did not reveal any reports of an association between genes or genetic variants on the haplotype and stunting.

A further limitation to previous studies is potential reverse causality in the association of stunting and malaria. As discussed by Arinaitwe et al. it is difficult to distinguish whether stunting increases the risk of malaria or whether malaria increases the risk of stunting (Arinaitwe et al., 2012). The Mendelian randomization design of this study solves part of this limitation. It enables us to see whether an increased frequency of malaria causes stunting. Specifically, any association between the sickle cell trait and stunting must come from an effect of malaria on stunting rather than the reverse. The sickle cell trait, which is determined at conception, only affects stunting through its effect on malaria. If malaria did not affect stunting, there would be no association between the sickle cell trait and stunting.

However, there are several additional factors that we cannot analyze or adjust for in our analysis that may have contributed to the differing findings between studies. For example, several studies were of cross-sectional design (Ehrhardt et al., 2006; Deribew et al., 2010; Crookston et al., 2010) and looked at a potential association between current malaria and stunting prevalence. Malaria at the time point of the study may or may not correlate to previous exposure. This correlation is likely to differ by transmission intensity of malaria and this varied from low-seasonal to high-perennial transmission. While the assessment of malaria incidence in the longitudinal studies was probably a more accurate measure of exposure, it seems plausible that the effect of malaria on growth is modulated by immunity and thereby may vary with age (McGregor et al., 1956; Bradley-Moore et al., 1985; ter Kuile et al., 2003; Deen et al., 2002; Arinaitwe et al., 2012; Snow et al., 1991; Fillol et al., 2009; Genton et al., 1998). In fact, a study from Tanzania found an effect modification by age with the strongest effect of malaria on stunting in children less than 1 year of age (Nyakeriga et al., 2004).

A further potential limitation of our model is the measurement of exposure. We have assumed that the simple sum of malaria episodes over a child's life is what affects the child being stunted at age two. It may be that a more complex function of a child's malaria history affects stunting; we plan to investigate this in future work. In addition, the population in this study was enrolled in a clinical trial and seen by medical personnel at close intervals. Prompt medical treatment and nutritional interventions were available free of charge during follow-up. It is possible that the effect of malaria on stunting in this population may differ from the general population and especially from populations where nutritional deficiencies are more common.

The interpretation of the effect ratio assumes that the effect HbAS has on stunting is solely mediated by a reduction of the number of malaria episodes. However, HbAS also reduces the severity of every malaria episode and the effect on stunting may partly be due to this (Kreuels et al., 2010). This would lead to an overestimation of the effect that is attributable to each malaria episode. However, the causality conclusion would not change and even the lower boundary of the 95% confidence interval for the effect, 0.044, still indicates a substantial effect of malaria on stunting.

Our analysis demonstrates the applicability of HbAS as an instrumental variable for the analysis of conditions related to malaria. As in all observational studies, research on the association of malaria with other medical conditions is often difficult due to the strong influence of confounders and randomized trials are almost always impractical. The method we propose can be applied to reanalyze previous studies in this area, specifically those where the genotyping of the sickle cell gene has already been performed (ter Kuile et al., 2003; Nyakeriga et al., 2004). We hope that our findings will encourage the application of MR to such analyses in the future. A potential further application of MR using HbAS is the elucidation of associations between malaria and other infections. One such analysis was performed by Scott et al. (2011) who used MR to analyze an association between malaria and bacteremia caused by Salmonella spp.

Our analysis provides evidence of a substantial causal effect of malaria episodes on stunting, at least in children less than 2 years of age in an area of high endemicity. Our findings will hopefully spur further research on this important epidemiological concern in sub-Saharan Africa and increase the application of sickle cell trait as an instrumental variable in malaria research.

CHAPTER 6 : Discussions

Throughout Chapters 2 to 4, we established results concerning the estimation of causal effects of the exposure on the outcome when invalid instruments are present. In Chapter 2 and 3, we provided results when we have multiple candidate instruments and proposed theoretical limits as well as propose an estimator that can consistently estimate the true causal effect and a confidence interval that has honest coverage. In Chapter 4, we dealt with the case when we have one single candidate instrument and proposed a nonparametric matching estimator. In Chapter 5, we applied our new method to a real data set concerning the causal effect of malaria on stunted growth.

As we seen in our work, there is much room for future work in the area of IV estimation with invalid instruments. Each chapter laid out some potential future works in each scenarios and the list below is a summary of those future directions.

1. Further extend our method on estimation with invalid instruments, specifically under the framework in Chapter 2, to encompass a wider class of models. Our current work is limited to linear, constant effects model and as such, our problem boils down to a mathematical problem of solving system of under-determined linear equations with constraints on the parameter space. We want to explore the theoretical limits and methods for estimation and identification under more general models that include arbitrary transformations of the instruments, such as $g(\mathbf{Z}_{i\cdot})$ in some function class $\mathcal{G}$, heterogeneous effects (i.e. β_i^* instead of a global β^*), and nonlinear outcome models (e.g. binary outcomes or survival outcomes). All these generalizations would be a departure from the usual system of under-determined linear equations and it is unknown whether estimation is possible. In line with this goal, developing an estimator that is robust to model mis-specification would be useful for researchers using IV methods.

2. Develop confidence intervals for IV methods that are both robust to weak instruments and invalid instruments. Weak instruments are, in essence, a near violation of (A1)

and there is a huge literature on weak instruments (see Stock et al. (2002) for a survey). However, there is very little literature on estimation in the presence of invalid instruments, except our work and work by Kolesár et al. (2013). Chapter 3 laid out some preliminary work, but it is currently unknown whether our CIs can be improved.

3. Explore sensitivity analysis when core IV assumptions are violated. In our work in Chapter 4, we used matching to controll for covariates to make assumption (A3) more plausible and we developed sensitivity analysis of our IV estimate should matching fail to make (A3) plausible. We would like to develop sensitivity analysis for other types of violations in IV assumptions, such as (A2), or other assumptions that may arise when we start considering heterogeneous causal effects.

4. Extend our method in Chapter 4 to multiple instruments. Currently, our matching algorithm in IV estimation can only handle binary instruments. However, non-binary instruments are also common in IV studies and we want to explore how to extend our method to this setting.

5. Apply our IV methods to various problems in the social sciences and health-related disciplines. Our work with the malaria data in Chapter 5 was a collaborative effort with medical professional where we applied our new IV matching method to solve a problem of interest in the medical community. We are interested in applying our methods to other settings in social science.

In conclusion, the current theory and methods behind IV estimation with invalid instruments, specifically instruments that violate (A2) and (A3), are very limited and there are many unanswered questions. Also, the work in this area has wide applications in fields where instrumental variables methods are used, which includes economics, biology, epidemiology, psychology, political science, sociology, and many others. It is our hope that the research in the area of IV estimation with invalid instruments will further the field of making causal conclusions from observational data, especially when one only has imperfect instruments.

APPENDIX

A.1. Proofs from Chapter 2

We adopt the following notations for the proofs. For any sets $A, B \subseteq \{1, \ldots, L\}$, denote $A \cap B$ to be the intersection of sets A and B, $A \cup B$ to be the union of sets A and B, and A^C and B^C to be the complement of sets A and B, respectively. If $A \subseteq B$, denote $B \setminus A$ to be the set that comprises of all the elements of B except those that are in A. Let $|A|$ and $|B|$ denote the cardinality of the sets A and B, respectively.

For any vector $\boldsymbol{\alpha} \in \mathbb{R}^L$ and set $A \subseteq \{1, \ldots, L\}$, denote $\boldsymbol{\alpha}_A \in \mathbb{R}^L$ to be the vector where all the elements except whose indices are in A are zero. Also, denote the jth element as α_j. Let $\text{supp}(\boldsymbol{\alpha}) \subseteq \{1, \ldots, L\}$ to be the support of the vector $\boldsymbol{\alpha}$ and $\text{supp}(\boldsymbol{\alpha})^C$ be the complement set. For any matrix $\mathbf{M} \in \mathbb{R}^{n \times L}$ and set $A \subseteq \{1, \ldots, p\}$, let $\mathbf{M}_A \in \mathbb{R}^{n \times L}$ be an n by $|A|$ matrix where the columns are specified by set A.

A.1.1. Proof of Theorem 2.1

First, we prove that, β^* is a unique solution if and only if $\boldsymbol{\alpha}^*$ is a unique solution. Suppose β^* has a unique solution; that is, for any two solutions $\boldsymbol{\alpha}^{(1)}$ $\beta^{(1)}$ and $\boldsymbol{\alpha}^{(2)}, \beta^{(2)}$, in equation (2.7)

$$\boldsymbol{\alpha}^{(1)} + \boldsymbol{\gamma}^* \beta^{(1)} = \boldsymbol{\Gamma}^* \tag{A.1a}$$

$$\boldsymbol{\alpha}^{(2)} + \boldsymbol{\gamma}^* \beta^{(2)} = \boldsymbol{\Gamma}^* \tag{A.1b}$$

we have $\beta^{(1)} = \beta^{(2)}$. Subtracting $\boldsymbol{\gamma}^* \beta^{(1)}$ from equations (A.1) gives $\boldsymbol{\alpha}^{(1)} = \boldsymbol{\alpha}^{(2)}$. Now, suppose $\boldsymbol{\alpha}^*$ is unique, which implies $\boldsymbol{\alpha}^{(1)} = \boldsymbol{\alpha}^{(2)}$. Again, subtracting $\boldsymbol{\alpha}^{(1)}$ from (A.1) reveals $\beta^{(1)} = \beta^{(2)}$.

Second, we prove the necessary and sufficient conditions for Theorem 2.1. Suppose the conditions on $\boldsymbol{\gamma}^*$ and $\boldsymbol{\Gamma}^*$ hold, specifically $q_m = q_{m'}$ for any $m \neq m'$, but there are two distinct sets of parameters, $\boldsymbol{\alpha}^{(1)}, \beta^{(1)}$ and $\boldsymbol{\alpha}^{(2)}, \beta^{(2)}$ that solve the moment equation in equation (A.1). Let $A^{(1)}$ =supp($\boldsymbol{\alpha}^{(1)}$) and $A^{(2)}$ =supp($\boldsymbol{\alpha}^{(2)}$) be the sets of invalid instruments for the two distinct parameter sets, not equal to each other; if the supports are equal to each other, we have the degenerate case whereby from equation (A.1), for any $j \in A^{(1)} = A^{(2)}$ $\gamma_j^* \beta^{(1)} = \Gamma_j^*$ and $\gamma_j^* \beta^{(2)} = \Gamma_j^*$, which implies that $\beta^{(1)} = \beta^{(2)}$ and $\boldsymbol{\alpha}^{(1)} = \boldsymbol{\alpha}^{(2)}$, a contradiction. Because the number of invalid instruments, s, is less than U, $s < U$, the number of valid instruments, $L - s$, must be greater than $L - U$, $L - s > L - U$. Thus, $|(A^{(1)})^C|, |(A^{(2)})^C| > L - U$.

Now, pick any subsets, $(A^{(1')})^C$ and $(A^{(2')})^C$, of $(A^{(1)})^C$ and $(A^{(2)})^C$, respectively, where $|(A^{(1')})^C| = |(A^{(2')})^C| = L - U + 1$. These subsets $(A^{(1')})^C$ and $(A^{(2')})^C$ inherit the following property from their larger sets $(A^{(1)})^C$ and $(A^{(2)})^C$, respectively.

$$\begin{aligned} \alpha_j^{(1)} + \gamma_j^* \beta^{(1)} = \gamma_j^* \beta^{(1)} = \Gamma_j^*, &\quad j \in (A^{(1')})^C \subseteq (A^{(1)})^C \\ \alpha_k^{(2)} + \gamma_k^* \beta^{(2)} = \gamma_k^* \beta^{(2)} = \Gamma_k^*, &\quad k \in (A^{(2')})^C \subseteq (A^{(2)})^C \end{aligned}$$

The condition on $\boldsymbol{\gamma}^*$ and $\boldsymbol{\Gamma}^*$ in Theorem 2.1 state that for any sets C_m with size $|C_m| = L - U + 1$ and with the property that $\gamma_j q_m = \Gamma_j, j \in C_m$, we have $q_m = q_{m'}$ for any m, m'. The subsets we constructed, $(A^{(1')})^C$ and $(A^{(2')})^C$, satisfy these condition with constants $q_{1'} = \beta^{(1)}$ and $q_{2'} = \beta^{(2)}$. Hence, $\beta^{(1)} = q_{1'} = q_{2'} = \beta^{(2)}$, which is a contradiction. Hence, the two sets of parameters $\boldsymbol{\alpha}^{(1)}, \beta^{(1)}$ and $\boldsymbol{\alpha}^{(2)}, \beta^{(2)}$ are identical to each other and the solution is unique.

Now, suppose the solution is unique. Then, we show that the conditions on $\boldsymbol{\gamma}^*$ and $\boldsymbol{\Gamma}^*$ must hold. Pick any two sets $A^{(1)}, A^{(2)} \subseteq \{1, \ldots, L\}$ with their complements having the size $|(A^{(1)})^C| = |(A^{(2)})^C| = L - U + 1$ and corresponding constants q_1 and q_2, respectively, defined in the theorem. We have to show that $q_1 = q_2$ for any pair of two sets.

Note that at least one set of these sets and its corresponding constant q must exist because at the true parameter values, $\boldsymbol{\alpha}^*$ and β^*, equation (2.7) is satisfied. Specifically, if

A^* =supp($\boldsymbol{\alpha}^*$) where, by $s < U$, $|(A^*)^C| = |\text{supp}(\boldsymbol{\alpha}^*)^C| > L - U$, we can take any subset $(A^{(*')})^C \subseteq (A^*)^C$ of size $|(A^{(*')})^C| = L - U + 1$. For any $j \in (A^{(*')})^C$, by equation (2.7), $\gamma_j^* \beta^* = \Gamma_j^*$ and thus, its corresponding constant $q_{*'}$ is $q_{*'} = \beta^*$. If there is exactly one set $A^{(1)}$, the condition holds automatically.

