Confidence Interval for Causal Effects with Possibly Invalid Instruments Even After Controlling for Many Confounders*

Zijian Guo[1], Hyunseung Kang[2], T. Tony Cai[1]and Dylan S. Small[1]

[1]Department of Statistics,The Wharton School, University of Pennsylvania
[2]Economics, Stanford Graduate School of Business, Stanford University

March 15, 2016

Abstract

The instrumental variable (IV) method is a popular method to estimate causal effects of a treatment on an outcome by using variables known as instruments to extract unconfounded variation in treatment. These instruments must satisfy the core assumptions, including (A1) association with the treatment, (A2) no direct effect on the outcome, and (A3) ignorability. These assumptions may be made more plausible by conditioning on many, possibly high-dimensional covariates, an opportunity increasingly made available by the compilation of large data sets. However, even after conditioning on a large number of covariates, it's possible that the putative instruments are invalid due to their direct effects on the outcome (violating (A2)) or

*Address for correspondence: Zijian Guo, Department of Statistics, The Wharton School, University of Pennsylvania, Jon M. Huntsman Hall, Philadelphia, PA 19104-6340, USA. Email: zijguo@wharon.upenn.edu. Zijian Guo is a Ph.D. student (Email: zijguo@wharon.upenn.edu); Hyunseung Kang is a Postdoctoral Research Fellow (Email: hskang@stanford.edu); T. Tony Cai is Dorothy Silberberg Professor of Statistics (E-mail: tcai@wharton.upenn.edu); Dylan S. Small is Professor (Email: dsmall@wharton.upenn.edu). The research of Hyunseung Kang was supported in part by NSF Grant DMS-1502437. The research of T. Tony Cai was supported in part by NSF Grants DMS-1208982 and DMS-1403708, and NIH Grant R01 CA127334. The research of Dylan S. Small was supported in part by NSF Grant SES-1260782.

violations of ignorability (violating (A3)) and the resulting inference on the causal effect may be misleading.

We propose a general inference procedure that provides honest inference in the presence of invalid IVs even after controlling for many, possibly high dimensional covariates. We demonstrate our procedure on simulated data and find that it outperforms traditional methods, especially when the instruments are invalid. We also demonstrate the usefulness of our method by analyzing the causal effect of education on earnings from the Wisconsin Longitudinal Study.

Keywords: High-dimensional covariates; sparsity; de-biasing method; exclusion restriction; treatment effect.

1 Introduction

1.1 Motivation: Causal Inference with Large Observational Data, Unmeasured Confounding, and Instrumental Variables

In health and social sciences, there is a growing trend toward large collection of diverse types of observational data to study causal effects. For example, the Alzheimer's Disease Neuroimaging Initiative (ADNI), started in 2004, is a multi-center effort to understand Alzheimer's disease by collecting large, often high-dimensional, and different types of data, including clinical, imaging, genetic, and biomarker data (Weiner et al., 2015). As another example, the Wisconsin Longitudinal Study (WLS), started in 1957, is a longitudinal survey of high school graduates from Wisconsin that contains demographic, socioeconomic, clinical, and most recently, large genetic data (Herd et al., 2014).

A central problem in analyzing observational data for causal inference is dealing with unmeasured confounding, that is dealing with other potential explanations for the treatment's effect on the outcome that weren't measured in the data. For example, suppose we want to study the causal effect of years of education on earnings from the WLS data; this question is of long standing interest in economics (Angrist and Krueger, 1991; Card, 1993, 1999) and we will revisit this question in Section 6. There are many confounders that can potentially

affect an individual's years of education and earnings, including innate ability and family background, and often, it's impractical to measure and control for all possible confounders for the effect of education on earnings.

One popular method to deduce causal effects in the presence of unmeasured confounding is by using instrumental variables (IV) analysis. An IV analysis requires variables called instruments that (A1) are related to the exposure (A2) have no direct pathway to the outcome and (A3) are not related to unmeasured variables that affect the exposure and the outcome (see Section 2.2 for details). Variables that satisfy these assumptions are referred to as valid instruments. The challenge in IV analysis is to find valid instruments and in practice, one often has a handful of candidate instruments where some of them are invalid.

To some degree, the large data sets bring some hope in finding valid instruments, in particular because (A3) may be more plausible after controlling for many covariates (see Hernán and Robins (2006) and Baiocchi et al. (2014) for discussion of the need to control for covariates for an instrument to be valid in some circumstances). For example, in studying the effect of education on earnings in WLS, a person's proximity to a college when growing up has been used as an instrument (Card, 1993, 1999). However, proximity to a college when growing up may be related to a person's socioeconomic status, characteristics of a person's high school and other covariates that may affect a person's earnings, and thus these covariates need to be controlled for in order for proximity to college to be a valid IV.

1.2 Problem: Invalid Instruments Even After Controlling for Many Confounders

Despite the promise that large data sets may bring in terms of finding valid instruments by conditioning on large number of covariates, some of the IVs may still turn out to be invalid and subsequent analysis assuming that all the IVs are valid after conditioning can be misleading. For example, suppose for studying the causal effect of education on earnings, we used proximity as an IV and to make sure the IV satisfies (A3), we control for confounders like high school test scores of the student, high school size, individual's genetic makeup, family education, and family's socioeconomic status. But, if living close to college had other

benefits beyond getting more education, say by being exposed to many programs available to high school students for job preparation and employers who come to the area to discuss employment opportunities for college students, then the IV, proximity to college, can directly affect individual's earning potential and violate (A2) (Card, 1999).

Our paper attempts to tackle the problem of causal effect estimation under this setup. Specifically, motivated by availability of data sets with many covariates that can make plausible the IV assumptions, but the fact that invalid instruments may still be present even after controlling for many covariates, we explore inferring causal effects that are robust to invalid instruments even after many, possibly high-dimensional controls in the form of confidence interval estimation.

1.3 Prior work and our contributions

There is prior work in IV literature with high dimensional covariates and/or instruments (Belloni et al., 2012; Chernozhukov et al., 2015; Fan and Liao, 2014; Gautier and Tsybakov, 2011). Unfortunately, all of them assume that after controlling for confounders, all the IVs are valid. Recently, there is work on estimating causal effects in the presence of invalid instruments Kang et al. (2015); Kolesár et al. (2015), but these methods do not incorporate controlling for high dimensional covariates. More importantly, Kang et al. (2015) only provides a consistent estimate of the treatment effect and not confidence intervals.

Our contribution is a unification of both lines of work, examining invalid instruments even after controlling for a high dimensional number of covariates. First, we propose a general confidence interval estimator that provides honest coverage of the causal effect of the exposure on the outcome in the presence of possibly invalid instruments even after controlling for high dimensional covariates. Second, we provide theoretical guarantees of our confidence interval. An interesting feature of the theoretical analysis is that the presence of invalid instruments forces us to consider assumptions on instruments globally and individually. Third, we conduct a simulation study to assess the performance of our confidence interval under various settings. Finally, we provide a real data analysis using our method, specifically revisiting the question about the causal effect of years of schooling on income

using data from the Wisconsin Longitudinal Study.

2 Model

2.1 Notation

To define causal effects, the potential outcome 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 $\mathbf{z} \in \mathbb{R}^{p_z}$. Let $D_i^{(\mathbf{z})} \in \mathbb{R}$ be the potential exposure if the individual had instruments $\mathbf{z} \in \mathbb{R}^{p_z}$. 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 $\mathbf{Z}_{i\cdot} \in \mathbb{R}^{p_z}$ and exposure D_i. We also denote pre-instrument covariates for each individual i as $\mathbf{X}_{i\cdot} \in \mathbb{R}^{p_x}$. In total, n sets of outcome, exposure, and instruments, denoted as $(Y_i, D_i, \mathbf{Z}_{i\cdot}, \mathbf{X}_{i\cdot})$, 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/treatment, $\mathbf{Z}$ to be a n by p_z matrix of instruments where row i consists of $\mathbf{Z}_{i\cdot}$, and $\mathbf{X}$ to be an n by p_x matrix of covariates where row i consists of $\mathbf{X}_{i\cdot}$. Let $\mathbf{W}$ be an n by $p = p_z + p_x$ matrix where $\mathbf{W}$ is a result of concatenating the matrices $\mathbf{Z}$ and $\mathbf{X}$. For any vector $\mathbf{v} \in \mathbb{R}^p$, let $\mathbf{v}_j$ denote the jth element of $\mathbf{v}$. Let $\|\mathbf{v}\|_1$, $\|\mathbf{v}\|_2$, and $\|\mathbf{v}\|_\infty$ denote the usual $1, 2$ and ∞-norms, respectively. Let $\|\mathbf{v}\|_0$ denote the number of non-zero elements in $\mathbf{v}$ and $\text{supp}(\mathbf{v}) \subseteq \{1, \ldots, p\}$, is defined as $\{j : \mathbf{v}_j \neq 0\}$.

For any n by p matrix $\mathbf{M} \in \mathbb{R}^{n \times p}$, we denote the (i, j) element of matrix $\mathbf{M}$ as M_{ij}, the ith row as $\mathbf{M}_{i\cdot}$, and the jth column as $\mathbf{M}_{\cdot j}$. Let $\mathbf{M}^\intercal$ be the transpose of $\mathbf{M}$ and $\|\mathbf{M}\|_\infty$ represent the element-wise matrix sup norm of matrix $\mathbf{M}$. For a sequence of random variables X_n, we use $X_n \xrightarrow{p} X$ to represent that X_n converges to X in probability. Finally, for any two sequences a_n and b_n, we will write $a_n \gg b_n$ if $\limsup \frac{b_n}{a_n} = 0$ and write $a_n \ll b_n$ if $b_n \gg a_n$. Also, for a set J, $|J|$ denotes its cardinality.