Suppose there are two or more sets and let $A^{(1)}$ and $^{(2)}$ be any pair of the sets. Based on the sets $A^{(1)}$ and $A^{(2)}$ and their corresponding constants q_1 and q_2, we construct the following sets of parameters $\boldsymbol{\alpha}^{(1)}, \beta^{(1)}$ and $\boldsymbol{\alpha}^{(2)}, \beta^{(2)}$

$$\beta^{(1)} = q_1, \quad \alpha_j^{(1)} = \begin{cases} 0 & j \in (A^{(1)})^C \\ \Gamma_j^* - q_1\gamma_j^* & j \in A^{(1)} \end{cases}$$

$$\beta^{(2)} = q_2, \quad \alpha_j^{(2)} = \begin{cases} 0 & j \in (A^{(2)})^C \\ \Gamma_j^* - q_2\gamma_j^* & j \in A^{(2)} \end{cases}$$

The cardinality of $\boldsymbol{\alpha}^{(1)}$ and $\boldsymbol{\alpha}^{(2)}$ are less than U. In addition, they satisfy the moment equation in equation (2.7).

$$\alpha_j^{(1)} + \gamma_j^* \beta^{(1)} = \begin{cases} \gamma_j^* q_1 = \Gamma_j^* & j \in (A^{(1)})^C \\ \Gamma_j^* - q_1\gamma_j^* + \gamma_j^* q_1 = \Gamma_j^* & j \in A^{(1)} \end{cases}$$

$$\alpha_j^{(2)} + \gamma_j^* \beta^{(2)} = \begin{cases} \gamma_j^* q_2 = \Gamma_j^* & j \in (A^{(2)})^C \\ \Gamma_j^* - q_2\gamma_j^* + \gamma_j^* q_2 = \Gamma_j^* & j \in A^{(2)} \end{cases}$$

Since the equation has only one unique solution, this implies that $\beta^{(1)} = \beta^{(2)}$, or $q_1 = q_2$. Since this holds for any two sets $(A^{(1)})^C, (A^{(2)})^C$ with constants q_1 and q_2 and cardinality $L - U + 1$, we arrive at the condition $q_m = q_{m'}$ for any m, m'. □

A.1.2. Proof of Corollary 2.1

Consider any two sets C_m and $C_{m'}$ with the constants q_m and $q_{m'}$ in Theorem 2.1. Take an element j from the intersection $C_m \cap C_{m'}$; this intersection is non-empty because $|C_m| = |C_{m'}| = L - U + 1 \geq L/2 + 1$. At element $j \in C_m \cap C_{m'}$, we have $\gamma_j^* q_m = \Gamma_j^*$ and $\gamma_j^* q_{m'} = \Gamma_j^*$, which implies $q_m = q_{m'}$. Since this holds for any two sets C_m and $C_{m'}$, $q_m = q_{m'}$ for m, m', the condition in Theorem 2.1 always holds whenever $U \geq L/2$ and we have identification. □

A.1.3. Proof of Theorem 2.2

We begin by introducing some notations and terminologies. For $\boldsymbol{\alpha} \in \mathbb{R}^p$ and $s \in \{1, \dots, p\}$, $\boldsymbol{\alpha}_{\max(s)}$ is defined as the vector where all but the largest s elements set to zero and $\boldsymbol{\alpha}_{-\max(s)}$ is defined as $\boldsymbol{\alpha} - \boldsymbol{\alpha}_{\max(s)}$.

Definition A.1. The restricted orthogonal constant (ROC) of single matrix of order k_1 and k_2, denoted as $\theta_{k_1,k_2}(\mathbf{M})$, is the smallest $\theta_{k_1,k_2}(\mathbf{M})$ where for any k_1-sparse vector $\boldsymbol{\alpha}_1$ and k_2-sparse vector $\boldsymbol{\alpha}_2$ with non-overlapping support, we have

$$|\langle \mathbf{M}\boldsymbol{\alpha}_1, \mathbf{M}\boldsymbol{\alpha}_2 \rangle| \leq \theta_{k_1,k_2}(\mathbf{M}) \|\boldsymbol{\alpha}_1\|_2 \|\boldsymbol{\alpha}_2\|_2.$$

Next, we introduce two lemmas. The first lemma relates the RIP and ROC constants.

Lemma A.1. *For any matrix* $\mathbf{M}$ *and positive integers* s_1 *and* s_2*,*

$$\theta_{s_1,s_2}(\mathbf{M}) \leq \frac{1}{2} \left(\delta^+_{s_1+s_2}(\mathbf{M}) - \delta^-_{s_1+s_2}(\mathbf{M}) \right).$$

Proof. For any vectors x and y with disjoint supports and $\|x\|_2 = \|y\|_2 = 1$, we must have

$x+y$, $x-y$ are both (s_1+s_2)-sparse and $\|x+y\|_2^2 = \|x-y\|_2^2 = 2$. Hence,

$$\begin{aligned}
|\langle \mathbf{M}x, \mathbf{M}y\rangle| =& \frac{1}{4}\left|\|\mathbf{M}(x+y)\|_2^2 - \|\mathbf{M}(x-y)\|_2^2\right| \\
=& \frac{1}{4}\max\left\{\|\mathbf{M}(x+y)\|_2^2 - \|\mathbf{M}(x-y)\|_2^2, \|\mathbf{M}(x-y)\|_2^2 - \|\mathbf{M}(x+y)\|_2^2\right\} \\
\leq& \frac{1}{4}\max\Big\{\delta^+_{s_1+s_2}(\mathbf{M})\|x+y\|_2^2 - \delta^-_{s_1+s_2}(\mathbf{M})\|x-y\|_2^2, \\
& \delta^+_{s_1+s_2}(\mathbf{M})\|x-y\|_2^2 - \delta^-_{s_1+s_2}(\mathbf{M})\|x+y\|_2^2\Big\} \\
\leq& \frac{1}{2}\left(\delta^+_{s_1+s_2}(\mathbf{M}) - \delta^-_{s_1+s_2}(\mathbf{M})\right),
\end{aligned}$$

which implies $\theta_{s_1,s_2}(\mathbf{M}) \leq \frac{1}{2}\left(\delta^+_{s_1+s_2}(\mathbf{M}) - \delta^-_{s_1+s_2}(\mathbf{M})\right)$. □

The second lemma proves a standard property of the Lasso.

Lemma A.2. *Suppose we have the model $Y_i = \mathbf{Z}_{i.}^T\boldsymbol{\alpha}^* + \epsilon_i$ where $\boldsymbol{\alpha}^*$ is s-sparse. Further suppose that matrix $\mathbf{Z}$ has upper and lower RIP constants $\delta_s^+(\mathbf{Z})$ and $\delta_s^-(\mathbf{Z})$, respectively. Define $\hat{\boldsymbol{\alpha}}$ as the Lasso estimator*

$$\hat{\boldsymbol{\alpha}}_\lambda = \operatorname*{argmin}_{\boldsymbol{\alpha}} \frac{1}{2}\|Y - \mathbf{Z}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1 \tag{A.2}$$

and let $h = \hat{\boldsymbol{\alpha}}_\lambda - \boldsymbol{\alpha}^*$ measure the errors of the estimator.

If $r\|\mathbf{Z}^T\boldsymbol{\epsilon}\|_\infty \leq \lambda$ for some $r > 1$, we have

$$\|h_{-\max(s)}\|_1 \leq \frac{r+1}{r-1}\|h_{\max(s)}\|_1. \tag{A.3}$$

Furthermore, if $(r+1)\delta_{2s}^+(\mathbf{Z}) < (3r-1)\delta_{2s}^-(\mathbf{Z})$,

$$\|h_{\max(s)}\|_2 \leq \frac{2\lambda\sqrt{s}(r-1)(r+1)/r}{(3r-1)\delta_{2s}^-(\mathbf{Z}) - (r+1)\delta_{2s}^+(\mathbf{Z})}. \tag{A.4}$$

Proof. Since $\hat{\boldsymbol{\alpha}}_\lambda$ is the minimizer of (A.2) , we have

$$\frac{1}{2}\|Y - \mathbf{Z}\hat{\boldsymbol{\alpha}}_\lambda\|_2^2 + \lambda\|\hat{\boldsymbol{\alpha}}_\lambda\|_1 \leq \frac{1}{2}\|y - \mathbf{Z}\boldsymbol{\alpha}^*\|_2^2 + \lambda\|\boldsymbol{\alpha}^*\|_1.$$

By the assumed model $Y_i = \mathbf{Z}_i^T \boldsymbol{\alpha}^* + \epsilon_i$, we have

$$\frac{1}{2}\left(\|\boldsymbol{\epsilon} - \mathbf{Z}h\|_2^2 - \|\boldsymbol{\epsilon}\|_2^2\right) \le \lambda(\|\boldsymbol{\alpha}^*\|_1 - \|\hat{\boldsymbol{\alpha}}_\lambda\|_1). \tag{A.5}$$

For the upper bound of (A.5), the fact that $\boldsymbol{\alpha}^*$ is s-sparse gives a useful bound. Specifically,

$$\begin{aligned}
\|\boldsymbol{\alpha}^*\|_1 - \|\hat{\boldsymbol{\alpha}}_\lambda\|_1 &= \|\boldsymbol{\alpha}^*_{supp(\boldsymbol{\alpha}^*)}\|_1 - \|\hat{\boldsymbol{\alpha}}_{supp(\boldsymbol{\alpha}^*)}\|_1 - \|\hat{\boldsymbol{\alpha}}_{supp(\boldsymbol{\alpha}^*)^c}\|_1 \\
&\le \|\boldsymbol{\alpha}^*_{supp(\boldsymbol{\alpha}^*)} - \hat{\boldsymbol{\alpha}}_{supp(\boldsymbol{\alpha}^*)}\|_1 - \|h_{supp(\boldsymbol{\alpha}^*)^c}\|_1 \\
&\le \|h_{supp(\boldsymbol{\alpha}^*)}\|_1 - \|h_{supp(\boldsymbol{\alpha}^*)^c}\|_1 \\
&\le \|h_{\max(s)}\|_1 - \|h_{-\max(s)}\|_1.
\end{aligned}$$

For the lower bound of (A.5), $\|\boldsymbol{\epsilon} - \mathbf{Z}h\|_2^2 - \|\boldsymbol{\epsilon}\|_2^2$, we can simplify as

$$\begin{aligned}
\frac{1}{2}\left(\|\boldsymbol{\epsilon} - \mathbf{Z}h\|_2^2 - \|\boldsymbol{\epsilon}\|_2^2\right) &= -\frac{1}{2}(\mathbf{Z}h)^T(2\boldsymbol{\epsilon} - \mathbf{Z}h) \ge -h^T\mathbf{Z}^T\boldsymbol{\epsilon} \ge -\|\mathbf{Z}^T\boldsymbol{\epsilon}\|_\infty \|h\|_1 \\
&= -\|\mathbf{Z}^T\boldsymbol{\epsilon}\|_\infty(\|h_{\max(s)}\|_1 + \|h_{-\max(s)}\|_1).
\end{aligned}$$

Hence, by (A.5) and the condition $r\|\mathbf{Z}^T\boldsymbol{\epsilon}\|_\infty \le \lambda$ where $r > 1$, we have

$$r(\|h_{\max(s)}\|_1 - \|h_{-\max(s)}\|_1) \ge -(\|h_{\max(s)}\|_1 + \|h_{-\max(s)}\|_1).$$

which yields (A.3), the first part of the theorem.

For (A.4), the second part of the theorem, suppose $(r+1)\delta^+_{2s}(\mathbf{Z}) < (3r-1)\delta^-_{2s}(\mathbf{Z})$ holds. By the Karush-Kuhn-Tucker (KKT) condition of the minimization problem in (A.2), we we have $\|\mathbf{Z}^T(y - \mathbf{Z}\hat{\boldsymbol{\alpha}})\|_\infty \le \lambda$ and

$$\|\mathbf{Z}^T\mathbf{Z}h\|_\infty \le \|\mathbf{Z}^T(y - \mathbf{Z}\hat{\boldsymbol{\alpha}})\|_\infty + \|\mathbf{Z}^T(y - \mathbf{Z}\boldsymbol{\alpha}^*)\|_\infty \le \lambda + \|\mathbf{Z}^T\boldsymbol{\epsilon}\|_\infty.$$

Lemma 5.1 in Cai and Zhang (2013a) with $\lambda = \max(\|h_{-\max(s)}\|_\infty, \|h_{-\max(s)}\|_1/s)$ implies

$$\begin{aligned}
|\langle \mathbf{Z}h_{\max(s)}, \mathbf{Z}h_{-\max(s)}\rangle| &\leq \theta_{s,s}(\mathbf{Z})\|h_{\max(s)}\|_2 \cdot \sqrt{s} \cdot \max(\|h_{-\max(s)}\|_\infty, \|h_{-\max(s)}\|_1/s) \\
&\leq \sqrt{s}\theta_{s,s}(\mathbf{Z})\|h_{\max(s)}\|_2 \cdot \frac{r+1}{r-1}\|h_{\max(s)}\|_1/s \\
&\leq \theta_{s,s}(\mathbf{Z})\frac{r+1}{r-1}\|h_{\max(s)}\|_2^2,
\end{aligned}$$

where the last inequality uses (A.3). We then have

$$\begin{aligned}
\sqrt{s}(\lambda + \|\mathbf{Z}^T\boldsymbol{\epsilon}\|_\infty)\|h_{\max(s)}\|_2 \geq (\lambda + \|\mathbf{Z}^T\boldsymbol{\epsilon}\|_\infty)\|h_{\max(s)}\|_1 &\geq \langle \mathbf{Z}^T\mathbf{Z}h, h_{\max(s)}\rangle \\
&= \langle \mathbf{Z}h_{\max(s)}, \mathbf{Z}h_{\max(s)}\rangle + \langle \mathbf{Z}h_{\max(s)}, \mathbf{Z}h_{-\max(s)}\rangle \\
&\geq \|\mathbf{Z}h_{\max(s)}\|_2^2 - \theta_{s,s}\frac{r+1}{r-1}\|h_{\max(s)}\|_2^2 \\
&= \left(\delta_{2s}^-(\mathbf{Z}) - \theta_{s,s}(\mathbf{Z})\frac{r+1}{r-1}\right)\|h_{\max(s)}\|_2^2 \\
&\geq \left(\frac{3r-1}{2(r-1)}\delta_{2s}^-(\mathbf{Z}) - \frac{r+1}{2(r-1)}\delta_{2s}^+\right)\|h_{\max(s)}\|_2^2,
\end{aligned}$$

where the last inequality uses Lemma A.1. Moving $\|h_{\max(s)}\|$ to the right hand side and using the condition $r\|\mathbf{Z}^T\boldsymbol{\epsilon}\|_\infty \leq \lambda$ where $r > 1$ yields (A.4). $\square$

Now we move on to the proof of Theorem 2.2. Section 2.3.6 in the main paper states that the original estimation method can be reinterpreted as a two-step method where the first step is the Lasso step and the second step is a dot product. The proof will first analyze step 1 using the lemmas about Lasso performance and use it to analyze step 2.

First, in lieu of step 1, the model in equation (2.3) from the original paper can be modified to

$$\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{P}_{\mathbf{Z}}Y = \mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{Z}\boldsymbol{\alpha}^* + \mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{P}_{\mathbf{Z}}\boldsymbol{\epsilon}. \tag{A.6}$$

Here, $\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{Z}$ becomes the design matrix, $\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{P}_{\mathbf{Z}}Y$ becomes the outcome, and $\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{P}_{\mathbf{Z}}\boldsymbol{\epsilon}$ is the new error term. In addition, from the condition $3\|\mathbf{Z}^T\mathbf{P}_{\hat{\mathbf{D}}^\perp}\boldsymbol{\epsilon}\| \leq \lambda$, we have

$$\lambda \geq 3\|\mathbf{Z}^T(I - \mathbf{P}_{\hat{\mathbf{D}}})\boldsymbol{\epsilon}\|_\infty = 3\|\mathbf{Z}^T(\mathbf{P}_\mathbf{Z} - \mathbf{P}_{\hat{\mathbf{D}}})\boldsymbol{\epsilon}\|_\infty = 3\|\mathbf{Z}^T(I - \mathbf{P}_{\hat{\mathbf{D}}})\mathbf{P}_\mathbf{Z}\boldsymbol{\epsilon}\|_\infty = 3\|(\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{Z})^T\mathbf{P}_\mathbf{Z}\boldsymbol{\epsilon}\|_\infty.$$

Second, note that (A.9) is in terms of the RIP constants of $\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{Z}$. To relate the RIP constants of $\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{Z}$ with that of $\mathbf{Z}$, we see that for any $2s$-sparse vector $x \in \mathbb{R}^L$, $\|\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{Z}x\|_2^2 = \|\mathbf{Z}x\|_2^2 - \|\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z}x\|_2^2 \leq \|\mathbf{Z}x\|_2^2 \leq \delta_{2s}^+(\mathbf{Z})\|x\|_2^2$. By the definition of $\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{Z})$, this implies

$$\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{Z}) \leq \delta_{2s}^+(\mathbf{Z}). \tag{A.7}$$

In addition, we have $\|\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{Z}x\|_2^2 = \|\mathbf{Z}x\|_2^2 - \|\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z}x\|_2^2 \geq \delta_{2s}^-(\mathbf{Z})\|x\|_2^2 - \delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z})\|x\|_2^2$. By the definition of $\delta_{2s}^-(\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{Z})$, this also implies

$$\delta_{2s}^-(\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{Z}) \geq \delta_{2s}^-(\mathbf{Z}) - \delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z}). \tag{A.8}$$

Combining (A.7), (A.8) with assumption that $2\delta_{2s}^-(\mathbf{Z}) > \delta_{2s}^+(\mathbf{Z}) + 2\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z})$, we know $2\delta_{2s}^-(\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{Z}) > \delta_{2s}^-(\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{Z})$. By Lemma A.2, where we set $r = 3$ in assumption $r\|\mathbf{Z}^T\boldsymbol{\epsilon}\|_\infty \leq \lambda$ and the model is rewritten as (A.6),

$$\|h_{\max(s)}\|_2 \leq \frac{4/3\lambda\sqrt{s}}{2\delta_{2s}^-(\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{Z}) - \delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{Z})} \tag{A.9}$$

and

$$\|h_{-\max(s)}\|_1 \leq 2\|h_{\max(s)}\|_1. \tag{A.10}$$

Combining the RIP relations established by (A.7) and (A.8), we can rewrite (A.9) as

$$\|h_{\max(s)}\|_2 \leq \frac{4/3\lambda\sqrt{s}}{2\delta_{2s}^-(\mathbf{Z}) - \delta_{2s}^+(\mathbf{Z}) - 2\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z})}. \tag{A.11}$$

Third, we establish a bound for $\|\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z}h\|_2$. This bound is needed to bound step 2 in Section

2.3.5 of the original paper because

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

Rearranging terms and taking norms on both sides give

$$\|\hat{\beta}_\lambda - \beta^*\|_2 \leq \frac{\|\hat{\mathbf{D}}^T \mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z}h\|_2}{\|\hat{\mathbf{D}}\|_2^2} + \frac{\|\hat{\mathbf{D}}^T \mathbf{P}_{\hat{\mathbf{D}}}\boldsymbol{\epsilon}\|_2}{\|\hat{\mathbf{D}}\|_2^2} \leq \frac{\|\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z}h\|_2}{\|\hat{\mathbf{D}}\|_2} + \frac{|\hat{\mathbf{D}}^T \boldsymbol{\epsilon}|}{\|\hat{\mathbf{D}}\|_2^2}. \tag{A.12}$$

Hence, a bound on $\|\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z}h\|_2$ is necessary to bound $\|\hat{\beta}_\lambda - \beta^*\|_2$. To start off, we apply Lemma 1.1 in Cai and Zhang (2013b) to represent $h_{-\max(s)}$ as a weighted mean of s-sparse vectors. This lemma allows us to convert the bound for $h_{\max(s)}$ in (A.11) to the bound for $\|\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z}h\|_2$. Specifically, the lemma states we can find $\lambda_i \geq 0$ and s-sparse $v_i \in \mathbb{R}^L$ where $i = 1, \ldots, N$ such that $\sum_{i=1}^N \lambda_i = 1$ and $h_{-\max(s)} = \sum_{i=1}^N \lambda_i v_i$. Hence, $h = \sum_{i=1}^N \lambda_i (h_{\max(s)} + v_i)$. Furthermore, we have

$$supp(v_i) \subseteq supp(h_{-\max(s)}), \quad \|v_i\|_\infty \leq \max\left(\|h_{-\max(s)}\|_\infty, \frac{\|h_{-\max(s)}\|_1}{s}\right)$$

and

$$\|v_i\|_1 = \|h_{-\max(s)}\|_1,$$

which yields

$$\|v_i\|_\infty \leq \max\left(\frac{\|h_{\max(s)}\|_1}{s}, \frac{2\|h_{\max(s)}\|_1}{s}\right) = \frac{2\|h_{\max(s)}\|_1}{s}, \quad \|v_i\|_1 \leq 2\|h_{\max(s)}\|_1$$

and $\|h_{\max(s)} + v_i\|_2^2 = \|h_{\max(s)}\|_2^2 + \|v_i\|_2^2 \leq \|h_{\max(s)}\|_2^2 + \|v_i\|_1 \|v_i\|_\infty \leq 5\|h_{\max(s)}\|_2^2$. Combining all these together with (A.11), we have

$$\begin{aligned}
\|\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z}h\|_2 &\leq \sum_{i=1}^{N} \lambda_i \|\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z}(h_{\max(s)} + v_i)\|_2 \leq \sum_{i=1}^{N} \lambda_i \sqrt{5\delta_{2s}^{+}(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z})} \|h_{\max(s)}\|_2 \\
&\leq \sqrt{5\delta_{2s}^{+}(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z})} \frac{4/3\lambda\sqrt{s}}{2\delta_{2s}^{-}(\mathbf{Z}) - \delta_{2s}^{+}(\mathbf{Z}) - 2\delta_{2s}^{+}(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z})} \\
&= \frac{4\sqrt{5}/3\lambda\sqrt{s\delta_{2s}^{+}(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z})}}{2\delta_{2s}^{-}(\mathbf{Z}) - \delta_{2s}^{+}(\mathbf{Z}) - 2\delta_{2s}^{+}(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z})}.
\end{aligned}$$

Finally, using the relation (A.12) gives us the desired bound for Theorem 2.2. □

Of independent interest is that the proof of Theorem 2.2 can be generalized to a matrix of $\mathbf{D}$ instead of a vector of $\mathbf{D}$. That is, the proof can consider models where there are more than one endogenous variables in the data-generating model. However, for clarity of presentation, we don't explore this route.

A.1.4. Proof of Corollary 2.2

Now, we establish Corollary 2.2 as a corollary to Theorem 2.2. Specifically, the task is to convert the RIP constants $\delta_{2s}^{+}(\mathbf{Z})$, $\delta_{2s}^{-}(\mathbf{Z})$, $\delta_{2s}^{+}(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z})$ and the constraint of $2\delta_{2s}^{-}(\mathbf{Z}) - \delta_{2s}^{+}(\mathbf{Z}) - 2\delta_{2s}^{+}(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z}) > 0$ into μ and a similar constraint on s. To do this, note that for any s-sparse vector $\boldsymbol{\alpha}$

$$\begin{aligned}
\|\mathbf{Z}\boldsymbol{\alpha}\|_2^2 &= \sum_{j \in supp(\boldsymbol{\alpha})} \|\mathbf{Z}_{.j}\|_2^2 \boldsymbol{\alpha}_j^2 + \sum_{i<j, i,j \in supp(\boldsymbol{\alpha})} 2\boldsymbol{\alpha}_i \boldsymbol{\alpha}_j \langle \mathbf{Z}_{.i}, \mathbf{Z}_{.j} \rangle \\
&\leq \sum_{j \in supp(\boldsymbol{\alpha})} \boldsymbol{\alpha}_j^2 + \sum_{i<j, i,j \in supp(\boldsymbol{\alpha})} (\boldsymbol{\alpha}_i^2 + \boldsymbol{\alpha}_j^2)\mu \\
&= (1 + (s-1)\mu) \sum_{j \in supp(\boldsymbol{\alpha})} \boldsymbol{\alpha}_j^2 = (1 + (s-1)\mu)\|\boldsymbol{\alpha}\|_2^2
\end{aligned}$$

and

$$\begin{aligned}
\|\mathbf{Z}\boldsymbol{\alpha}\|_2^2 &= \sum_{j \in supp(\boldsymbol{\alpha})} \|\mathbf{Z}_{.j}\|_2^2 \boldsymbol{\alpha}_j^2 + \sum_{i<j, i,j \in supp(\boldsymbol{\alpha})} 2\boldsymbol{\alpha}_i \boldsymbol{\alpha}_j \langle \mathbf{Z}_{.i}, \mathbf{Z}_{.j} \rangle \\
&\geq \sum_{j \in supp(\boldsymbol{\alpha})} \boldsymbol{\alpha}_j^2 - \sum_{i<j, i,j \in supp(\boldsymbol{\alpha})} (\boldsymbol{\alpha}_i^2 + \boldsymbol{\alpha}_j^2)\mu \\
&= (1 - (s-1)\mu)\|\boldsymbol{\alpha}\|_2^2.
\end{aligned}$$

The upper and lower bounds on $\|\mathbf{Z}\boldsymbol{\alpha}\|_2^2$ imply

$$\delta_s^+(\mathbf{Z}) \leq (1 + (s-1)\mu), \quad \text{and} \quad \delta_s^-(\mathbf{Z}) \geq (1 - (s-1)\mu);$$

For $\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{Z}$ and all $2s$-sparse vector x, we have

$$\begin{aligned}
\|\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z}x\|_2^2 &\leq \left(\sum_{j \in supp(x)} \|\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z}_{.j}x_j\|_2 \right)^2 \leq 2s \sum_{j \in supp(x)} \|\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z}_{.j}x_j\|_2^2 \\
&= 2s \sum_{j \in supp(x)} \|\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z}_{.j}\|_2^2 x_j^2 = 2s \sum_{j \in supp(x)} \frac{\|\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z}_{.j}\|_2^2}{\|\mathbf{Z}_{.j}\|_2^2} \|\mathbf{Z}_{.j}x_j\|_2^2 \\
&\leq 2s\rho^2 \delta_1^+(\mathbf{Z}) \sum_{j \in supp(x)} x_j^2 \leq 2s\rho^2 \delta_{2s}^+(\mathbf{Z})\|x\|_2^2.
\end{aligned}$$

Again, by the definition of $\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z})$, this implies that

$$\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z}) \leq 2s\rho^2 \delta_{2s}^+(\mathbf{Z}). \tag{A.13}$$

Under the condition $s < \min\left(\frac{1}{12\mu}, \frac{1}{10\rho^2}\right)$, the denominator of the bound in Theorem 2.2

becomes

$$\begin{aligned}2\delta_{2s}^{-}(\mathbf{Z}) - \delta_{2s}^{+}(\mathbf{Z}) - 2\delta_{2s}^{+}(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z}) &\geq 2\delta_{2s}^{-}(\mathbf{Z}) - (1+4s\rho^2)\delta_{2s}^{+}(\mathbf{Z}) \\ &\geq 2(1-(2s-1)\mu) - (1+4s\rho^2)(1+(2s-1)\mu) \\ &= 1 - 6s\mu + 3\mu - 4s\rho^2 - 8s^2\rho^2\mu + 4s\rho^2\mu \\ &\geq 1 - 6s\mu - 5s\rho^2 > 0.\end{aligned}$$