2.2 Model and Instrumental Variables Assumptions

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) and Kang et al. (2015). For two possible values of the exposure d', d and instruments $\mathbf{z}', \mathbf{z}$, we assume the following potential outcomes model

$$Y_i^{(d',\mathbf{z}')} - Y_i^{(d,\mathbf{z})} = (\mathbf{z}' - \mathbf{z})^\intercal \boldsymbol{\kappa}^* + (d' - d)\beta^*, \quad E(Y_i^{(0,\mathbf{0})} \mid \mathbf{Z}_{i.}, \mathbf{X}_{i.}) = Z_{i.}^\intercal \boldsymbol{\eta}^* + X_{i.}^\intercal \boldsymbol{\phi}^* \tag{1}$$

where $\boldsymbol{\kappa}^*, \beta^*, \boldsymbol{\eta}^*$, and $\boldsymbol{\phi}^*$ 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 $\boldsymbol{\phi}^*$ represents the impact of covariates on the baseline potential outcome $Y_i^{(0,\mathbf{0})}$. The parameter $\boldsymbol{\kappa}^*$ represents violation of (A2), the direct effect of the instruments on the outcome. If (A2) holds, then $\boldsymbol{\kappa}^* = 0$. The parameter $\boldsymbol{\eta}^*$ represents violation of (A3), the presence of unmeasured confounding between the instrument and the outcome. If (A3) holds, then $\boldsymbol{\eta}^* = 0$.

Let $\boldsymbol{\pi}^* = \boldsymbol{\kappa}^* + \boldsymbol{\eta}^*$ and $\epsilon_{i1} = Y_i^{(0,\mathbf{0})} - E(Y_i^{(0,\mathbf{0})} \mid \mathbf{Z}_{i.}, \mathbf{X}_{i.})$. When we combine equation (1) along with the definition of ϵ_{i1}, the observed data model becomes

$$Y_i = \mathbf{Z}_{i.}^\intercal \boldsymbol{\pi}^* + D_i\beta^* + \mathbf{X}_{i.}^\intercal \boldsymbol{\phi}^* + \epsilon_{i1}, \quad E(\epsilon_{i1} \mid \mathbf{Z}_{i.}, \mathbf{X}_{i.}) = 0. \tag{2}$$

The observed model is also known as the under-identified single-equation linear model in econometrics (page 83 of Wooldridge (2010)). This model is not a usual regression model because D_i might be correlated with ϵ_{i1}. In particular, the parameter β^* measures the causal effect of changing D on Y rather than an association. The parameter $\boldsymbol{\pi}^*$ in model (2) combines both the violation of (A2), represented by $\boldsymbol{\kappa}^*$, and the violation of (A3), represented by $\boldsymbol{\eta}^*$. If both (A2) and (A3) are satisfied, then $\boldsymbol{\kappa}^* = \boldsymbol{\eta}^* = 0$ and $\boldsymbol{\pi}^* = 0$. Hence, $\boldsymbol{\pi}^*$ captures the violations of (A2) and (A3). We formalize this notion with the following definition.

Definition 1. *Suppose we have p_z candidate instruments along with the models* (1)–(2)*. We say that instrument $j = 1, \ldots, p_z$ satisfies* (A2) − (A3) *if $\boldsymbol{\pi}_j^* = 0$.*

We also assume a linear association/observational model between the endogenous variable D_i, the instruments $\mathbf{Z}_{i.}$, and the covariates $\mathbf{X}_{i.}$.

$$D_i = \mathbf{Z}_{i.}^\intercal \boldsymbol{\gamma}^* + \mathbf{X}_{i.}^\intercal \boldsymbol{\psi}^* + \epsilon_{i2}, \quad E(\epsilon_{i2} | \mathbf{Z}_{i.}, \mathbf{X}_{i.}) = 0 \tag{3}$$

Each element $\boldsymbol{\gamma}_j^*$ is the partial correlation between the jth instrument and D. The parameter $\boldsymbol{\psi}^*$ represents the association between the covariates and D_i. Also, unlike the models (1)-(2), we do not need a causal model between D_i, $\mathbf{Z}_{i.}$, and $\mathbf{X}_{i.}$; only the association model (3) is sufficient for our method. Finally, for notation, we let $s_{\text{z2}} = \|\boldsymbol{\pi}^*\|_0$, $s_{\text{x2}} = \|\boldsymbol{\phi}^*\|_0$, $s_{\text{z1}} = \|\boldsymbol{\gamma}^*\|_0$ and $s_{\text{x1}} = \|\boldsymbol{\psi}^*\|_0$.

Based on model (3), we can define a set of instruments that satisfy (A1).

Definition 2. *Suppose we have p_{z} candidate instruments along with the model* (3)*. We say that instrument $j = 1, \ldots, p_z$ satisfies* (A1)*, or is an individually strong IV, if $\boldsymbol{\gamma}_j^* \neq 0$ and denote $\mathcal{S}^*$ to be the set of these instruments.*

Typically, satisfying (A1) has been defined in a global sense where (A1) is satisfied if $\boldsymbol{\gamma}^* \neq 0$ (Wooldridge, 2010). However, this global definition can be misleading in the presence of multiple candidate instruments. For example, it's possible that $\boldsymbol{\gamma}_1^* \gg 0$ while $\boldsymbol{\gamma}_j^* = 0$ for all $j \neq 1$ so that only the first instrument has an effect on the exposure while the rest do not. Using the global definition would imply that all the p_z instruments satisfy (A1) while Definition 2 makes it explicit and, perhaps less ambiguous, that it's only the first instrument $j = 1$ that satisfies (A1). Nevertheless, both the traditional global definition and Definition 2 are equivalent if $\boldsymbol{\gamma}_j^* \neq 0$ for all j, that is where we only include relevant instruments, which is typically the case in practice and is the scenario studied by Kang et al. (2015). Finally, when there is only one candidate instrument so that $p_{\text{z}} = 1$, both definitions are equivalent to the definition presented in Holland (1988) and both become a special case of the definition presented in Angrist et al. (1996) under an additive, linear, constant effects model. In short, Definition 2 agrees with most definitions of satisfying (A1) in the literature. Also, in economics, Definition 2 is the same as the notion of non-redundant instruments (Cheng and Liao, 2015).

Combining Definition 2 and 1, we can formally define a set of valid instruments, i.e. the set of instruments that satisfy all (A1)-(A3), as follows.

Definition 3. *Suppose we have p_{z} candidate instruments along with the models* (1)–(3)*. We say that instrument $j = 1, \ldots, p_z$ is valid, i.e. satisfies all* (A1)$-$(A3)*, if $\boldsymbol{\pi}_j^* = 0$ and $\boldsymbol{\gamma}_j^* \neq 0$. Let $\mathcal{V}^*$ be the set of valid instruments.*

When there is only one instrument, $p_z = 1$, Definition 3 of a valid instrument is identical to the definition of a valid instrument in Holland (1988). Specifically, Definition 2 satisfies assumption (A1) that the instrument is related to the exposure. Also, assumption (A2), the exclusion restriction, which means $Y_i^{(d,\mathbf{z})} = Y_i^{(d,\mathbf{z}')}$ for all $d, \mathbf{z}, \mathbf{z}'$, is equivalent to $\boldsymbol{\kappa}^* = 0$ and assumption (A3), no unmeasured confounding, which means $Y_i^{(d,\mathbf{z})}$ and $D_i^{(\mathbf{z})}$ are independent of $Z_{i\cdot}$ for all d and $\mathbf{z}$, is equivalent to $\boldsymbol{\eta}^* = 0$, implying $\boldsymbol{\pi}^* = \boldsymbol{\kappa}^* + \boldsymbol{\eta}^* = \mathbf{0}$. Definition 3 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, $p_{\mathrm{z}} > 1$, are present, our models (1)–(3) and Definition 3 can be viewed as a generalization of the definition of valid instruments in Holland (1988).

While models presented above make strong assumptions, specifically the additive, constant, linear effects, we can also extend the model in (2) to include heterogeneous causal effects and non-linear effects (Kang et al., 2015). Furthermore, both the observed models in (2) and (3) are very popular and standard models to study instrumental variables in econometrics (Wooldridge, 2010), with the important exception that (i) the model in (2) allows for possibly invalid instruments and (ii) we allow the number of covariates p_x (and even the number of instruments p_z) to be larger than the sample size n. In short, our setup allows us to model motivating data sets discussed in Section 1.1 by allowing for many pre-instrument covariates $\mathbf{X}_{i\cdot}$ (e.g. genetic data from an observational data) to make the IV assumptions (A1)-(A3) more plausible, but also leaves the possibility that the instruments may still be invalid even after controlling for many covariates.

3 Confidence Interval Estimation

3.1 Estimating the Center and the Width of the Interval

There are two estimands that comprise estimating the confidence interval under our setting: the center, $\widehat{\beta}$, and the width of the interval. We start the discussion with estimating the center of the interval.

For estimating the center $\widehat{\beta}$, it's instructive to split the estimation into two parts. The

first part involves dealing with the problem posed by the high-dimensional nature of our covariates. The second part involves dealing with the problem posed by the invalid instruments that may be present despite our best efforts to avoid them by adding many covariates.