For the numerator of the bound in Theorem 2, we have

$$\begin{aligned}\frac{4\sqrt{5}}{3}\lambda\sqrt{s\delta_{2s}^{+}(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z})} \leq \frac{4\sqrt{5}}{3}\lambda\sqrt{2s^2\rho^2\delta_{2s}^{+}(\mathbf{Z})} &\leq \frac{4\sqrt{10}}{3}\lambda s\rho\sqrt{1+(2s-1)\mu} \\ \leq \frac{4\sqrt{10}}{3}\lambda s\rho\sqrt{1+2s\mu} &\leq \frac{4\sqrt{10}}{3}\lambda s\rho\sqrt{1+1/6} = \frac{4\sqrt{105}}{9}\lambda s\rho.\end{aligned}$$

Combining them together leads to the desired bound. Note that one can improve the constants in the constraint of s with a bit more care on the above inequalities. □

A.1.5. Proof of Theorem 2.3

The original estimation method can be rewritten as follows

$$\begin{aligned}\hat{\boldsymbol{\alpha}}_\lambda, \hat{\beta}_\lambda =& \underset{\boldsymbol{\alpha},\beta}{\text{argmin}}\ \frac{1}{2}||\mathbf{P}_{\mathbf{Z}}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha} - \mathbf{D}\beta)||_2^2 + \lambda||\boldsymbol{\alpha}||_1 \\ =& \underset{\boldsymbol{\alpha},\beta}{\text{argmin}}\ \frac{1}{2}||(\mathbf{P}_{\hat{\mathbf{D}}} + \mathbf{P}_{\hat{\mathbf{D}}^\perp})\mathbf{P}_{\mathbf{Z}}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha} - \mathbf{D}\beta)||_2^2 + \lambda||\boldsymbol{\alpha}||_1 \\ =& \underset{\boldsymbol{\alpha},\beta}{\text{argmin}}\ \frac{1}{2}||\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{P}_{\mathbf{Z}}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha} - \mathbf{D}\beta)||_2^2 + \frac{1}{2}||\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{P}_{\mathbf{Z}}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha} - \mathbf{D}\beta)||_2^2 + \lambda||\boldsymbol{\alpha}||_1 \\ =& \underset{\boldsymbol{\alpha},\beta}{\text{argmin}}\ \frac{1}{2}||\mathbf{P}_{\hat{\mathbf{D}}}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha}) - \hat{\mathbf{D}}\beta||_2^2 + \frac{1}{2}||\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{P}_{\mathbf{Z}}\mathbf{Y} - \mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{Z}\boldsymbol{\alpha}||_2^2 + \lambda||\boldsymbol{\alpha}||_1.\end{aligned}$$

The first term, $\frac{1}{2}||\mathbf{P}_{\hat{\mathbf{D}}}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha}) - \hat{\mathbf{D}}\beta||_2^2$ is always zero for any given $\boldsymbol{\alpha} \in \mathbb{R}^L$ because $\mathbf{P}_{\hat{\mathbf{D}}}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha})$ lies in the span of $\hat{\mathbf{D}}$ and thus, we can pick β such that the first term is zero. The second term, $\frac{1}{2}||\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{P}_{\mathbf{Z}}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha})||_2^2 + \lambda||\boldsymbol{\alpha}||_1$, is the traditional Lasso problem where the outcome is $\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{P}_{\mathbf{Z}}\mathbf{Y}$ and the design matrix is $\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{Z}$. Hence, the minimizer for this

Lasso problem is also the minimizer for the original method. □

A.2. Proofs from Chapter 4

A.2.1. Proof of Theorem 4.1

Proof. By (A3), we have

$$
\begin{aligned}
&E(R_{ij}|Z_{ij}=1,\mathcal{F},\mathcal{Z})-E(R_{ij}|Z_{ij}=0,\mathcal{F},\mathcal{Z})\\
=&r_{1ij}^{(d_{1ij})}-r_{0ij}^{(d_{0ij})}\\
=&\sum_{k=0}^{M}r_{1ij}^{(k)}\chi(d_{1ij}=k)-\sum_{k=0}^{M}r_{0ij}^{(k)}\chi(d_{0ij}=k)\\
=&\sum_{k=0}^{M}r_{1ij}^{(k)}\{\chi(d_{1ij}\geq k)-\chi(d_{1ij}\geq k+1)\}-\sum_{k=0}^{M}r_{0ij}^{(k)}\{\chi(d_{0ij}\geq k)-\chi(d_{0ij}\geq k+1)\}
\end{aligned}
$$

By (A2), $r_{1ij}^{(k)}=r_{0ij}^{(k)}$ for all k. Then, we have

$$
\begin{aligned}
&\sum_{k=0}^{M}r_{ij}^{(k)}\{\chi(d_{1ij}\geq k)-\chi(d_{1ij}\geq k+1)-\chi(d_{0ij}\geq k)+\chi(d_{0ij}\geq k+1)\}\\
=&\sum_{k=0}^{M}r_{ij}^{(k)}\{\chi(d_{1ij}\geq k)-\chi(d_{0ij}\geq k)\}-\sum_{k=0}^{M}r_{ij}^{(k)}\{\chi(d_{1ij}\geq k+1)-\chi(d_{0ij}\geq k+1)\}\\
=&\sum_{k=1}^{M}r_{ij}^{(k)}\{\chi(d_{1ij}\geq k)-\chi(d_{0ij}\geq k)\}-\sum_{k=1}^{M}r_{ij}^{(k-1)}\{\chi(d_{1ij}\geq k)-\chi(d_{0ij}\geq k)\}\\
=&\sum_{k=1}^{M}(r_{ij}^{(k)}-r_{ij}^{(k-1)})\{\chi(d_{1ij}\geq k)-\chi(d_{0ij}\geq k)\}
\end{aligned}
$$

By monotonicity, $d_{1ij} \geq d_{0ij}$ for all i, j. Then,

$$\begin{aligned}
&\sum_{k=1}^{M}(r_{ij}^{(k)} - r_{ij}^{(k-1)})\{\chi(d_{1ij} \geq k) - \chi(d_{0ij} \geq k)\} \\
=&\sum_{k=1}^{M}(r_{ij}^{(k)} - r_{ij}^{(k-1)})\chi\{\chi(d_{1ij} \geq k) - \chi(d_{0ij} \geq k) = 1\} \\
=&\sum_{k=1}^{M}(r_{ij}^{(k)} - r_{ij}^{(k-1)})\chi(d_{1ij} \geq k > d_{0ij})
\end{aligned}$$

Similarly, by (A3), the expected differences between $Z_{ij} = 1$ and $Z_{ij} = 0$ for the exposure D_{ij} can be written as

$$\begin{aligned}
&E(D_{ij}|Z_{ij} = 1, \mathcal{F}, \mathcal{Z}) - E(D_{ij}|Z_{ij} = 0, \mathcal{F}, \mathcal{Z}) \\
=&d_{1ij} - d_{0ij} \\
=&\sum_{k=0}^{M} k\chi(d_{1ij} = k) - \sum_{k=0}^{M} k\chi(d_{0ij} = k) \\
=&\sum_{k=0}^{M} k\{\chi(d_{1ij} \geq k) - \chi(d_{1ij} \geq k+1)\} - \sum_{k=0}^{M} k\{\chi(d_{0ij} \geq k) - \chi(d_{0ij} \geq k+1)\} \\
=&\sum_{k=0}^{M} k\{\chi(d_{1ij} \geq k) - \chi(d_{1ij} \geq k+1) - \chi(d_{0ij} \geq k) + \chi(d_{0ij} \geq k+1)\} \\
=&\sum_{k=0}^{M} k\{\chi(d_{1ij} \geq k) - \chi(d_{0ij} \geq k)\} - \sum_{k=0}^{M} k\{\chi(d_{1ij} \geq k+1) - \chi(d_{0ij} \geq k+1)\} \\
=&\sum_{k=1}^{M} k\{\chi(d_{1ij} \geq k) - \chi(d_{0ij} \geq k)\} - \sum_{k=1}^{M}(k-1)\{\chi(d_{1ij} \geq k) - \chi(d_{0ij} \geq k)\} \\
=&\sum_{k=1}^{M}\{\chi(d_{1ij} \geq k) - \chi(d_{0ij} \geq k)\}
\end{aligned}$$

By monotonicity, we have

$$\sum_{k=1}^{M}\{\chi(d_{1ij} \geq k) - \chi(d_{0ij} \geq k)\} = \sum_{k=1}^{M}\chi(d_{1ij} \geq k > d_{0ij})$$

Thus, we end up with

$$\begin{aligned}
&\frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i} E(R_{ij}|Z_{ij}=1,\mathcal{F},\mathcal{Z}) - E(R_{ij}|Z_{ij}=0,\mathcal{F},\mathcal{Z})}{\sum_{i=1}^{I}\sum_{j=1}^{n_i} E(D_{ij}|Z_{ij}=1,\mathcal{F},\mathcal{Z}) - E(D_{ij}|Z_{ij}=0,\mathcal{F},\mathcal{Z})}\\
=&\frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i} r_{1ij}^{(d_{1ij})} - r_{0ij}^{(d_{0ij})}}{\sum_{i=1}^{I}\sum_{j=1}^{n_i} d_{1ij} - d_{0ij}}\\
=&\frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i}\sum_{k=1}^{M}(r_{ij}^{(k)} - r_{ij}^{(k-1)})\chi(d_{1ij} \geq k > d_{0ij})}{\sum_{i=1}^{I}\sum_{j=1}^{n_i}\sum_{k=1}^{M}\chi(d_{1ij} \geq k > d_{0ij})}
\end{aligned}$$

□

A.2.2. Proof of Theorem 4.2

We require the following two Lemmas. Lemma A.3 characterizes the moments of the test statistics in (4.5). Lemma A.4 derives the bias of $S^2(\lambda_0)$ in estimating the variance of $T(\lambda_0)$.

Lemma A.3. *The expected value and the variance of the test statistic in equation* (4.5) *are*

$$\begin{aligned}
E\{T(\lambda_0)|\mathcal{F},\mathcal{Z}\} &= \frac{1}{I}(\lambda - \lambda_0)\sum_{i=1}^{I}\sum_{j=1}^{n_i}(d_{1ij} - d_{0ij})\\
Var\{T(\lambda_0)|\mathcal{F},\mathcal{Z}\} &= \frac{1}{I^2}\sum_{i=1}^{I}\frac{1}{n_i}\sum_{i=1}^{n_i}(a_{ij,\lambda_0} - \bar{a}_{i,\lambda_0})^2
\end{aligned}$$

where

$$a_{ij,\lambda_0} = \frac{n_i}{m_i}y_{1ij,\lambda_0} + \frac{n_i}{n_i - m_i}y_{0ij,\lambda_0}, \quad \bar{a}_{i,\lambda_0} = \frac{1}{n_i}\sum_{i=1}^{n_i} a_{ij,\lambda_0}$$

Proof. Let $y_{0ij,\lambda_0} = r_{0ij}^{(d_{0ij})} - \lambda_0 d_{0ij}$ and $y_{1ij,\lambda_0} = r_{1ij}^{(d_{1ij})} - \lambda_0 d_{1ij}$. Then, $V_i(\lambda_0)$ becomes

$$\begin{aligned}
V_i(\lambda_0) &= \frac{n_i}{m_i}\sum_{j=1}^{n_i} Z_{ij}(R_{ij} - \lambda_0 D_{ij}) - \frac{n_i}{n_i - m_i}\sum_{j=1}^{n_i}(1 - Z_{ij})(R_{ij} - \lambda_0 D_{ij})\\
&= \frac{n_i}{m_i}\sum_{j=1}^{n_i} Z_{ij}y_{1ij,\lambda_0} - \frac{n_i}{n_i - m_i}\sum_{j=1}^{n_i}(1 - Z_{ij})y_{0ij,\lambda_0}
\end{aligned}$$

By assumption (A3) of IV in the main manuscript, Z_{ij} are independent within each strata.

Then, for any $i = 1, \ldots, I$ and for $j, k = 1, ..., n_i$ where $j \neq k$

$$E(Z_{ij}|\mathcal{F}, \mathcal{Z}) = \frac{m_i}{n_i}, \quad E(Z_{ij}Z_{ik}|\mathcal{F}, \mathcal{Z}) = \frac{m_i(m_i-1)}{n_i(n_i-1)} = \frac{m_i - 1}{n_i}$$

where the second equality is true because in full matching, $m_i = 1$ and $n_i = m_i - 1$ or $m_i = n_i - 1$ and $n_i = 1$. Then, the expectation of $V_i(\lambda_0)$ and the test statistic $T(\lambda_0)$ are

$$\begin{aligned}
E\{V_i(\lambda_0)|\mathcal{F}, \mathcal{Z}\} &= \sum_{j=1}^{n_i}(r_{1ij}^{(d_{1ij})} - r_{0ij}^{(d_{0ij})}) - \lambda_0(d_{1ij} - d_{0ij}) \\
E\{T(\lambda_0)|\mathcal{F}, \mathcal{Z}\} &= \frac{1}{D}\sum_{i=1}^{I} E\{V_i(\lambda_0)|\mathcal{F}, \mathcal{Z}\} = \frac{1}{D}(\lambda - \lambda_0)\sum_{i=1}^{I}\sum_{j=1}^{n_i}(d_{1ij} - d_{0ij})
\end{aligned}$$

For variance of $V_i(\lambda_0)$, Proposition 2 in Rosenbaum (2002, Sec. 2.4.4) gives us

$$\begin{aligned}
&Var\{V_i(\lambda_0)|\mathcal{F}, \mathcal{Z}\} \\
=&Var\left\{\sum_{j=1}^{n_i} Z_{ij}\left(\frac{n_i}{m_i}y_{1ij,\lambda_0} + \frac{n_i}{n_i - m_i}y_{0ij,\lambda_0}\right)|\mathcal{F}, \mathcal{Z}\right\} \\
=&\sum_{j=1}^{n_i}\left(\frac{m_i}{n_i} - \frac{m_i^2}{n_i^2}\right)a_{ij,\lambda_0}^2 + \left(\frac{m_i - 1}{n_i} - \frac{m_i^2}{n_i^2}\right)\sum_{j \neq k} a_{ij,\lambda_0}a_{ik,\lambda_0} \\
=&\left(\frac{m_i}{n_i} - \frac{m_i^2}{n_i^2} - \frac{m_i - 1}{n_i} + \frac{m_i^2}{n_i^2}\right)\sum_{j=1}^{n_i} a_{ij,\lambda_0}^2 + \left(\frac{m_i - 1}{n_i} - \frac{m_i^2}{n_i^2}\right)\sum_{j,k} a_{ij,\lambda_0}a_{ik,\lambda_0} \\
=&\frac{1}{n_i}\sum_{j=1}^{n_i} a_{ij,\lambda_0}^2 + \frac{n_i(m_i - 1) - m_i^2}{n_i^2}\sum_{j,k} a_{ij,\lambda_0}a_{ik,\lambda_0} \\
=&\frac{1}{n_i}\sum_{j=1}^{n_i} a_{ij,\lambda_0}^2 - \frac{1}{n_i^2}\sum_{j,k} a_{ij,\lambda_0}a_{ik,\lambda_0} \\
=&\frac{1}{n_i}\sum_{i=1}^{n_i}(a_{ij,\lambda_0} - \bar{a}_{i,\lambda})^2
\end{aligned}$$

Finally, the variance of $T(\lambda_0)$ is given by

$$Var\{T(\lambda_0)|\mathcal{F}, \mathcal{Z}\} = \frac{1}{D^2}\sum_{i=1}^{I} Var\{V_i(\lambda_0)|\mathcal{F}, \mathcal{Z}\} = \frac{1}{D^2}\sum_{i=1}^{I}\frac{1}{n_i}\sum_{j=1}^{n_i}(a_{ij,\lambda_0} - \bar{a}_{i,\lambda})^2$$

□

Lemma A.4. *Let $\mu_{i,\lambda_0} = E\{V_i(\lambda_0)|\mathcal{F},\mathcal{Z}\}$ and $\mu_{\lambda_0} = E\{T(\lambda_0)|\mathcal{F},\mathcal{Z}\}$. The bias of* (4.6) *in estimating the variance of the test statistic in* (4.5) *is*

$$E\{S^2(\lambda_0)|\mathcal{F},\mathcal{Z}\} - Var\{T(\lambda_0)|\mathcal{F},\mathcal{Z}\} = \frac{1}{I(I-1)}\sum_{i=1}^{I}(\mu_{i,\lambda_0} - \mu_{\lambda_0})^2 \tag{A.14}$$

Proof. Let $v^2_{i,\lambda_0} = Var\{V_i(\lambda_0)|\mathcal{F},\mathcal{Z}\}$. Under the generalized effect ratio, the bias of the estimator (4.6) is

$$\begin{aligned}
&E\{S^2(\lambda_0)|\mathcal{F},\mathcal{Z}\}\\
=&\frac{1}{I(I-1)}\sum_{i=1}^{I} E[\{V_i(\lambda_0) - T(\lambda_0)\}^2|\mathcal{F},\mathcal{Z}]\\
=&\frac{1}{I(I-1)}\sum_{i=1}^{I} E\{V_i^2(\lambda_0)|\mathcal{F},\mathcal{Z}\} + E\{T^2(\lambda_0)|\mathcal{F},\mathcal{Z}\} - 2E\{V_i(\lambda_0)T(\lambda_0)|\mathcal{F},\mathcal{Z}\}\\
=&\frac{1}{I(I-1)}\sum_{i=1}^{I}(\mu^2_{i,\lambda_0} + v_{i,\lambda_0}) + \left(\mu^2_{\lambda_0} + \frac{1}{I^2}\sum_{j=1}^{I} v_{j,\lambda_0}\right)\\
&-\frac{2}{I}\left(\mu^2_{i,\lambda_0} + v_{i,\lambda_0} + \sum_{j\neq i}\mu_{i,\lambda_0}\mu_{j,\lambda_0}\right)\\
=&\frac{1}{I(I-1)}\sum_{i=1}^{I}\left(v_{i,\lambda_0} - \frac{2}{I}v_{i,\lambda_0} + \frac{1}{I^2}\sum_{j=1}^{I} v_{j,\lambda_0}\right)\\
&+\frac{1}{I(I-1)}\sum_{i=1}^{I}\left(\mu^2_{i,\lambda_0} + \mu^2_{\lambda_0} - \frac{2}{I}\sum_{j=1}^{I}\mu_{i,\lambda_0}\mu_{j,\lambda_0}\right)\\
=&\left(\frac{I^2 - 2I + I}{I(I-1)}\right)\frac{1}{D^2}\sum_{i=1}^{n} v_{i,\lambda_0} + \frac{1}{I(I-1)}\sum_{i=1}^{I}(\mu_{i,\lambda_0} - \mu_{\lambda_0})^2\\
=&\frac{1}{I^2}\sum_{i=1}^{I} v_{i,\lambda_0} + \frac{1}{I(I-1)}\sum_{i=1}^{I}(\mu_{i,\lambda_0} - \mu_{\lambda_0})^2
\end{aligned}$$

□

Now, we can prove the Theorem as follows. We use the same notation adopted in the proof of Lemma A.4, mainly $\mu_{i,\bar{\lambda}}$, $\mu_{\bar{\lambda}}$, and $v^2_{i,\bar{\lambda}}$. In addition, let $q_{i,\bar{\lambda}} = E\{V_i^2(\bar{\lambda})|\mathcal{F},\mathcal{Z}\}$, and $v_{\bar{\lambda}} = Var\{T(\bar{\lambda})|\mathcal{F},\mathcal{Z}\}$. First, $\sum_{i=1}^I V_i^2(\bar{\lambda})/I$ is an unbiased estimator for $\sum_{i=1}^I q_{i,\bar{\lambda}}/I$. In addition,

$$Var\left\{\frac{1}{I}\sum_{i=1}^I V_i^2(\bar{\lambda})|\mathcal{F},\mathcal{Z}\right\} \leq \frac{1}{I^2}\sum_{i=1}^I E\{V_i^4(\bar{\lambda})|\mathcal{F},\mathcal{Z}\}$$

By the fourth moment condition in (4.7), we have $\sum_{i=1}^I V_i^2(\bar{\lambda})/I - \sum_{i=1}^I q_{i,\bar{\lambda}}/I \to 0$ in probability. Similarly, the same fourth moment condition in (4.7) and the same reasoning gives $T(\bar{\lambda}) - \mu_{\bar{\lambda}} \to 0$ in probability because of the growth of the variance of $T(\bar{\lambda})$ is controlled by the moment condition. Since $\mu_{\bar{\lambda}} = 0$ for all I under the null hypothesis, we have, by the continuous mapping theorem, $T^2(\bar{\lambda}) \to 0$ in probability. Combining all these convergence results, we get that for $\epsilon > 0$ and $\delta > 0$, there exists I^* such that

$$\text{for } I \geq I^*: P\left\{\frac{1}{I}\sum_{i=1}^I V_i^2(\bar{\lambda}) - \frac{1}{I}\sum_{i=1}^I q_{i,\bar{\lambda}} < -\frac{\epsilon}{2}\right\} < \frac{\delta}{2}, \quad P\left\{T^2(\bar{\lambda}) < -\frac{\epsilon}{2}\right\} < \frac{\delta}{2}$$

and

$$\begin{aligned}
&P\left\{IS^2(\bar{\lambda}) - Iv_{\bar{\lambda}} < -\epsilon\right\} \\
=&P\left[\frac{I}{I-1}\left\{\frac{1}{I}\sum_{i=1}^I V_i^2(\bar{\lambda}) - T^2(\bar{\lambda})\right\} - Iv_{\bar{\lambda}} < -\epsilon\right] \\
=&P\left[\frac{I}{I-1}\left\{\frac{1}{I}\sum_{i=1}^I V_i^2(\bar{\lambda}) - \frac{1}{I}\sum_{i=1}^I q_{i,\bar{\lambda}} + \frac{1}{I}\sum_{i=1}^I q_{i,\bar{\lambda}} - T^2(\bar{\lambda})\right\} - Iv_{\bar{\lambda}} < -\epsilon\right] \\
=&P\left[\frac{I}{I-1}\left\{\frac{1}{I}\sum_{i=1}^I V_i^2(\bar{\lambda}) - \frac{1}{I}\sum_{i=1}^I q_{i,\bar{\lambda}} - T^2(\bar{\lambda})\right\} - Iv_{\bar{\lambda}} + \frac{1}{I-1}\sum_{i=1}^I q_{i,\bar{\lambda}} < -\epsilon\right] \\
\leq&P\left[\frac{I}{I-1}\left\{\frac{1}{I}\sum_{i=1}^I V_i^2(\bar{\lambda}) - \frac{1}{I}\sum_{i=1}^I q_{i,\bar{\lambda}} - T^2(\bar{\lambda})\right\} < -\epsilon\right] \\
\leq&\frac{\delta}{2} + \frac{\delta}{2} = \delta
\end{aligned}$$

Stated in words, $IS^2(\bar{\lambda})$ will over-estimate $Iv_{\bar{\lambda}}$ with high probability.

Second, under the null hypothesis $H_0 : \lambda = \bar{\lambda}$ and from Lemma A.3, $\sum_{i=1}^I \mu_{i,\bar{\lambda}}/I = 0$.

Hence, we can rewrite the test statistic as

$$T(\bar{\lambda}) = \frac{1}{I}\sum_{i=1}^{I} V_i(\bar{\lambda}) = \frac{1}{I}\sum_{i=1}^{I}[V_i(\bar{\lambda}) - \mu_{i,\bar{\lambda}}]$$

where the test statistic becomes a sum of independent random variables $V_i(\bar{\lambda}) - \mu_{i,\bar{\lambda}}$ with mean zero and variance $v_{i,\bar{\lambda}}$.

Finally, combining the two facts, under the null $H_0 : \lambda = \bar{\lambda}$, we have

$$\frac{T(\bar{\lambda})}{S(\bar{\lambda})} = \left[\frac{\frac{1}{I}\sum_{i=1}^{I}\{V_i(\bar{\lambda}) - \mu_{i,\bar{\lambda}}\}}{\sqrt{\frac{1}{I^2}\sum_{i=1}^{I} v_{i,\bar{\lambda}}}}\right]\left\{\frac{\sqrt{\frac{1}{I^2}\sum_{i=1}^{I} v_{i,\bar{\lambda}}}}{\sqrt{S^2(\bar{\lambda})}}\right\}$$

By conditions specified in Breiman (1992, pg 186) for the central limit theorem with non-identical distributions, the first parenthesis term converges to the standard Normal distribution. From our result about $IS^2(\bar{\lambda})$ overestimating $Iv_{\bar{\lambda}}$, the second parenthesis term will be smaller than 1 with high probability. Hence, taking the sup of the entire expression, we obtain

$$\limsup_{I\to\infty} P\left\{\frac{T(\bar{\lambda})}{S(\bar{\lambda})} \leq -t|\mathcal{F}, \mathcal{Z}\right\} \leq \Phi(-t), \quad \limsup_{I\to\infty} P\left\{\frac{T(\bar{\lambda})}{S(\bar{\lambda})} \geq t|\mathcal{F}, \mathcal{Z}\right\} \leq \Phi(-t)$$

where $\Phi()$ is the standard normal distribution. □

A.2.3. Proof of Corollary 4.1

First, we see that $T(\lambda)/S(\lambda) = q$ implies $T^2(\lambda) = q^2 S^2(\lambda)$. This expression can be rewritten as

$$T^2(\lambda) = \frac{q^2}{I(I-1)}\sum_{i=1}^{I}(V_i(\lambda) - T(\lambda))^2 = \frac{q^2}{I(I-1)}\left\{\sum_{i=1}^{I} V_i^2(\lambda) - IT^2(\lambda)\right\} \tag{A.15}$$

Rearranging the terms in (A.15), we get

$$T^2(\lambda)\left(1+\frac{q^2}{I-1}\right)=\frac{q^2}{I(I-1)}\sum_{i=1}^{I}V_i^2(\lambda)$$

Second, we can re-express $V_i(\lambda)$ as follows.

$$\begin{aligned}
V_i(\lambda) &= \sum_{j=1}^{n_i}\left(\frac{n_i}{m_i}+\frac{n_i}{n_i-m_i}\right)Z_{ij}R_{ij}-\sum_{j=1}^{n_i}\frac{n_i}{n_i-m_i}R_{ij}\\
&\quad-\sum_{j=1}^{n_i}\left(\frac{n_i}{m_i}+\frac{n_i}{n_i-m_i}\right)\lambda Z_{ij}D_{ij}+\sum_{j=1}^{n_i}\frac{n_i}{n_i-m_i}\lambda D_{ij}\\
&=\sum_{j=1}^{n_i}\frac{n_i^2}{m_i(n_i-m_i)}Z_{ij}R_{ij}-\left(\sum_{j=1}^{n_i}\frac{n_i}{n_i-m_i}R_{ij}\right)\left(\frac{1}{m_i}\sum_{j=1}^{n_i}Z_{ij}\right)\\
&\quad-\sum_{j=1}^{n_i}\frac{n_i^2}{m_i(n_i-m_i)}\lambda Z_{ij}D_{ij}+\left(\sum_{j=1}^{n_i}\frac{n_i}{n_i-m_i}\lambda D_{ij}\right)\left(\frac{1}{m_i}\sum_{j=1}^{n_i}Z_{ij}\right)\\
&=\frac{n_i^2}{m_i(n_i-m_i)}\left(\sum_{j=1}^{n_i}Z_{ij}R_{ij}-\frac{1}{n_i}\sum_{j=1}^{n_i}R_{ij}\sum_{j=1}^{n_i}Z_{ij}\right)\\
&\quad-\lambda\frac{n_i^2}{m_i(n_i-m_i)}\left(\sum_{j=1}^{n_i}Z_{ij}D_{ij}-\frac{1}{n_i}\sum_{j=1}^{n_i}D_{ij}\sum_{j=1}^{n_i}Z_{ij}\right)\\
&=\frac{n_i^2}{m_i(n_i-m_i)}\sum_{j=1}^{n_i}(Z_{ij}-\bar{Z}_{i.})(R_{ij}-\bar{R}_{i.})-\lambda\frac{n_i^2}{m_i(n_i-m_i)}\sum_{j=1}^{n_i}(Z_{ij}-\bar{Z}_{i.})(D_{ij}-\bar{D}_{i.})
\end{aligned}$$