First, to deal with the high dimensional covariates $\mathbf{X}_{i.}$, we formulate the observed models (2) and (3) as reduced-forms models where both models are only functions of $\mathbf{Z}_{i.}$ and $\mathbf{X}_{i.}$

$$Y_i = \mathbf{Z}_{i.}^\intercal \mathbf{\Gamma}^* + \mathbf{X}_{i.}^\intercal \Psi^* + e_{i1} \tag{4}$$

$$D_i = \mathbf{Z}_{i.}^\intercal \boldsymbol{\gamma}^* + \mathbf{X}_{i.}^\intercal \boldsymbol{\psi}^* + \epsilon_{i2} \tag{5}$$

where $\mathbf{\Gamma}^* = \beta^* \boldsymbol{\gamma}^* + \boldsymbol{\pi}^*$ and $\Psi^* = \boldsymbol{\phi}^* + \beta^* \boldsymbol{\psi}^*$ are the parameters of the reduced-form model and $e_{i1} = \beta^* \epsilon_{i2} + \epsilon_{i1}$ is the reduced-form error term in (4). The errors in the reduced-models have the property that $E(e_{i1}|\mathbf{Z}_{i.}, \mathbf{X}_{i.}) = 0$ and $E(\epsilon_{i2}|\mathbf{Z}_{i.}, \mathbf{X}_{i.}) = 0$ along with the covariance matrix $\mathbf{\Theta}^*$ with elements $\mathbf{\Theta}^*_{11} = Var(e_{i1}|\mathbf{Z}_{i.}, \mathbf{X}_{i.})$, $\mathbf{\Theta}^*_{22} = Var(\epsilon_{i2}|\mathbf{Z}_{i.}, \mathbf{X}_{i.})$, and $\mathbf{\Theta}^*_{12} = \text{Cov}(e_{i1}, \epsilon_{i2}|\mathbf{Z}_{i.}, \mathbf{X}_{i.})$. Thus, each equation in the reduced-form model is a usual (high dimensional) regression model with (high dimensional) covariates $\mathbf{Z}_{i.}$ and $\mathbf{X}_{i.}$ and outcomes Y_i and D_i, respectively.

There are many methods in the literature to estimate the parameters of high dimensional regression models like the reduced-form models in (4) and (5). One approach is the scaled Lasso estimator proposed by Sun and Zhang (2012),

$$\{\widehat{\mathbf{\Gamma}}, \widehat{\mathbf{\Psi}}, \widehat{\mathbf{\Theta}}_{11}\} = \underset{\mathbf{\Gamma} \in \mathbb{R}^{p_z}, \mathbf{\Psi} \in \mathbb{R}^{p_x}, \mathbf{\Theta}_{11} \in \mathbb{R}^+}{\text{argmin}} \frac{\|\mathbf{Y} - \mathbf{Z}\mathbf{\Gamma} - \mathbf{X}\mathbf{\Psi}\|_2^2}{2n\sqrt{\mathbf{\Theta}_{11}}} + \frac{\sqrt{\mathbf{\Theta}_{11}}}{2} + \frac{\lambda_0}{\sqrt{n}} \left(\sum_{j=1}^{p_z} \|\mathbf{Z}_{.j}\|_2 |\mathbf{\Gamma}_j| + \sum_{j=1}^{p_x} \|\mathbf{X}_{.j}\|_2 |\mathbf{\Psi}_j| \right) \tag{6}$$

for the reduced model in (4) and

$$\{\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\psi}}, \widehat{\mathbf{\Theta}}_{22}\} = \underset{\mathbf{\Gamma} \in \mathbb{R}^{p_z}, \mathbf{\Psi} \in \mathbb{R}^{p_x}, \mathbf{\Theta}_{22} \in \mathbb{R}^+}{\text{argmin}} \frac{\|\mathbf{D} - \mathbf{Z}\boldsymbol{\gamma} - \mathbf{X}\boldsymbol{\psi}\|_2^2}{2n\sqrt{\mathbf{\Theta}_{22}}} + \frac{\sqrt{\mathbf{\Theta}_{22}}}{2} + \frac{\lambda_0}{\sqrt{n}} \left(\sum_{j=1}^{p_z} \|\mathbf{Z}_{.j}\|_2 |\boldsymbol{\gamma}_j| + \sum_{j=1}^{p_x} \|\mathbf{X}_{.j}\|_2 |\boldsymbol{\psi}_j| \right) \tag{7}$$

for the reduced model in (5). The term λ_0 in both estimation problems (6) and (7) represents the penalty term in the scaled Lasso estimator and we choose $\lambda_0 = \sqrt{2.05 \log p/n}$. Here, the constant 2.05 is chosen out of convenience and will be for the remainder of the text. But, 2.05 can be any arbitrary number greater than 2 and our results will hold. Also, we can estimate $\mathbf{\Theta}^*_{12}$ from the estimation problems (6) and (7) by $\widehat{\mathbf{\Theta}}_{12} = 1/n \left(\mathbf{Y} - \mathbf{Z}\widehat{\mathbf{\Gamma}} - \mathbf{X}\widehat{\mathbf{\Psi}} \right)^\intercal \left(\mathbf{D} - \mathbf{Z}\widehat{\boldsymbol{\gamma}} - \mathbf{X}\widehat{\boldsymbol{\psi}} \right)$.

Unfortunately, most penalized estimators for high dimensional regression problems are biased and the scaled Lasso estimators are no exception. In our case, using the estimates, say $\widehat{\mathbf{\Gamma}}$ and $\widehat{\boldsymbol{\gamma}}$, are biased for the parameters that they estimate $\mathbf{\Gamma}^*$ and $\boldsymbol{\gamma}^*$. Thankfully, recent works by Javanmard and Montanari (2014); van de Geer et al. (2014); Zhang and Zhang (2014) and Cai and Guo (2016) allow us to de-bias these biased estimates. Specifically, let $\mathbf{W}$ be the concatenated matrix of the instruments $\mathbf{Z}$ and the covariates $\mathbf{X}$. Suppose we solve p_z optimization problems where the solution to each p_z optimization problem, denoted as $\widehat{\mathbf{u}}^{[j]} \in \mathbb{R}^p$, $j = 1, \ldots, p_z$, is

$$\widehat{\mathbf{u}}^{[j]} = \underset{\mathbf{u}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n}\|\mathbf{W}\mathbf{u}\|_2^2 \quad \text{s.t.} \quad \|\frac{1}{n}\mathbf{W}^\intercal\mathbf{W}\mathbf{u} - \mathbf{I}_{\cdot j}\|_\infty \leq \lambda_n. \tag{8}$$

The tuning parameter λ_n is chosen to be $12M_1^2\sqrt{\log p/n}$ with M_1 defined as the largest eigenvalue of the covariance matrix $\mathbf{\Sigma} = \mathrm{Cov}(\mathbf{W}_{i\cdot})$. Let $\widehat{\mathbf{U}}$ denote the concatenation of the p_z solutions to the optimization problem, i.e. $\widehat{\mathbf{U}} = (\widehat{\mathbf{u}}^{[1]}, \ldots, \widehat{\mathbf{u}}^{[p_z]})^\intercal$. Then, the debiased estimates of $\widehat{\mathbf{\Gamma}}$ and $\widehat{\boldsymbol{\gamma}}$, denoted as $\widetilde{\mathbf{\Gamma}}$ and $\widetilde{\boldsymbol{\gamma}}$, are

$$\widetilde{\mathbf{\Gamma}} = \widehat{\mathbf{\Gamma}} + \frac{1}{n}\widehat{\mathbf{U}}\mathbf{W}^\intercal\left(\mathbf{Y} - \mathbf{Z}\widehat{\mathbf{\Gamma}} - \mathbf{X}\widehat{\mathbf{\Psi}}\right), \quad \widetilde{\boldsymbol{\gamma}} = \widehat{\boldsymbol{\gamma}} + \frac{1}{n}\widehat{\mathbf{U}}\mathbf{W}^\intercal\left(\mathbf{D} - \mathbf{Z}\widehat{\boldsymbol{\gamma}} - \mathbf{X}\widehat{\boldsymbol{\psi}}\right). \tag{9}$$

In short, to overcome the high dimensionality of the covariates, we used scaled Lasso along with de-biasing methods on the reduced-form models to obtain de-biased estimates $\widetilde{\mathbf{\Gamma}}$ and $\widetilde{\boldsymbol{\gamma}}$ of the reduced-form parameters.

The second step in estimating $\widehat{\beta}$, the center of our confidence interval, is to deal with the problem posed by invalid IVs. Specifically, we need to (i) find IVs that satisfy (A1), that is the set $\mathcal{S}^*$ in Definition 2, and (ii) use them to find valid IVs that satisfy (A1)-(A3), that is the set $\mathcal{V}^*$ in Definition 3. First, an estimate of $\mathcal{S}^*$, denoted as $\widetilde{\mathcal{S}}$, can be found by thresholding the coefficients of the de-biased estimate $\widetilde{\boldsymbol{\gamma}}$

$$\widetilde{\mathcal{S}} = \left\{ j : |\widetilde{\boldsymbol{\gamma}}_j| \geq \frac{\sqrt{\widehat{\mathbf{\Theta}}_{22}}\|\mathbf{W}\widehat{\mathbf{u}}^{[j]}\|_2}{\sqrt{n}}\sqrt{\frac{2.05\log p_z}{n}} \right\}. \tag{10}$$

The threshold is based on the noise level of $\widetilde{\boldsymbol{\gamma}}_j$ in (9) (represented by $\sqrt{\widehat{\mathbf{\Theta}}_{22}}\|\mathbf{W}\widehat{\mathbf{u}}^{[j]}\|_2/n$), adjusted by dimensionality of the instrument size (represented by $\sqrt{2.05\log p_z}$).