Immediately, we also have $V_i(\lambda) = G_i - \lambda H_i$. Then, we can rewrite $\sum_{i=1}^{I} V_i^2(\lambda)$ and $T^2(\lambda)$ as follows

$$\begin{aligned}
\sum_{i=1}^{I} V_i^2(\lambda) &= \sum_{i=1}^{I} (G_i - \lambda H_i)^2 \\
&= \sum_{i=1}^{I} G_i^2 - 2\lambda \sum_{i=1}^{I} G_i H_i + \lambda^2 \sum_{i=1}^{I} H_i^2 \\
T^2(\lambda) &= \frac{1}{I^2} \left\{ \sum_{i=1}^{I} V_i(\lambda) \right\}^2 \\
&= \frac{1}{I^2} \left\{ \sum_{i=1}^{I} (G_i - \lambda H_i) \right\}^2 \\
&= \frac{1}{I^2} \left\{ \left(\sum_{i=1}^{I} G_i \right)^2 - 2\lambda \sum_{i=1}^{I} G_i \sum_{i=1}^{I} H_i + \lambda^2 \left(\sum_{i=1}^{I} H_i \right)^2 \right\}
\end{aligned}$$

Overall, we can rewrite the equation (A.15) as

$$\begin{aligned}
&\frac{1}{I^2} \left\{ \left(\sum_{i=1}^{I} G_i \right)^2 - 2\lambda \sum_{i=1}^{I} G_i \sum_{i=1}^{I} H_i + \lambda^2 \left(\sum_{i=1}^{I} H_i \right)^2 \right\} \left(1 + \frac{q^2}{I-1} \right) \\
=&\frac{q^2}{I(I-1)} \left(\sum_{i=1}^{I} G_i^2 - 2\lambda \sum_{i=1}^{I} G_i H_i + \lambda^2 \sum_{i=1}^{I} H_i^2 \right)
\end{aligned}$$

Finally, we pull out the coefficients associated with λ^2 and λ, denoted as A_2 and A_1, respectively. The remaining term are constants and we denote them as A_0. All A_2, A_1, and

A_0 are explicitly written below.

$$\begin{aligned}
A_2 &= \frac{1}{I^2}\left(\sum_{i=1}^{I} H_i\right)^2 + \frac{q^2}{I(I-1)}\left\{\frac{1}{I}\left(\sum_{i=1}^{I} H_i\right)^2 - \sum_{i=1}^{I} H_i^2\right\} \\
&= \bar{H}_.^2 - \frac{q^2}{I(I-1)}\sum_{i=1}^{I}(H_i - \bar{H}_.)^2 \\
A_1 &= -2\left[\frac{1}{I^2}\sum_{i=1}^{I} G_i \sum_{i=1}^{I} H_i + \frac{q^2}{I(I-1)}\left\{\frac{1}{I}\sum_{i=1}^{I} G_i \sum_{i=1}^{I} H_i - \sum_{i=1}^{I} G_i H_i\right\}\right] \\
&= -2\left[\bar{G}_. \bar{H}_. - \frac{q^2}{I(I-1)}\left\{\sum_{i=1}^{I}(G_i - \bar{G}_.)(H_i - \bar{H}_.)\right\}\right] \\
A_0 &= \frac{1}{I^2}\left(\sum_{i=1}^{I} G_i\right)^2 + \frac{q^2}{I(I-1)}\left\{\frac{1}{I}\left(\sum_{i=1}^{I} G_i\right)^2 - \sum_{i=1}^{I} G_i^2\right\} \\
&= \bar{G}_.^2 - \frac{q^2}{I(I-1)}\sum_{i=1}^{I}(G_i - \bar{G}_.)^2
\end{aligned}$$

□

If $q = 0$ in Corollary 4.1, there is only one solution to the quadratic equation since

$$A_2\lambda^2 + A_1\lambda + A_0 = \bar{H}_.^2\lambda^2 - 2\bar{H}_.\bar{G}_.\lambda + \bar{G}_.^2 = (\bar{H}_.\lambda - \bar{G}_.)^2 = 0$$

This gives us an explicit formula for the estimator of the effect ratio, denoted as $\hat{\lambda}$.

$$\hat{\lambda} = \frac{\bar{G}.}{\bar{H}.} = \frac{\sum_{i=1}^{I} \frac{n_i^2}{m_i(n_i - m_i)} \sum_{j=1}^{n_i}(R_{ij} - \bar{R}_{i.})(Z_{ij} - \bar{Z}_{i.})}{\sum_{i=1}^{I} \frac{n_i^2}{m_i(n_i - m_i)} \sum_{j=1}^{n_i}(D_{ij} - \bar{D}_{i.})(Z_{ij} - \bar{Z}_{i.})} \tag{A.16}$$

□

A.2.4. Proof of Theorem 4.3

First, for all $i = 1, \ldots, I$ and $j = 1, \ldots, n_i$, we have

$$Z_{ij} - \bar{Z}_{i.} = \begin{cases} 1 - \frac{m_i}{n_i} & \text{if } Z_{ij} = 1 \\ -\frac{m_i}{n_i} & \text{if } Z_{ij} = 0 \end{cases}$$

Furthermore,

$$\sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_{i.}) = 0, \quad \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_{i.})^2 = \frac{m_i(n_i - m_i)}{n_i}$$

Second, for fixed Z_{ij}, we have the following expected values for J_i

$$\begin{aligned} E(J_i) =& 0 \\ E(J_i^2) =& Var\left\{\sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_{i.})(\epsilon_{ij} - \bar{\epsilon}_{i.})\right\} \\ =& \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_{i.})^2 Var(\epsilon_{ij} - \bar{\epsilon}_{i.}) + \sum_{j,k} (Z_{ij} - \bar{Z}_{i.})(Z_{ik} - \bar{Z}_{k.}) Cov(\epsilon_{ij} - \bar{\epsilon}_{i.}, \epsilon_{ik} - \bar{\epsilon}_{i.}) \\ =& (1 - \frac{1}{n_i})\sigma_{i,R}^2 \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_{i.})^2 - \frac{1}{n_i}\sigma_{i,R}^2 \sum_{j,k} (Z_{ij} - \bar{Z}_{i.})(Z_{ik} - \bar{Z}_{k.}) \\ =& \sigma_{i,R}^2 \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_{i.})^2 - \frac{1}{n_i}\sigma_{i,R}^2 \left\{\sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_{i.})\right\}^2 \\ =& \sigma_{i,R}^2 \frac{m_i(n_i - m_i)}{n_i} \end{aligned}$$

For the third moment, for each i, let $k_1, \ldots, k_{n_i}$ be non-negative integers and define the multinomial coefficient as follows.

$$\binom{3}{k_1, \ldots, k_{n_i}} = \frac{3!}{k_1! \cdots k_{n_i}!}$$

Then, we have

$$
\begin{aligned}
E(|J_i^3|) =& E| \left\{ \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_{i.})(\epsilon_{ij} - \bar{\epsilon}_{i.}) \right\}^3 | \\
=& E| \sum_{k_1 + \cdots + k_{n_i} = 3} \binom{3}{k_1, \ldots, k_{n_i}} \prod_{j=1}^{n_i} \left\{ (Z_{ij} - \bar{Z}_{i.})(\epsilon_{ij} - \bar{\epsilon}_{i.}) \right\}^{k_j} | \\
\leq & \sum_{k_1 + \cdots + k_{n_i} = 3} \binom{3}{k_1, \ldots, k_{n_i}} \prod_{j=1}^{n_i} |Z_{ij} - \bar{Z}_{i.}|^{k_j} E|\epsilon_{ij} - \bar{\epsilon}_{i.}|^{k_j} < \infty
\end{aligned}
$$

because third moments exist and are bounded for all ϵ_{ij} and n_i is bounded. Third, based on these moment calculations, it immediately follows that

$$
E\left[\sum_{i=1}^{I} \left\{ \frac{n_i^2}{(m_i)(n_i - m_i)} J_i \right\}^2 \right] = \sum_{i=1}^{I} \left\{ \frac{n_i^4}{(m_i)^2 (n_i - m_i)^2} \right\} \left\{ \frac{m_i(n_i - m_i)}{n_i} \sigma_{i,R}^2 \right\} = s_I^2
$$

Then, by Theorem 9.2 in Chapter 9, Section 3 of Breiman (1992) (pg 187), the sum of J_i weighted by $n_i^2/m_i(n_i - m_i)$ is a standard Normal distribution

$$
\frac{\sum_{i=1}^{I} \frac{n_i^2}{m_i(n_i - m_i)} J_i}{s_I} \rightarrow N(0,1)
$$

Fourth, for H_i, we have the following moments

$$
\begin{aligned}
E(H_i) =& \gamma m_i (1 - \frac{m_i}{n_i}) \\
Var(H_i) =& Var\left(\sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_{i.})(D_{ij} - \bar{D}_{i.}) \right) \\
=& (1 - \frac{1}{n_i}) \sigma_{i,D}^2 \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_{i.})^2 - \frac{1}{n_i} \sigma_{i,D}^2 \sum_{j,k} (Z_{ij} - \bar{Z}_{i.})(Z_{ik} - \bar{Z}_{k.}) \\
=& \sigma_{i,D}^2 \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_{i.})^2 - \frac{1}{n_i} \sigma_{i,D}^2 \left(\sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_{i.}) \right)^2 \\
=& \sigma_{i,D}^2 \frac{m_i(n_i - m_i)}{n_i}
\end{aligned}
$$

Fifth, by Theorem C in page 27 of Serfling (1980),

$$\begin{aligned}
&\frac{1}{I}\sum_{i=1}^{I}\frac{n_i^2}{m_i(n_i-m_i)}H_i-\gamma\frac{1}{I}\sum_{i=1}^{I}E\left\{\frac{n_i^2}{m_i(n_i-m_i)}H_i\right\}\\
=&\frac{1}{I}\sum_{i=1}^{I}\frac{n_i^2}{m_i(n_i-m_i)}H_i-\gamma\frac{1}{I}\sum_{i=1}^{I}n_i\to 0
\end{aligned}$$

Finally, combining all these facts together, we can rewrite the effect ratio estimator as follows.

$$\begin{aligned}
\hat{\beta}&=\frac{\sum_{i=1}^{I}\frac{n_i^2}{m_i(n_i-m_i)}\sum_{j=1}^{n_i}(Z_{ij}-\bar{Z}_{i.})(R_{ij}-\bar{R}_{i.})}{\sum_{i=1}^{I}\frac{n_i^2}{m_i(n_i-m_i)}\sum_{j=1}^{n_i}(Z_{ij}-\bar{Z}_{i.})(D_{ij}-\bar{D}_{i.})}\\
&=\beta+\frac{\sum_{i=1}^{I}\frac{n_i^2}{m_i(n_i-m_i)}\sum_{j=1}^{n_i}(Z_{ij}-\bar{Z}_{i.})(\epsilon_{ij}-\bar{\epsilon}_{i.})}{\sum_{i=1}^{I}\frac{n_i^2}{m_i(n_i-m_i)}H_i}\\
&=\beta+\frac{\sum_{i=1}^{I}\frac{n_i^2}{m_i(n_i-m_i)}J_i}{\sum_{i=1}^{I}\frac{n_i^2}{m_i(n_i-m_i)}H_i}
\end{aligned}$$

which leads to

$$\sqrt{I}(\hat{\beta}-\beta)=\left\{\frac{\sum_{i=1}^{I}\frac{n_i^2}{m_i(n_i-m_i)}J_i}{s_I}\right\}\left\{\frac{\frac{1}{\sqrt{I}}s_I}{\frac{1}{I}\sum_{i=1}^{I}\frac{n_i^2}{m_i(n_i-m_i)}H_i}\right\}$$

Finally, using Slutsky's Theorem, $\sqrt{I}(\hat{\beta}-\beta)$ converges to a Normal distribution with mean 0 and stated asymptotic variance. □

BIBLIOGRAPHY

A. Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231 – 263, 2003.

M. Aidoo, D. J. Terlouw, M. S. Kolczak, P. D. McElroy, F. O. ter Kuile, S. Kariuki, B. L. Nahlen, A. A. Lal, and V. Udhayakumar. Protective effects of the sickle cell gene against malaria morbidity and mortality. *The Lancet*, 359(9314):1311–1312, 2002.

A. C. Allison. Polymorphism and natural selection in human populations. *Cold Spring Harbor Symposia on Quantitative Biology*, 29:137–149, 1964.

E. Anderson. True fact: The lack of pirates is causing global warming. `http://www.forbes.com/sites/erikaandersen/2012/03/23/true-fact-the-lack-of-pirates-is-causing-global-warming/`, 2012. Accessed: 2015-03-31.

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

D. W. K. Andrews. Consistent moment selection procedures for generalized method of moments estimation. *Econometrica*, 67(3):543–563, 1999.

D. W. K. Andrews, M. J. Moreira, and J. H. Stock. Optimal two-sided invariant similar tests for instrumental variables regression. *Econometrica*, 74(3):715–752, 2006.

J. D. Angrist and G. W. Imbens. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430):431–442, 1995.

J. D. Angrist and A. B. Krueger. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014, 1991.

J. D. Angrist and A. B. Krueger. Instrumental variables and the search for identification: From supply and demand to natural experiments. *The Journal of Economic Perspectives*, 15(4):69–85, 2001.

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

E. Arinaitwe, A. Gasasira, W. Verret, J. Homsy, H. Wanzira, A. Kakuru, and T. G. Sandison. The association between malnutrition and the incidence of malaria among young hiv-infected and-uninfected ugandan children: a prospective study. *Malaria Journal*, 11: 90, 2012.

M. T. Ashcroft, P. Desai, and S. A. Richardson. Growth, behaviour, and educational achievement of jamaican children with sickle-cell trait. *British Medical Journal*, 1(6022): 1371–1373, 1976.

M. T. Ashcroft, P. Desai, G. A. Grell, B. E. Serjeant, and G. R. Serjeant. Heights and weights of west indian children with the sickle cell trait. *Archives of Disease in Childhood*, 53(7):596–598, 1978.

M. Baiocchi, D. S. Small, S. Lorch, and P. R. Rosenbaum. Building a stronger instrument in an observational study of perinatal care for premature infants. *Journal of the American Statistical Association*, 105(492):1285–1296, 2010.

M. Baiocchi, J. Cheng, and D. S. Small. Instrumental variable methods for causal inference. *Statistics in Medicine*, 33(13):2297–2340, 2014.

R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

S. Beckers, A. V. Peeters, F. De Freitas, I. L. Mertens, S. L. Verhulst, D. Haentjens, K. N. Desager, L. F. Van Gaal, and W. Van Hul. Association study and mutation analysis of adiponectin shows association of variants in apm1 with complex obesity in women. *Annals of Human Genetics*, 73(5):492–501, 2009.

A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012.

A. Björklund and R. Moffitt. The estimation of wage gains and welfare gains in self-selection models. *The Review of Economics and Statistics*, pages 42–49, 1987.

K. A. Bollen. Instrumental variables in sociology and the social sciences. *Annual Review of Sociology*, 38:37–72, 2012.

N. Bouatia-Naji, D. Meyre, S. Lobbens, K. Séron, F. Fumeron, B. Balkau, B. Heude, B. Jouret, P. E. Scherer, C. Dina, J. Weill, and P. Froguel. Acdc/adiponectin polymorphisms are associated with severe childhood and adult obesity. *Diabetes*, 55(2):545–550, 2006.

H. E. Bouis and L. J. Haddad. *Agricultural Commercialization, Nutrition, and the Rural Poor*. Lynne Rienner Publishers, 1990.

H. E. Bouis and L. J. Haddad. Are estimates of calorie-income elasticities too high?: A recalibration of the plausible range. *Journal of Development Economics*, 39(2):333–364, 1992.

J. Bound, D. A. Jaeger, and R. M. Baker. Problems with instrumental variables estimation when the correlation between instruments and the endogenous variable is weak. *Journal of the American Statistical Association*, 90:443–450, 1995.

A. M. Bradley-Moore, B. M. Greenwood, A. K. Bradley, B. R. Kirkwood, and H. M. Gilles. Malaria chemoprophylaxis with chloroquine in young nigerian children. iii. its effect on nutrition. *Annals of Tropical Medicine and Parasitology*, 79(6):575–584, 1985.

L. Breiman. *Probability.* SIAM: Society for Industrial and Applied Mathematics, 1992.

P. Brennan. Commentary: Mendelian randomization and gene–environment interaction. *International Journal of Epidemiology*, 33(1):17–21, 2004.

M. A. Brookhart and S. Schneeweiss. Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. *The International Journal of Biostatistics*, 3(1):14, 2007.

P. Bühlmann and S. van der Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer, 2011.

T. T. Cai and A. Zhang. Compressed sensing and affine rank minimization under restricted isometry. *IEEE Transactions on Signal Processing*, 61(13):3279–3290, 2013a.

T. T. Cai and A. Zhang. Sharp rip bound for sparse signal and low-rank matrix recovery. *Applied and Computational Harmonic Analysis*, 35(1):74–93, 2013b.

T. T. Cai, J. Fan, and T. Jiang. Distributions of angles in random packing on spheres. *The Journal of Machine Learning Research*, 14(1):1837–1864, 2013.

E. J. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

D. Card. Using geographic variations in college proximity to estimate the return to schooling. In L. N. Christofides, E. K. Grant, and R. Swidinsky, editors, *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp*. University of Toronto Press, 1995.

J. Cawley and C. Meyerhoefer. The medical care costs of obesity: an instrumental variables approach. *Journal of Health Economics*, 31(1):219–230, 2012.

J. Cheng. Using the instrumental propensity score in observational studies for causal effects. *Joint Statistical Meeting Presentation*, 2011.

J. Cheng, J. Qin, and B. Zhang. Semiparametric estimation and inference for distributional and general treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(4):881–904, 2009.

R. Cholera, N. J. Brittain, M. R. Gillrie, T. M. Lopera-Mesa, S. A. S. Diakit, T. Arie, M. A. Krause, A. Guindo, A. Tubman, H. Fujioka, D. A. Diallo, O. K. Doumbo, M. Ho,

T. E. Wellems, and R. M. Fairhurst. Impaired cytoadherence of plasmodium falciparum-infected erythrocytes containing sickle hemoglobin. *Proceedings of the National Academy of Sciences*, 105(3):991–996, 2008.

W. G. Cochran. The planning of observational studies of human populations. *Journal of the Royal Statistical Society: Series A (General)*, pages 234–266, 1965.

B. T. Crookston, S. C. Alder, I. Boakye, R. M. Merrill, J. H. Amuasi, C. A. Porucznik, J. B. Stanford, T. T. Dickerson, K. A. Dearden, D. C. Hale, J. Sylverken, B. S. Snow, A. Osei-Akoto, and D. Ansong. Exploring the relationship between chronic undernutrition and asymptomatic malaria in ghanaian children. *Malaria Journal*, 9(39), 2010.

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

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

R. Davidson and J. G. MacKinnon. *Estimation and Inference in Econometrics.* Oxford University Press, New York, 1993.

J. L. Deen, G. E. L. Walraven, and L. von Seidlein. Increased risk for malaria in chronically malnourished children under 5 years of age in rural gambia. *Journal of Tropical Pediatrics*, 48(2):78–83, 2002.

A. Deribew, F. Alemseged, F. Tessema, L. Sena, Z. Birhanu, A. Zeynudin, M. Sudhakar, N. Abdo, K. Deribe, and S. Biadgilign. Malaria and under-nutrition: A community based study among under-five children at risk of malaria, south-west ethiopia. *PLoS One*, 5(5): e10775, 2010.

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

C. Dina, D. Meyre, S. Gallina, E. Durand, A. Krner, P. Jacobson, L. M. S. Carlsson, W. Kiess, V. Vatin, C. Lecoeur, J. Delplanque, E. Vaillant, F. Pattou, J. Ruiz, J. Weill, C. Levy-Marchal, F. Horber, N. Potoczna, S. Hercberg, C. Le Stunff, P. Bougnéres, P. Kovacs, M. Marre, B. Balkau, S. Cauchi, J.-C. Chévre, and P. Froguel. Variation in FTO contributes to childhood obesity and severe adult obesity. *Nature Genetics*, 39(6): 724–726, 2007.

D. L. Donoho. For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006.

J.-M. Dufour. Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica*, pages 1365–1387, 1997.

J.-M. Dufour. Identification, weak instruments, and statistical inference in econometrics. *The Canadian Journal of Economics / Revue canadienne d'Economique*, 36(4):767–808, 2003.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

S. Ehrhardt, G. D. Burchard, C. Mantel, J. P. Cramer, S. Kaiser, M. Kubo, R. N. Otchwemah, U. Bienzle, and F. P. Mockenhaupt. Malaria, anemia, and malnutrition in african childrendefining intervention priorities. *Journal of Infectious Diseases*, 194(1):108–114, 2006.

F. Fillol, J. B. Sarr, D. Boulanger, B. Cisse, C. Sokhna, G. Riveau, K. B. Simondon, and F. Remoué. Impact of child malnutrition on the specific anti-plasmodium falciparum antibody response. *Malaria Journal*, 8(1):116, 2009.

M. J. Friedman. Erythrocytic mechanism of sickle cell resistance to malaria. *Proceedings of the National Academy of Sciences*, 75(4):1994–1997, 1978.

M. Frölich. Nonparametric iv estimation of local average treatment effects with covariates. *Journal of Econometrics*, 139(1):35–75, 2007.

M. Frölich and B. Melly. Estimation of quantile treatment effects with stata. *Stata Journal*, 10(3):423–457, 2010.

J. L. Gastwirth, A. M. Krieger, and P. R. Rosenbaum. Asymptotic separability in sensitivity analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62 (3):545–555, 2000.

E. Gautier and A. B. Tsybakov. High-dimensional instrumental variables regression and confidence sets. *arXiv preprint arXiv:1105.2454*, 2011.

L. A. Gennetian, K. Magnuson, and P. A. Morris. From statistical associations to causation: What developmentalists can learn from instrumental variables techniques coupled with experimental data. *Developmental Psychology*, 44(2):381, 2008.

B. Genton, F. Al-Yaman, M. Ginny, J. Taraika, and M. P. Alpers. Relation of anthropometry to malaria morbidity and immunity in papua new guinean children. *The American Journal of Clinical Nutrition*, 68(3):734–41, 1998.

A. Ghansah, K. A. Rockett, T. G. Clark, M. D. Wilson, K. A. Koram, A. R. Oduro, L. Amenga-Etego, T. Anyorigiya, A. Hodgson, P. Milligan, W. O. Rogers, and D. P. Kwiatkowski. Haplotype analyses of haemoglobin c and haemoglobin s and the dynamics of the evolutionary response to malaria in kassena-nankana district of ghana. *PLoS ONE*, 7(4):e34565, 04 2012.

M. M. Glymour, E. J. Tchetgen Tchetgen, and J. M. Robins. Credible mendelian randomization studies: approaches for evaluating the instrumental variable assumptions. *American Journal of Epidemiology*, 175(4):332–339, 2012.

B. B. Hansen. Full matching in an observational study of coaching for the sat. *Journal of the American Statistical Association*, 99(467):609–618, 2004.

B. B. Hansen and S. O. Klopfer. Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, 15(3):609–627, 2006.

L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, pages 1029–1054, 1982.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, volume 2. Springer, 2009.

R. M. Hauser. Survey response in the long run: The wisconsin longitudinal study. *Field Methods*, 17(1):3–29, 2005.

A. Haviland, D. S. Nagin, and P. R. Rosenbaum. Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods*, 12 (3):247, 2007.

J. J. Heckman and R. Robb Jr. Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics*, 30(1):239–267, 1985.

M. A. Hernán and J. M. Robins. Instruments for causal inference: An epidemiologist's dream? *Epidemiology*, 17(4):360–372, 2006.

A. V. S. Hill, C. E. M. Allsopp, D. Kwiatkowski, N. M. Anstey, P. Twumasi, P. A. Rowe, S. Bennett, D. Brewster, A. J. McMichael, and B. M. Greenwood. Common west african hla antigens are associated with protection from severe malaria. *Nature*, 352(6336):595–600, 1991.

J. L. Hodges and E. L. Lehmann. Estimation of location based on ranks. *Annals of Mathematical Statistics*, 34(2):598–611, 1963.

P. W. Holland. Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology*, 18(1):449–484, 1988.

J.-P. Hwang, S.-J. Tsai, C.-J. Hong, C.-H. Yang, J.-F. Lirng, and Y.-M. Yang. The val66met polymorphism of the brain-derived neurotrophic-factor gene is associated with geriatric depression. *Neurobiology of Aging*, 27(12):1834–1837, 2006.

G. W. Imbens and D. B. Rubin. Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, 25(1):305–327, 1997.

Institute of Medicine of the National Academies: Immunization Safety Review Committee. *Immunization Safety Review: Vaccines and Autism.* National Academies Press, 2004.

J. B. Kadane and T. W. Anderson. A comment on the test of overidentifying restrictions. *Econometrica*, 45(4):1027–1031, 1977.

H. Kang, B. Kreuels, O. Adjei, R. Krumkamp, J. May, and D. S. Small. The causal effect of malaria on stunting: a mendelian randomization and matching approach. *International Journal of Epidemiology*, 42(5):1390–1398, 2013.

H. Kang, A. Zhang, T. T. Cai, and D. S. Small. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, 2015.

M. B. Katan. Apoupoprotein e isoforms, serum cholesterol, and cancer. *The Lancet*, 327 (8479):507–508, 1986.

L. J. Keele and J. Morgan. Stronger instruments by design. *Working Paper*, 2013.

F. Kleibergen. Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica*, 70(5):1781–1803, 2002.

F. Kleibergen. Generalizing weak instrument robust iv statistics towards multiple parameters, unrestricted covariance matrices and identification statistics. *Journal of Econometrics*, 139(1):181–216, 2007.

R. Kobbe, C. Kreuzberg, S. Adjei, B. Thompson, I. Langefeld, P. A. Thompson, H. H. Abruquah, B. Kreuels, M. Ayim, W. Busch, F. Marks, K. Amoah, E. Opoku, C. G. Meyer, O. Adjei, and J. May. A randomized controlled trial of extended intermittent preventive antimalarial treatment in infants. *Clinical Infectious Diseases*, 45(1):16–25, 2007.

M. Kolesár, R. Chetty, J. N. Friedman, E. L. Glaeser, and G. W. Imbens. Identification and inference with many invalid instruments. *National Bureau of Economic Research*, page No. w17519, 2013.

T. C. Koopmans, H. Rubin, and R. B. Leipnik. Measuring the equation systems of dynamic economics. In *Statistical Inference in Dynamic Economic Models*. John Wiley & Sons, Inc., 1950.

E. L. Korenromp, J. R. M. Armstrong-Schellenberg, B. G. Williams, B. L. Nahlen, and R. W. Snow. Impact of malaria control on childhood anaemia in africa–a quantitative review. *Tropical Medicine & International Health*, 9(10):1050–1065, 2004.