Next, we estimate the set of valid instruments $\mathcal{V}^*$ that satisfy (A1)-(A3) by estimating $\boldsymbol{\pi}^*$. Specifically, by Definition 1, the set of instruments where $\boldsymbol{\pi}_j^* = 0$ satisfy (A2) and (A3) and consequently, intersecting this set with $\widetilde{\mathcal{S}}$, which are estimates of IVs that satisfy (A1), allow us to estimate the set of valid instruments $\mathcal{V}^*$ that satisfy (A1)-(A3). To estimate $\boldsymbol{\pi}^*$, for each strong instrument j in $\widetilde{\mathcal{S}}$, we define $\widehat{\beta}^{[j]}$ to be the estimate of β^* from using this strong instrument by dividing the reduced-form estimates, i.e. $\widehat{\beta}^{[j]} = \widetilde{\boldsymbol{\Gamma}}_j / \widetilde{\boldsymbol{\gamma}}_j$, and $\widehat{\boldsymbol{\pi}}^{[j]}$ to be the estimate of $\boldsymbol{\pi}^*$ using this jth instrument's estimate of β^*, i.e. $\widehat{\boldsymbol{\pi}}^{[j]} = \widetilde{\boldsymbol{\Gamma}} - \widehat{\beta}^{[j]}\widetilde{\boldsymbol{\gamma}}$. For each $\widehat{\boldsymbol{\pi}}^{[j]}$ in $j \in \widetilde{\mathcal{S}}$, we threshold each element of $\widehat{\boldsymbol{\pi}}^{[j]}$ to create the thresholded estimate $\widetilde{\boldsymbol{\pi}}^{[j]}$,

$$\widetilde{\boldsymbol{\pi}}_k^{[j]} = \widehat{\boldsymbol{\pi}}_k^{[j]} \mathbf{1}\left(k \in \widetilde{\mathcal{S}} \ \cap \ |\widehat{\pi}_k^{[j]}| \geq 2.05 \sqrt{\widehat{\boldsymbol{\Theta}}_{11} + (\widehat{\beta}^{[j]})^2 \widehat{\boldsymbol{\Theta}}_{22} - 2\widehat{\beta}^{[j]} \widehat{\boldsymbol{\Theta}}_{12}} \frac{\|\mathbf{W}(\widehat{\mathbf{u}}^{[k]} - \frac{\widetilde{\gamma}_k}{\widetilde{\gamma}_j} \widehat{\mathbf{u}}^{[j]})\|_2}{\sqrt{n}} \sqrt{\frac{\log p_z}{n}} \right) \tag{11}$$

for all $1 \leq k \leq p_z$. Each thresholded estimate $\widetilde{\boldsymbol{\pi}}^{[j]}$ is obtained by looking at the elements of the un-thresholded estimate, $\widehat{\boldsymbol{\pi}}^{[j]}$, and examining whether each element of it exceeds the noise threshold, similar to the thresholding strategy in (10). In the end, we have $|\widetilde{\mathcal{S}}|$ estimates of $\widetilde{\boldsymbol{\pi}}^{[j]}$ based on each individually strong instrument in $\widetilde{\mathcal{S}}$, where each $\widetilde{\boldsymbol{\pi}}^{[j]}$ represents an estimate of $\boldsymbol{\pi}^*$.

The final estimate of β^* is by picking the $\widetilde{\boldsymbol{\pi}}^{[j]}$ that has the most valid instruments, or equivalently choosing $\widetilde{\boldsymbol{\pi}}^{[j^*]}$, $j^* \in \widetilde{\mathcal{S}}$, where $j^* = \underset{j}{\operatorname{argmin}} \|\widetilde{\boldsymbol{\pi}}^{[j]}\|_0$; if there is a non-unique solution, we choose $\widetilde{\boldsymbol{\pi}}^{[j]}$ with the smallest ℓ_1 norm, the closest convex norm of ℓ_0. We estimate the set of valid instruments $\widetilde{\mathcal{V}} \subseteq \{1, \ldots, p_z\}$ as those elements of $\widetilde{\boldsymbol{\pi}}^{[j^*]}$ that are zero,

$$\widetilde{\mathcal{V}} = \widetilde{\mathcal{S}} \setminus \operatorname{supp}\left(\widetilde{\boldsymbol{\pi}}^{[j^*]}\right) \quad \text{where} \quad j^* = \underset{j}{\operatorname{argmin}} \ \|\widetilde{\boldsymbol{\pi}}^{[j]}\|_0. \tag{12}$$

Then, using the estimated $\widetilde{\mathcal{V}}$, we obtain our estimate of β^* and consequently, the center of our confidence interval as

$$\widehat{\beta} = \frac{\sum_{j \in \widetilde{\mathcal{V}}} \widetilde{\boldsymbol{\gamma}}_j \widetilde{\boldsymbol{\Gamma}}_j}{\sum_{j \in \widetilde{\mathcal{V}}} \widetilde{\boldsymbol{\gamma}}_j^2}. \tag{13}$$

Note that $\widehat{\beta}$ in (13) has the "correct" form in that if, by chance, we correctly estimated the set of valid instruments $\mathcal{V}^*$ and our debiased estimates of the reduced-form parameters, $\widetilde{\boldsymbol{\Gamma}}$ and $\widetilde{\boldsymbol{\gamma}}$, are perfect estimates of the reduced-form parameters, our estimate of β^* in (13) would become $\widehat{\beta} = \sum_{j \in \widetilde{\mathcal{V}}} \widetilde{\boldsymbol{\gamma}}_j \widetilde{\boldsymbol{\Gamma}}_j / \sum_{j \in \widetilde{\mathcal{V}}} \widetilde{\boldsymbol{\gamma}}_j^2 = \sum_{j \in \mathcal{V}^*} \boldsymbol{\gamma}_j^{*2} \beta^* / \sum_{j \in \mathcal{V}^*} \boldsymbol{\gamma}_j^{*2} = \beta^*$. Hence, our estimator in

(13) would identify β^*. Clearly, we would never have a perfect estimate of the set $\mathcal{V}^*$ or the reduced-form parameters in finite sample and Section 4 describes the properties of our estimate $\widehat{\beta}$ under these uncertainties.

Once we found an estimate for the center of our confidence interval, $\widehat{\beta}$, we estimate the width of the interval. In particular, we estimate the standard error of $\widehat{\beta}$ by $\sqrt{\widehat{\mathrm{V}}/n}$ with

$$\widehat{\mathrm{V}} = \frac{\left\| \sum_{j \in \widetilde{\mathcal{V}}} \widetilde{\boldsymbol{\gamma}}_j \frac{1}{\sqrt{n}} \mathbf{W} \widehat{\mathbf{u}}^{[j]} \right\|_2^2}{\left(\sum_{j \in \widetilde{\mathcal{V}}} \widetilde{\boldsymbol{\gamma}}_j^2 \right)^2} \left(\widehat{\boldsymbol{\Theta}}_{11} + \widehat{\beta}^2 \widehat{\boldsymbol{\Theta}}_{22} - 2\widehat{\beta} \widehat{\boldsymbol{\Theta}}_{12} \right). \tag{14}$$

In equation (14), we see some familiar variance expressions from the traditional IV literature. For example, the $\widehat{\boldsymbol{\Theta}}_{11} + \widehat{\beta}^2 \widehat{\boldsymbol{\Theta}}_{22} - 2\widehat{\beta} \widehat{\boldsymbol{\Theta}}_{12}$ in (15) is the usual IV variance estimator of the error term in our original model (2). But, our variance estimator is scaled by terms that depend on the estimated set of valid instruments $\widetilde{\mathcal{V}}$.

Finally, having both the estimates of the center, $\widehat{\beta}$ and the standard error of $\widehat{\beta}$, our robust confidence interval takes on the usual form

$$\left(\widehat{\beta} - (1 + \eta_0) z_{1-\alpha/2} \sqrt{\widehat{\mathrm{V}}/n}, \quad \widehat{\beta} + (1 + \eta_0) z_{1-\alpha/2} \sqrt{\widehat{\mathrm{V}}/n} \right), \tag{15}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution and η_0 is any positive constant. The constant η_0 serves to improve the finite sample performance of the confidence interval and in practice, we choose $\eta_0 = 0.05$. For completeness, our confidence interval estimation procedure is outlined in Procedure 1.

We now make a few comments about our confidence interval. First, Procedure 1 provides a general method to construct confidence intervals in the presence of invalid IVs and high dimensional covariates in that the resulting output will be valid if any, or both, of these two problems are present. For example, if the data is low-dimensional, but contains invalid instruments, one can still apply Procedure 1 without any modification and still obtain valid confidence intervals. If the data only contains valid instruments, unrealistic in practice, one can still apply Procedure 1, again without any modification and still obtain valid confidence intervals. However, depending on particular data sets one may have, the procedure can be modified for simplicity and we discuss a couple important cases in Section 3.2.

Second, in Kang et al. (2015), the authors provided an estimator for β^* under a special case of our setup where all the IVs are relevant, i.e. $\boldsymbol{\gamma}_j^* \neq 0$ for all j, but did not derive

Procedure 1 Confidence Interval for β^* under Invalid IVs with High-Dimensional Confounders

Input: Outcome $\mathbf{Y}$, treatment $\mathbf{D}$, instrument $\mathbf{Z}$, covariates $\mathbf{X}$, significance level α

STEP 1: Estimate the center, $\widehat{\beta}$

STEP 1a: Compute de-biased scaled Lasso estimates of reduced-form parameters, $\widetilde{\mathbf{\Gamma}}$ and $\widetilde{\boldsymbol{\gamma}}$, by (6)-(9)

STEP 1b: Estimate valid IVs, $\widetilde{\mathcal{V}}$, by estimating $\widetilde{\mathcal{S}}$ and $\widetilde{\boldsymbol{\pi}}^{[j]}$ for $j \in \widetilde{\mathcal{S}}$ in (10)-(12)

STEP 1c: Combine outputs from STEP 1a and STEP 1b to estimate $\widehat{\beta}$ in (13)

STEP 2: Estimate the width $(1+\eta_0)z_{1-\alpha/2}\sqrt{\widehat{\mathrm{V}}/n}$ with STEP 1a, STEP 1b, and (14)

STEP 3: Combine STEP 1 and STEP 2 to compute the confidence interval in (15)

Output: $1-\alpha$ Confidence interval for β^*

confidence intervals. We extend their work by not only provides an estimator for β^* under more general settings, but also providing robust confidence intervals under these general settings.

3.2 Special Cases

While Procedure 1 provides a general method to construct confidence intervals with possibly invalid IVs and many covariates, the procedure can be simplified depending on the data sets one may have. We discuss two special cases in this section; other cases are also discussed in Section 1 of the Supplementary Materials.

The first case is when we have low-dimensional covariates (i.e. the sample size exceeds the number of covariates), but we do not assume all the IVs are valid. Under this case, we can simply use the ordinary least square (OLS) estimates for the reduced forms instead of the debiased scaled Lasso estimates in STEP 1a of Procedure 1. Also, as a result of using OLS in STEP 1a, we need to modify STEP 1b slightly where we replace $\widehat{\mathbf{u}}^{[j]}$ from (8) with $\widehat{\mathbf{u}}^{[j]} = (\mathbf{W}^\intercal\mathbf{W}/n)^{-1}_{j\cdot}$, the jth row of the estimated precision matrix for $\mathbf{W}$. STEP 2 will remain unchanged except for using the OLS estimates of the noise variance. Finally, combining STEP1 and STEP2 as before would result in a robust confidence interval for β^*.

The second case is when we have high dimensional covariates, but we assume all the

instruments are valid. This setup was considered in Belloni et al. (2012); Chernozhukov et al. (2015); Fan and Liao (2014); Gautier and Tsybakov (2011). Under this case, our procedure doesn't change except in STEP 1b where $\widetilde{\mathcal{V}} = \widetilde{\mathcal{S}}$ and the resulting estimator of β^* is denoted as

$$\widehat{\beta}_0 = \frac{\sum_{j \in \widetilde{\mathcal{S}}} \widetilde{\boldsymbol{\gamma}}_j \widetilde{\boldsymbol{\Gamma}}_j}{\sum_{j \in \widetilde{\mathcal{S}}} \widetilde{\boldsymbol{\gamma}}_j^2}. \tag{16}$$

After this modification, we can proceed with STEP2 and STEP3 to obtain an estimate of the confidence interval for β^*

$$\left(\widehat{\beta}_0 - (1+\eta_0) z_{1-\alpha/2} \sqrt{\widehat{\mathrm{V}}_0 / n}, \quad \widehat{\beta}_0 + (1+\eta_0) z_{1-\alpha/2} \sqrt{\widehat{\mathrm{V}}_0 / n} \right). \tag{17}$$

Here, $\widehat{\mathrm{V}}_0$ is $\widehat{\mathrm{V}}$ in (14) except we replace $\widetilde{\mathcal{V}} = \widetilde{\mathcal{S}}$ and $\widehat{\beta}$ with $\widehat{\beta}_0$.

In summary, the modifications can be broadly categorized into two general themes, (i) the presence of all valid instruments and (ii) the presence of low dimensional covariates. The presence of all valid instruments replaces STEP 1b by setting $\widetilde{\mathcal{V}} = \widetilde{\mathcal{S}}$. The presence of low-dimensional covariates swaps out STEP 1a with OLS estimates of the reduced form parameters. These modifications streamline our procedure for easier use in practice. However, as mentioned before, one can simply use the general Procedure 1 without any modification under any of these cases to obtain valid confidence intervals for the treatment effect.

4 Theoretical Results

In this section, we assess the properties of the confidence interval proposed in Procedure 1. Before we begin, we first make the usual regularity assumptions in high dimensional inference (Bickel et al., 2009; Bühlmann and van de Geer, 2011; Cai and Guo, 2016).

(R1) (Sparsity): $\max\{s_{\mathrm{z2}}, s_{\mathrm{z1}}, s_{\mathrm{x2}}, s_{\mathrm{x1}}\} \leq s$.

(R2) (Coherence): $\mathbf{W}_{i\cdot}$ has a covariance matrix $\boldsymbol{\Sigma}^*$ where $1/M_1 \leq \lambda_{\min}(\boldsymbol{\Sigma}^*) \leq \lambda_{\max}(\boldsymbol{\Sigma}^*) \leq M_1$ for some constant $M_1 > 1$ and has bounded sub-Gaussian norm.

Assumption (R1) imposes sparsity on the parameters so that our problem is tractable in high dimensions. Assumption (R2) places a condition on the spectrum of the design matrix

$\mathbf{W}$ and the tail distribution of $\mathbf{W}_{i,\cdot}$, which is related to restricted eigenvalue condition in Bickel et al. (2009). For simplicity, we also assume that the sub-Gaussian norm of $\mathbf{W}_{i\cdot}$ is upper bounded by M_1, that is, $\sup_{\boldsymbol{v}\in S^{p-1}} \sup_{q\geq 1} (\mathbf{E}|\boldsymbol{v}^\intercal \mathbf{W}_{i\cdot}|^q/q)^{\frac{1}{q}} \leq M_1$ where S^{p-1} is the unit sphere in $\mathbb{R}^p$; see Vershynin (2012) for details on sub-Gaussian random variables and bounds.

Next, we make the usual regularity assumptions that are encountered in instrumental variables literature adjusted to deal with the high dimensionality of the problem.

(R3) (Exogeneity): $\mathbf{W}_{i\cdot}$ is independent of the error terms in (4) and (5).

(R4) (Normality): The error terms in (4) and (5) follow a bivariate normal distribution.

(R5) (Global IV Strength): The IVs are globally strong with $\|\boldsymbol{\gamma}^*_{\mathcal{V}^*}\|_2 = \sqrt{\sum_{j\in\mathcal{V}^*} \boldsymbol{\gamma}_j^2} \geq \delta \gg s_{z1} \log p/\sqrt{n}$, where $\mathcal{V}^*$ is the set of valid instruments defined in Definition 3.

Assumption (R3) is the usual instrumental variables assumption where we assume that the covariates and the IVs are exogenous (Wooldridge, 2010). Assumption (R4) states that the error terms (e_{i1}, ϵ_{i2}), are bivariate normals. Here, we take the normal errors assumption as a simple, first-step approach to analyzing the behaviors of our procedure. Finally, Assumption (R5) is essentially a more pointed statement about IV assumption (A1) where the global strength of instruments, measured by the ℓ_2 norm of $\boldsymbol{\gamma}^*$ among valid IVs $\mathcal{V}^*$, is bounded away from zero; if all the IVs are valid, then $\|\boldsymbol{\gamma}^*_{\mathcal{V}^*}\|_2 = \|\boldsymbol{\gamma}^*\|_2$. This assumption is satisfied by "strong instrument asymptotics" or "traditional asymptotics" in instrumental variables (Stock et al., 2002; Wooldridge, 2010) where we fix the number of instruments and let n go to infinity. This assumption is also satisfied so long as there is at least one IV that has a constant non-zero effect on the treatment, or a non-zero effect that doesn't diminish with sample size, which will hold in most empirical applications.

Overall, the regularity assumptions (R1)-(R5) are standard in the usual high dimensional and the instrumental variables literature. Consequently, if we only have valid instruments and use $\widehat{\beta}_0$ in (16) from our procedure, only assumptions (R1)-(R5) are needed for the consistency and asymptotic normality of $\widehat{\beta}_0$. This is formally stated in Theorem 1.

Theorem 1. *Suppose we have valid IVs, that is* $\boldsymbol{\pi}^* = 0$ *in* (2)*, and the assumptions* (R1) – (R5) *hold. The following property holds for the estimator* $\widehat{\beta}_0$*,*

$$\sqrt{n}\left(\widehat{\beta}_0 - \beta^*\right) = T^{\beta^*} + \Delta^{\beta^*}, \tag{18}$$

where $T^{\beta^*} \mid \mathbf{W} \sim N\left(0, \mathrm{V}_0\right)$, $\mathrm{V}_0 = 1/\|\boldsymbol{\gamma}^*\|_2^4 \times \left\| \sum_{j \in \mathcal{S}^*} \boldsymbol{\gamma}_j^* \frac{1}{\sqrt{n}} \mathbf{W} \widehat{\mathbf{u}}^{[j]} \right\|_2^2 (\boldsymbol{\Theta}_{11}^* + (\beta^*)^2 \boldsymbol{\Theta}_{22}^* - 2\beta^* \boldsymbol{\Theta}_{12}^*)$ *and* $\Delta^{\beta^*}/\sqrt{\mathrm{V}_0} \xrightarrow{p} 0$ *as* $\sqrt{s_{\mathrm{z}1}} s \log p/\sqrt{n} \to 0$*.*

Theorem 1 states that, when we only have valid IVs, $\widehat{\beta}_0$ defined in (16) is a consistent estimator of the treatment effect and the dominating part of the scaled difference $\sqrt{n}(\widehat{\beta}_0 - \beta)$ is asymptotically normal. Based on the asymptotic normality established in (18), the following theorem justifies the coverage property of the confidence interval proposed in (17) under the assumption that all instruments are valid.