M. S. Kramer, Y. Rooks, and H. A. Pearson. Growth and development in children with sickle-cell trait: a prospective study of matched pairs. *New England Journal of Medicine*, 299(13):686–689, 1978.

B. Kreuels, S. Ehrhardt, C. Kreuzberg, S. Adjei, R. Kobbe, G. Burchard, C. Ehmen, M. Ayim, O. Adjei, and J. May. Sickle cell trait (hbas) and stunting in children below two years of age in an area of high malaria transmission. *Malaria Journal*, 8(1):16, 2009.

B. Kreuels, C. Kreuzberg, R. Kobbe, M. Ayim-Akonor, P. Apiah-Thompson, B. Thompson, C. Ehmen, S. Adjei, I. Langefeld, O. Adjei, and J. May. Differing effects of hbs and hbc traits on uncomplicated falciparum malaria, anemia, and child growth. *Blood*, 115(22): 4551–4558, 2010.

D. A. Lawlor, R. M. Harbord, J. A. C. Sterne, N. Timpson, and G. Davey Smith. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27(8):1133–1163, 2008.

J. Little and M. J. Khoury. Mendelian randomisation: a new spin or real progress? *The Lancet*, 362(9388):930–931, 2003.

A. Martínez-Calleja, I. Quiróz-Vargas, I. Parra-Rojas, J. F. Muñoz-Valle, M. A. Leyva-Vázquez, G. Fernández-Tilapa, A. Vences-Velázquez, M. Cruz, E. Salazar-Martínez, and E. Flores-Alfaro. 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, 2012.

J. May, J. A. Evans, C. Timmann, C. Ehmen, W. Busch, T. Thye, T. Agbenyega, and R. D. Horstmann. Hemoglobin variants and disease manifestations in severe falciparum malaria. *Journal of the American Medical Association*, 297(20):2220–2226, 2007.

I. A. McGregor, H. M. Gilles, J. H. Walters, A. H. Davies, and F. A. Pearson. Effects of heavy and repeated malarial infections on gambian infants and children: Effects of erythrocytic parasitization. *The British Medical Journal*, 2(4994):686–692, 1956.

F. Mealli and B. Pacini. Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. *Journal of the American Statistical Association*, 108 (503):1120–1131, 2013.

A. Mikusheva. Robust confidence sets in the presence of weak instruments. *Journal of Econometrics*, 157(2):236–247, 2010.

M. J. Moreira. A conditional likelihood ratio test for structural models. *Econometrica*, 71 (4):1027–1048, 2003.

M. P. Murray. Avoiding invalid instruments and coping with weak instruments. *The Journal of Economic Perspectives*, 20(4):111–132, 2006.

B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

National Institute of Health. Clinical guidelines on the identification, evaluation, and treatment of overweight and obesity in adults: The evidence report. *Obesity Research*, 2: 51S–209S, 1998.

J. Neyman. On the application of probability theory to agricultural experiments. *Statistical Science*, 5:463–480, 1923.

S.-L. T. Normand, M. B. Landrum, E. Guadagnoli, J. Z. Ayanian, T. J. Ryan, P. D. Cleary, and B. J. McNeil. Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. *Journal of Clinical Epidemiology*, 54(4):387–398, 2001.

A. M. Nyakeriga, M. Troye-Blomberg, A. K. Chemtai, K. Marsh, and T. N. Williams. Malaria and nutritional status in children living on the coast of kenya. *Scandinavian Journal of Immunology*, 59(6):615–616, 2004.

E. L. Ogburn, A. Rotnitzky, and J. M. Robins. Doubly robust estimation of the local average treatment effect curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):373–396, 2015.

S. B. Omer, D. A. Salmon, W. A. Orenstein, M. P. deHart, and N. Halsey. Vaccine refusal, mandatory immunization, and the risks of vaccine-preventable diseases. *New England Journal of Medicine*, 360(19):1981–1988, 2009.

R. A. Price, W.-D. Li, and H. Zhao. Fto gene snps associated with extreme obesity in cases, controls and extremely discordant sister pairs. *BMC Medical Genetics*, 9(1):4, 2008.

N. Rehan. Growth status of children with and without sickle cell trait. *Clinical Pediatrics*, 20(11):705–709, 1981.

A. L. Rice, L. Sacco, A. Hyder, and R. E. Black. Malnutrition as an underlying cause of childhood deaths associated with infectious diseases in developing countries. *Bulletin of the World Health Organization*, 78(10):1207–1221, 2000. ISSN 0042-9686.

D. H. Richardson and D.-M. Wu. A note on the comparison of ordinary and two-stage least squares estimators. *Econometrica*, pages 973–981, 1971.

N. S. Roetker, J. A. Yonker, C. Lee, V. Chang, J. J. Basson, C. L. Roan, T. S. Hauser, R. M. Hauser, and C. S. Atwood. Multigene interactions and the prediction of depression in the wisconsin longitudinal study. *British Medical Journal Open*, 2(4):e000944, 2012.

P. R. Rosenbaum. A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society. Series B. Methodological*, 53(3):597–610, 1991.

P. R. Rosenbaum. *Observational Studies.* Springer Series in Statistics. Springer-Verlag, New York, second edition, 2002.

P. R. Rosenbaum. *Design of Observational Studies.* Springer Series in Statistics. Springer, New York, 2010.

P. R. Rosenbaum and J. H. Silber. Amplification of sensitivity analysis in matched observational studies. *Journal of the American Statistical Association*, 104(488):1398–1405, 2009.

T. J. Rothenberg. Identification in parametric models. *Econometrica*, pages 577–591, 1971.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

D. B. Rubin. Comment on "randomized analysis of experimental data: The fisher randomization test". *Journal of the American Statistical Association*, 75(371):591–593, 1980.

D. B. Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

D. B. Rubin. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, 26(1):20–36, 2007.

D. B. Rubin and R. P. Waterman. Estimating the causal effects of marketing interventions using propensity score methodology. *Statistical Science*, pages 206–222, 2006.

J. K. Rybakowski, A. Borkowska, M. Skibinska, A. Szczepankiewicz, P. Kapelski, A. Leszczynska-rodziewicz, P. M. Czerski, and J. Hauser. Prefrontal cognition in schizophrenia and bipolar illness in relation to val66met polymorphism of the brain-derived neurotrophic factor gene. *Psychiatry and Clinical Neurosciences*, 60(1):70–76, 2006.

T. H. Sach, G. R. Barton, M. Doherty, K. R. Muir, C. Jenkinson, and A. J. Avery. 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(1):189–196, 2006.

J. D. Sargan. The estimation of economic relationships using instrumental variables. *Econometrica*, pages 393–415, 1958.

J. A. G. Scott, J. A. Berkley, I. Mwangi, L. Ochola, S. Uyoga, A. Macharia, C. Ndila, B. S. Lowe, S. Mwarumba, E. Bauni, K. Marsh, and T. N. Williams. Relation between falciparum malaria and bacteraemia in kenyan children: a population-based, case-control study and a longitudinal study. *The Lancet*, 378(9799):1316–1323, 2011.

R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, 1980.

N. A. Sheehan, V. Didelez, P. R. Burton, and M. D. Tobin. Mendelian randomisation and causal inference in observational epidemiology. *PLoS Medicine*, 5(8):e177, 2008.

Y. Y. Shugart, L. Chen, I. N. M. Day, S. J. Lewis, N. J. Timpson, W. Yuan, M. R. Abdollahi, S. M. Ring, S. Ebrahim, J. Golding, D. A. Lawlor, and G. Davey Smith. 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(8):1050–1055, 2009.

D. S. Small. Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association*, 102(479):1049–1058, 2007.

D. S. Small, J. L. Gastwirth, A. M. Krieger, and P. R. Rosenbaum. Simultaneous sensitivity analysis for observational studies using full matching or matching with multiple controls. *Statistics and Its Interface*, 2:203–211, 2009.

R. W. Snow, P. Byass, F. C. Shenton, and B. M. Greenwood. The relationship between anthropometric measurements and measurements of iron status and susceptibility to malaria in gambian children. *Transactions of The Royal Society of Tropical Medicine and Hygiene*, 85(5):584–589, 1991.

N. Solovieff, C. Cotsapas, P. H. Lee, S. M. Purcell, and J. W. Smoller. Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*, 14(7):483–495, 2013.

A. J. Sovey and D. P. Green. Instrumental variables estimation in political science: A readers guide. *American Journal of Political Science*, 55(1):188–200, 2011.

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

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

E. A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1, 2010.

S. A. Swanson and M. A. Hernán. Commentary: How to report instrumental variable analyses (suggestions welcome). *Epidemiology*, 24(3):370–374, 2013.

Z. Tan. Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association*, 101(476):1607–1618, 2006.

F. O. ter Kuile, D. J. Terlouw, S. K. Kariuki, P. A. Phillips-Howard, L. B. Mirel, W. A. Hawley, J. F. Friedman, Y. P. Shi, M. S. Kolczak, A. A. Lal, J. M. Vulule, and B. L. Nahlen. Impact of permethrin-treated bed nets on malaria, anemia, and growth in infants in an area of intense perennial malaria transmission in western kenya. *The American Journal of Tropical Medicine and Hygiene*, 68(4 suppl):68–77, 2003.

D. C. Thomas and D. V. Conti. Commentary: the concept of mendelian randomization. *International Journal of Epidemiology*, 33(1):21–25, 2004.

G. Thorleifsson, G. B. Walters, D. F. Gudbjartsson, V. Steinthorsdottir, P. Sulem, A. Helgadottir, U. Styrkarsdottir, S. Gretarsdottir, S. Thorlacius, I. Jonsdottir, T. Jonsdottir, E. J. Olafsdottir, G. H. Olafsdottir, T. Jonsson, F. Jonsson, K. Borch-Johnsen, T. Hansen, G. Andersen, T. Jorgensen, T. Lauritzen, K. K. Aben, A. L. Verbeek, N. Roeleveld, E. Kampman, L. R. Yanek, L. C. Becker, L. Tryggvadottir, T. Rafnar, D. M. Becker, J. Gulcher, L. A. Kiemeney, O. Pedersen, A. Kong, U. Thorsteinsdottir, and K. Stefansson. Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nature Genetics*, 41(1):18–24, 2008.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

R. J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7: 1456–1490, 2013.

N. J. Timpson, D. A. Lawlor, R. M. Harbord, T. R. Gaunt, I. N. M. Day, L. J. Palmer, A. T. Hattersley, S. Ebrahim, G. Lowe, A. Rumley, and G. Davey Smith. C-reactive protein and its role in metabolic syndrome: Mendelian randomisation study. *The Lancet*, 366(9501):1954–1959, 2005.

G. W. Torrance. Utility approach to measuring health-related quality of life. *Journal of Chronic Diseases*, 40(6):593–600, 1987.

K. Trakas, P. I. Oh, S. Singh, N. Risebrough, and N. H. Shear. The health status of obese individuals in canada. *International Journal of Obesity*, 25(5):662–668, 2001.

J. A. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52(3):1030–1051, 2006.

O. Ukkola, E. Ravussin, P. Jacobson, L. Sjöström, and C. Bouchard. Mutations in the adiponectin gene in lean and obese subjects from the swedish obese subjects cohort. *Metabolism*, 52(7):881–884, 2003.

United States Surgeon General. *The Health Consequences of Smoking.* US Department of Health and Human Services, 1988.

J. Wang and E. Zivot. Inference on structural parameters in instrumental variables regression with weak instruments. *Econometrica*, pages 1389–1404, 1998.

G. L. Wehby, R. L. Ohsfeldt, and J. C. Murray. mendelian randomization equals instrumental variable analysis with genetic instruments. *Statistics in Medicine*, 27(15):2745–2749, 2008.

WHO Multicentre Growth Reference Study Group. WHO child growth standards based on length/height, weight and age. *Acta Paediatrica. Supplement*, 450:76–85, 2006.

M. Willcox, A. Björkman, J. Brohult, P. Pehrson, L. Rombo, and E. Bengtsson. A case-control study in northern liberia of plasmodium falciparum malaria in haemoglobin s and beta-thalassaemia traits. *Annals of Tropical Medicine and Parasitology*, 77(3):239–246, 1983.

T. N. Williams, T. W. Mwangi, D. J. Roberts, N. D. Alexander, D. J. Weatherall, S. Wambua, M. Kortok, R. W. Snow, and K. Marsh. An immune basis for malaria protection by the sickle cell trait. *PLoS medicine*, 2(5):e128, 2005.

J. M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT press, 2nd ed. edition, 2010.

World Health Organization. World malaria report 2014. *World Health Organization*, 2014.

F. Yang, J. R. Zubizarreta, D. S. Small, S. Lorch, and P. R. Rosenbaum. Dissonant conclusions when testing the validity of an instrumental variable. *The American Statistician*, 68(4):253–263, 2014.

W.-S. Yang, P.-L. Tsou, W.-J. Lee, D.-L. Tseng, C.-L. Chen, C.-C. Peng, K.-C. Lee, M.-J. Chen, C.-J. Huang, T.-Y. Tai, and L.-M. Chuang. Allele-specific differential expression of a common adiponectin gene polymorphism related to obesity. *Journal of Molecular Medicine*, 81(7):428–434, 2003.

W.-S. Yang, Y.-C. Yang, C.-L. Chen, I.-L. Wu, J.-Y. Lu, F.-H. Lu, T.-Y. Tai, and C.-J. Chang. Adiponectin snp276 is associated with obesity, the metabolic syndrome, and diabetes in the elderly. *The American Journal of Clinical Nutrition*, 86(2):509–513, 2007.

J. R. Zubizarreta, D. S. Small, N. K. Goyal, S. Lorch, and P. R. Rosenbaum. Stronger instruments via integer programming in an observational study of late preterm birth outcomes. *The Annals of Applied Statistics*, 7(1):25–50, 2013.