Theorem 2. *Suppose we have valid IVs, that is* $\boldsymbol{\pi}^* = 0$ *in* (2) *and the assumptions* (R1) – (R5) *hold. Assuming* $\sqrt{s_{\mathrm{z}1}} s \log p/\sqrt{n} \to 0$*, the confidence interval given in* (17) *has asymptotically coverage probability* $1 - \alpha$*, i.e.,*

$$\mathbf{P}\left\{ \beta \in \left(\widehat{\beta}_0 - (1+\eta_0) z_{1-\alpha/2} \sqrt{\widehat{\mathrm{V}}_0/n}, \quad \widehat{\beta}_0 + (1+\eta_0) z_{1-\alpha/2} \sqrt{\widehat{\mathrm{V}}_0/n} \right) \right\} \to 1 - \alpha, \tag{19}$$

for any given positive constant $\eta_0 > 0$*.*

Theorem 2 is similar to a result given in Chernozhukov et al. (2015), who studied IV estimators in high dimensional regime where all the instruments are valid. However, there are some notable differences between our results and those in Chernozhukov et al. (2015) where different sparsity and instrument-covariate modeling assumptions are required. A simulation study is carried out in Section 5 to compare our procedure to that of the oracle.

To analyze our procedure in the case of invalid instruments, we make three additional assumptions that are not part of the usual instrumental variables literature and may be of interest for future theoretical work. We label these assumptions as "IN" to denote that they are specific to the case of invalid IVs.

(IN1) (50% Rule) The number of valid IVs is more than half of the number of individually strong IVs, that is $|\mathcal{V}^*| > \frac{1}{2}|\mathcal{S}^*|$.

(IN2) (Individual IV Strength) Among IVs in $\mathcal{S}^*$, we have $\min_{j\in\mathcal{S}^*}\left|\boldsymbol{\gamma}_j^*\right| \geq \delta_{\min} \gg \sqrt{\log p/n}$.

(IN3) (Strong violation) Among IVs in the set $\mathcal{S}^* \setminus \mathcal{V}^*$, we have

$$\min_{j\in\mathcal{S}^*\setminus\mathcal{V}^*}\left|\frac{\boldsymbol{\pi}_j^*}{\boldsymbol{\gamma}_j^*}\right| \geq \frac{12(1+|\beta^*|)}{\delta_{\min}}\sqrt{\frac{M_1\log p_{\mathrm{z}}}{n\lambda_{\min}(\boldsymbol{\Theta}^*)}}. \tag{20}$$

Assumption (IN1) is the generalization of the 50% rule in Kang et al. (2015) and Han (2008) to deal with high dimensional and redundant IVs. In a nutshell, (IN1) states that the number of invalid instruments is not too large so that we can detect the invalid IVs from valid IVs, without knowing a priori which IVs are valid or invalid; see Kang et al. (2015) for details. Assumption (IN2) requires individual IV strength to be bounded away from zero by $\delta_{\min}$ so that all the IVs selected in $\widetilde{\mathcal{S}}$ are strong. This contrasts with the valid IV case, specifically Assumption (R5), where (R5) only required the global IV strength to be bounded away from zero. Finally, (IN3) requires IVs that violate IV Assumptions (A2) and/or (A3), specifically $\boldsymbol{\pi}_j^*/\boldsymbol{\gamma}_j^*$, to be large so that we can correctly identify IVs that violate (A2) and (A3). Overall, the collection of assumptions (IN1)-(IN3) allows detection of each valid IVs beyond the noise level of the data. In particular, (IN1)-(IN3) assures that STEP 1b of Procedure 1 has a high probability of correctly selecting the valid instruments. If all the instruments are valid, we avoid the STEP 1b of Procedure 1 and thus, we do not need Assumptions (IN1), (IN2) and (IN3).

With the above assumptions, Theorem 3 states that our general Procedure 1 produces a consistent and asymptotic normal estimate of β^*.

Theorem 3. *Suppose the assumptions* (R1)−(R5) *and* (IN1)−(IN2) *hold. As* $\sqrt{s_{\mathrm{z1}}}s\log p/\sqrt{n}\to 0$*, with probability larger than* $1-c\left(p^{-c}+\exp\left(-cn\right)\right)$*,*

$$\left|\widehat{\beta}-\beta^*\right| \leq C\frac{1}{\delta_{\min}}\sqrt{\frac{\log p_{\mathrm{z}}}{n}}, \tag{21}$$

where $c, C>0$ *are constants independent of* n *and* p*. In addition, if* (IN3) *hold, we have*

$$\sqrt{n}\left(\widehat{\beta}-\beta^*\right) = T^{\beta^*}+\Delta^{\beta^*} \tag{22}$$

where $T^{\beta^*} \mid \mathbf{W}\sim N\left(0,\mathrm{V}\right)$*,* $\mathrm{V}=1/\|\boldsymbol{\gamma}^*\|_2^4\times\left\|\sum_{j\in\mathcal{V}^*}\gamma_j^*\frac{1}{\sqrt{n}}\mathbf{W}\widehat{\mathbf{u}}^{[j]}\right\|_2^2(\boldsymbol{\Theta}_{11}^*+(\beta^*)^2\boldsymbol{\Theta}_{22}^*-2\beta^*\boldsymbol{\Theta}_{12}^*)$*, and* $\Delta^{\beta^*}/\sqrt{\mathrm{V}}\xrightarrow{p}0$ *as* $\sqrt{s_{\mathrm{z1}}}s\log p/\sqrt{n}\to 0$*.*

Theorem 3 shows that the proposed estimator $\widehat{\beta}$ is a consistent estimator of the treatment effect and the asymptotic normality holds in the presence of invalid instruments. The consistency of our estimator, specifically (21) in Theorem 3, is established under relatively weaker assumptions. For asymptotic normality, Theorem 3 requires Assumption (IN3), which rules out instruments that weakly violate IV assumptions (A2) and (A3). In Section 3.2 of the Supplementary Materials, we explore the nature of (IN3), specifically whether (IN3) is necessary for honest coverage, by varying the degree of instruments that violate the IV assumptions (A2) and (A3), some of which violate (IN3). Also, the asymptotic normality from Theorem 3 allows us to show that our confidence interval (15) has correct coverage even with invalid instruments and high dimensional covariates.

Theorem 4. *Suppose the assumptions* (R1)−(R5) *and* (IN1)−(IN3) *hold. As* $\sqrt{s_{\mathrm{z1}}}s\log p/\sqrt{n} \to 0$*, the confidence interval given in* (15) *has asymptotically coverage probability* $1-\alpha$*, i.e.,*

$$\mathbf{P}\left\{\beta^* \in \left(\widehat{\beta} - (1+\eta_0)z_{1-\alpha/2}\sqrt{\widehat{\mathrm{V}}/n}, \quad \widehat{\beta} + (1+\eta_0)z_{1-\alpha/2}\sqrt{\widehat{\mathrm{V}}/n}\right)\right\} \to 1-\alpha, \tag{23}$$

for any given positive constant $\eta_0 > 0$*.*

We note that our confidence interval from the general procedure works well even if all the instruments turn out to be valid, but we do not a priori make this assumption. Specifically, in Section 5, we find that the confidence interval proposed in (15) has the correct coverage and similar average length to the oracle even when the instruments turn out to be valid.

5 Simulation

5.1 Setup

In addition to the theoretical analysis of our method in Section 4, we also conduct a simulation study to investigate the performance of our method and other comparators. The data generating process for the simulation follows the models (2) and (3) in Section 2.2 with $p_{\mathrm{z}} = 100$ instruments and $p_{\mathrm{x}} = 150$ covariates where $\mathbf{W}_{i\cdot}$ is a multivariate normal with mean zero and covariance $\mathbf{\Sigma}^*_{ij} = 0.5^{|i-j|}$ for $1 \le i, j \le 250$. The parameters for the models are: $\beta^* = 1$, $\phi^* = (0.6, 0.7, 0.8, \cdots, 1.5, 0, 0, \cdots, 0) \in \mathbb{R}^{150}$ so that $s_{\mathrm{x1}} = 10$,

$\psi^* = (1.1, 1.2, 1.3, \cdots, 2.0, 0, 0, \cdots, 0) \in \mathbb{R}^{150}$ so that $s_{\text{x2}} = 10$, and variance-covariance of the error terms are $Var(\epsilon_{i1}) = Var(\epsilon_{i2}) = 1.5$, and $Cov(\epsilon_{i1}, \epsilon_{i2}) = 0.75$. Instruments that satisfy Assumption (A1) are $\mathcal{S}^* = \{1, \ldots, 7\}$ and instruments that satisfy all three IV assumptions (A1)-(A3) are $\mathcal{V}^* = \{1, 2, 3, 4, 5\}$; thus instruments 6 and 7 only satisfy (A1), but do not satisfy (A2) and (A3). We fix these values throughout the entire simulation study.

The parameters we vary in the simulation study are: the sample size n, the strength of IVs via $\boldsymbol{\gamma}^*$, and violations of (A2) and (A3) via $\boldsymbol{\pi}^*$. For sample size, we let $n = (100, 200, 300, 1000, 3000)$. For IV strength, we set $\boldsymbol{\gamma}^*_{\mathcal{V}^*} = K(1, 1, 1, 1, \rho_1)$ and $\boldsymbol{\gamma}^*_{\mathcal{S}^* \setminus \mathcal{V}^*} = K(1, 1)$ and $\boldsymbol{\gamma}_{(\mathcal{S}^*)^C} = \mathbf{0}$, where we vary K (to be discussed later) and $\rho_1 = (0, 0.1, 0.2)$ across simulations. The value K controls the global strength of instruments, with higher $|K|$ indicating strong instruments in a global sense. The value ρ_1 controls the relative individual strength of instrument, specifically between the first four instruments in $\mathcal{V}^*$ and the fifth instrument. For example, $\rho_1 = 0.2$ implies that the fifth IV's individual strength is only 20% of the other four valid instruments, i.e IVs 1 to 4.

To specify K across simulations, we introduce a quantity we call the oracle concentration parameter (OCP) denoted as $C(\boldsymbol{\gamma}, \mathcal{V}^*, n)$.

$$C(\boldsymbol{\gamma}^*, \mathcal{V}^*, n) = \frac{n\boldsymbol{\gamma}^{*\intercal}_{\mathcal{V}^*} \left(\boldsymbol{\Sigma}^*_{\mathcal{V}^*\mathcal{V}^*} - \boldsymbol{\Sigma}^*_{\mathcal{V}^*(\mathcal{V}^*)^C} \boldsymbol{\Sigma}^{*-1}_{(\mathcal{V}^*)^C(\mathcal{V}^*)^C} \boldsymbol{\Sigma}^*_{(\mathcal{V}^*)^C\mathcal{V}^*} \right) \boldsymbol{\gamma}^*_{\mathcal{V}^*}}{|\mathcal{V}^*| \boldsymbol{\Theta}^*_{22}}, \tag{24}$$

where $\boldsymbol{\Sigma}^*_{I,J}$ denotes the submatrix containing $\boldsymbol{\Sigma}^*_{ij}$ for $i \in I$ and $j \in J$ and $\boldsymbol{\gamma}^*_{\mathcal{V}^*}$ denotes the subvector containing $\boldsymbol{\gamma}^*_j$ for $j \in \mathcal{V}^*$. In Section 3.1 of the Supplementary Materials, we discuss the OCP and its relation to the usual concentration parameter in the IV literature Stock and Wright (2000). In short, we define the OCP because the usual concentration parameter can be misleading when there are unknown invalid instruments and the OCP serves as a proxy for the usual concentration parameter.

Having defined the OCP, we can specify K as a function of n and $C(\boldsymbol{\gamma}^*, \mathcal{V}^*, n)$. Specifically, if n is set at a baseline of 100 and the simulation parameters $\mathcal{V}^*$, ρ_1, $\boldsymbol{\Sigma}^*$ and $\boldsymbol{\Theta}^*_{22}$ are specified as above, we can find K for a particular value of the expected oracle concentration parameter $C(\boldsymbol{\gamma}^*, \mathcal{V}^*, 100)$. Thus, by varying $C(\boldsymbol{\gamma}^*, \mathcal{V}^*, 100) = (50, 100, 150, 200, 250, 500, 1000)$, we vary K.

Finally, we vary $\boldsymbol{\pi}^*$, which controls the validity of the IVs by defining $\boldsymbol{\pi}^*_j = \rho_2 \boldsymbol{\gamma}^*_j$ for

$j = 6, 7$ and $\boldsymbol{\pi}_j^* = 0$ for all other j so that ρ_2 controls the magnitude of the violation of IV assumptions (A2) and (A3) from the 6th and 7th instruments. In the ideal case, we would have $\rho_2 = 0$ so that $\mathcal{S}^* = \mathcal{V}^* = \{1, 2, 3, 4, 5, 6, 7\}$. But, $\rho_2 \neq 0$ implies that the last two instruments do not satisfy (A2) and (A3). As such, we vary $\boldsymbol{\pi}^*$ by varying $\rho_2 = (0, 1, 2)$.

In summary, we vary n, the strength of IVs via $\boldsymbol{\gamma}^*$, and violations of (A2) and (A3) via $\boldsymbol{\pi}^*$ in our simulation study and our simulation are repeated 500 times. For high dimensional simulation settings $n \ll p$, we compare our procedure to $\widehat{\beta}_0$, which assumes IVs are valid. For low dimensional settings where $n \gg p$, we add two additional comparators, the two-stage least squares (2SLS) and OLS. 2SLS is the most popular IV method where one regresses $\mathbf{D}$ on $\mathbf{Z}$ and $\mathbf{X}$, and uses the predicted value of $\mathbf{D}$ in the regression of $\mathbf{Y}$ on $\mathbf{X}$ and $\mathbf{D}$. Note that the way we implement 2SLS mimics most practitioners' use of 2SLS by simply assuming all the instruments $\mathbf{Z}$ are valid. OLS is defined as where one regresses $\mathbf{Y}$ on $\mathbf{D}$ and $\mathbf{X}$. OLS will be biased because of confounding on $\mathbf{D}$. Finally, for both low and high dimensional settings, we have the oracle 2SLS where an oracle provides us with the true set of valid IVs, which will not occur in practice.

5.2 Results

We present the most representative results from our simulation study; all the simulations are in Section 3.2 of the Supplementary Materials. First, Figure 1 considers the high dimensional setting with $n = 200$ and three comparators, our procedure $\widehat{\beta}$ that's robust to invalid IVs, our procedure $\widehat{\beta}_0$ that assumes all valid IVs, and the oracle 2SLS. Columns "Weak" and "Strong" in the figure represent cases where $\rho_1 = 0.2$ and $\rho_1 = 0$, respectively. Columns "Valid" and "Invalid" represent cases where $\rho_2 = 0$ and $\rho_2 = 2$, respectively. The row "MAE" in the figure represents the median absolute error of the estimators, which measures the performance of the point estimators. The row "Coverage" represents the coverage performance of the confidence intervals. Finally, the row "Length" represents the average length of confidence intervals across simulations.

Both estimators $\widehat{\beta}$ and $\widehat{\beta}_0$ perform well in terms of estimation accuracy, coverage and length of confidence intervals and have similar performance to the benchmark, $\widehat{\beta}_{\text{oracle}}$ when

all the instruments are valid (i.e. first and second columns of Figure 1). For example, in the MAE and length plots, the solid lines, which represent our estimator, the dashed lines, which represent the our estimator assuming all valid IVs, and the dotted lines, which represent the oracle, overlap with each other. However, if the instruments are invalid (i.e. the third and fourth columns of Figure 1), $\widehat{\beta}_0$ is not consistent and loses coverage, which makes sense since $\widehat{\beta}_0$ assumes all the IVs are valid. However, our proposed estimator $\widehat{\beta}$ allows for possibly invalid instruments and performs as well as the oracle in terms of estimation accuracy and coverage. The average length of our robust confidence interval is only slightly larger than that of the oracle.

Figure 2 represents the same setting as Figure 1 except we now consider the low dimensional setting with $n = 1000$. As expected, the estimators $\widehat{\beta}$ and $\widehat{\beta}_0$ along with the traditional 2SLS estimator perform similarly to the oracle benchmark in terms of estimation accuracy, coverage and the length of confidence intervals when all the instruments are actually valid. For example, in the MAE plot of Figure 2, the solid, dashed, green and dotted lines, representing $\widehat{\beta}$, $\widehat{\beta}_0$, 2SLS and the oracle, respectively, overlap with each other. Note that OLS cannot deal with confounding and hence, produces a biased estimate. However, when the instruments are invalid, the traditional 2SLS estimator and $\widehat{\beta}_0$ are biased and fail to have the correct coverage. In contrast, the proposed estimator $\widehat{\beta}$ performs as well as the oracle estimator in terms of estimation accuracy and coverage, with the length of the proposed estimator being slightly longer than that for the oracle. We should also stress that throughout both figures, our proposed estimator $\widehat{\beta}$ performs similarly for individually weak and strong IVs (i.e. for different ρ_1s). This provides empirical support that the proposed estimator is not too sensitive to the individual instrument strength assumption (IN2) required in the theoretical analysis.

6 Data Analysis

To demonstrate our procedure 1 in real settings, we analyze the causal effect of years of education on yearly earnings, which has been studied extensively in economics using IV methods (Angrist and Krueger, 1991; Card, 1993, 1999). The data comes from the Wis-

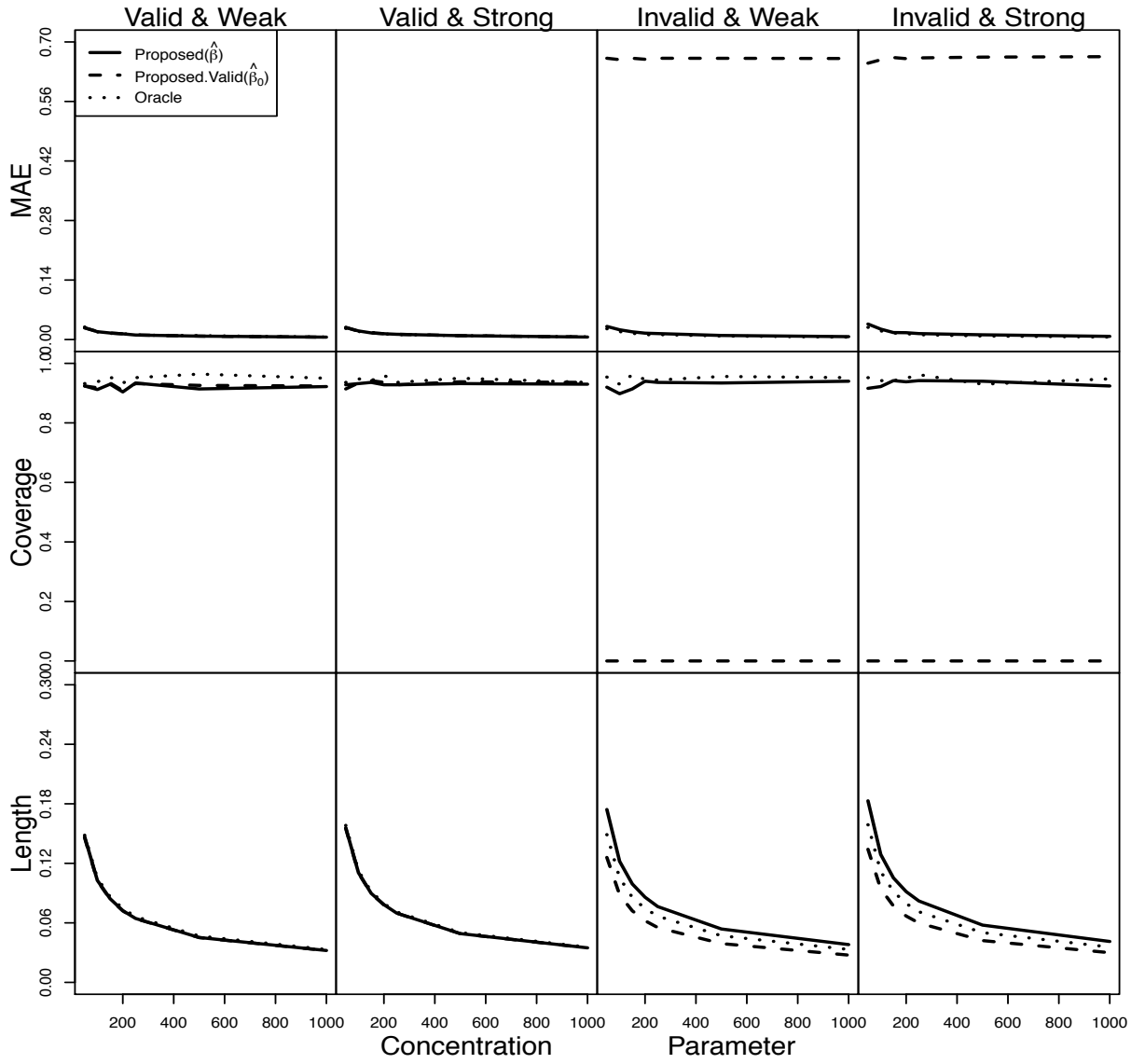

Figure 1: Comparison of different methods when $n = 200$. The x-axis represents the concentration parameter. On the y-axis, MAE represents Median Absolute Error of the estimators, Coverage represents coverage of the confidence intervals and Length represents the average length of confidence intervals. Proposed is our method allowing for invalid IVs and is represented by the solid line. Proposed.valid is our method that assumes all the IVs are valid and is represented by the dashed line. Oracle is the method that knows exactly which instruments are valid and is represented by the dotted line. The column labeled with Valid & Weak represents the case $\rho_1 = 0.2$ and $\rho_2 = 0$. The column labeled with Valid &Strong represents the case $\rho_1 = 0$ and $\rho_2 = 0$. The column labeled with Invalid &Weak represents the case $\rho_1 = 0.2$ and $\rho_2 = 2$. Finally, the column labeled with Invalid & Strong represents the case $\rho_1 = 0$ and $\rho_2 = 2$.

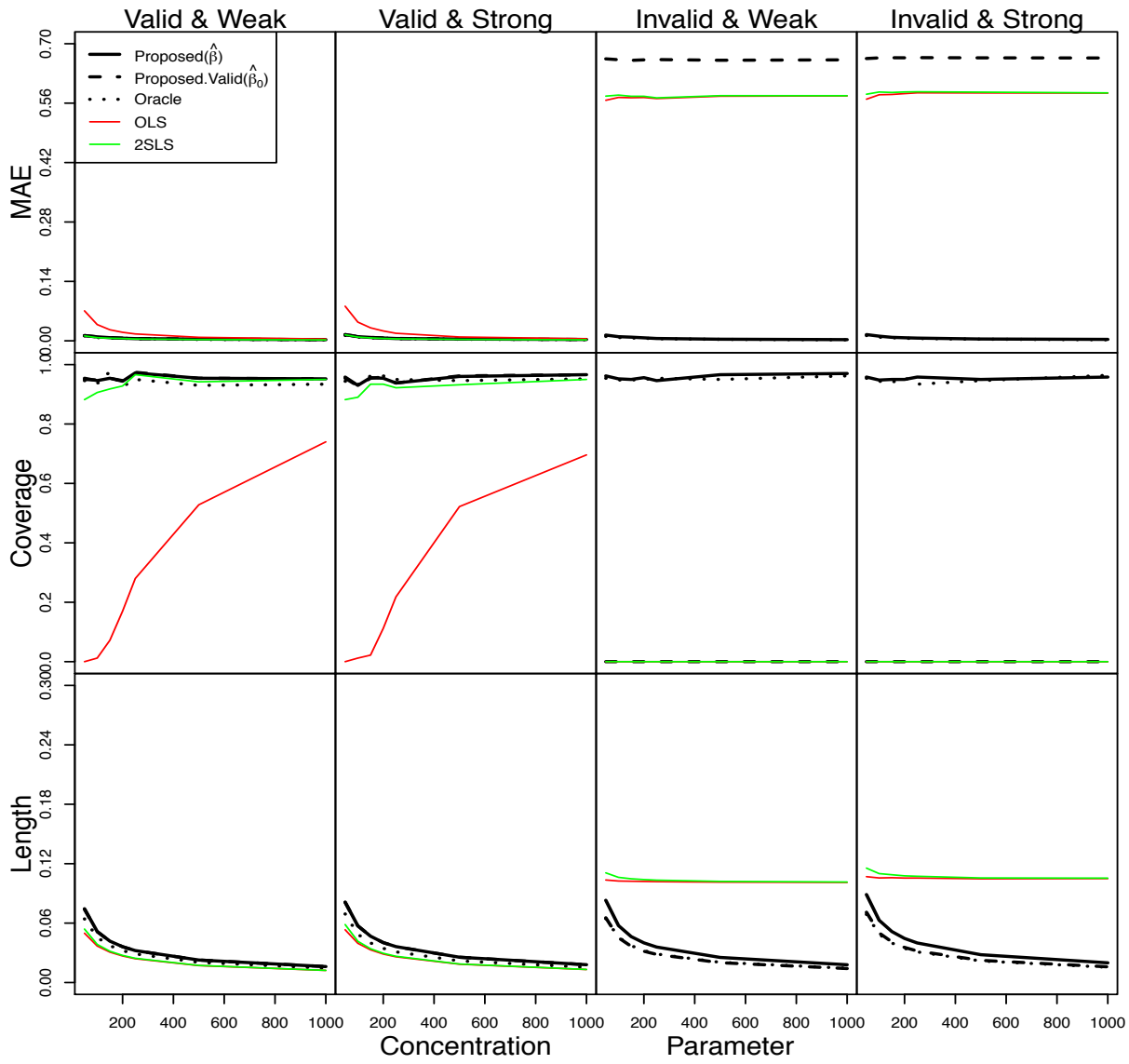

Figure 2: Comparison of different methods when $n = 1000$. The x-axis represents the concentration parameter. On the y-axis, MAE represents Median Absolute Error of the estimators, Coverage represents coverage of confidence intervals and Length represents the average length of confidence intervals. Proposed is our method allowing for invalid IVs and is represented by the solid line. Proposed.valid is our method that assumes all the IVs are valid and is represented by the dashed line. Oracle is the method that knows exactly which instruments are valid and is represented by the dotted line. The column labeled with Valid & Weak represents the case $\rho_1 = 0.2$ and $\rho_2 = 0$. The column labeled with Valid &Strong represents the case $\rho_1 = 0$ and $\rho_2 = 0$. The column labeled with Invalid &Weak represents the case $\rho_1 = 0.2$ and $\rho_2 = 2$. Finally, the column labeled with Invalid & Strong represents the case $\rho_1 = 0$ and $\rho_2 = 2$.

consin Longitudinal Study (WLS), a longitudinal study that has kept track of American high school graduates from Wisconsin since 1957, and we examine the relationship between graduates' earnings and education from the 1974 survey Hauser (2005), roughly 20 years after they graduated from high school. Our analysis includes $N = 3772$ individuals, 1784 males and 1988 females. For our outcome, we use imputed log total yearly earnings and for the treatment, we use the total years of education, all from the 1974 survey. The median total earnings is $\$9,200$ with a 25% quartile of $\$1,000$ and a 75% quartile of $\$15,320$ in 1974 dollars. The mean years of total education is 13.7 years with a standard deviation of 2.3 years.

We incorporate many covariates, including sex, graduate's hometown population, educational attainment of graduates' parents, graduates' family income, relative income in graduates' hometown, graduates' high school denomination, high school class size, all measured in 1957 when the participants were high school seniors. We also include 81 genetic covariates, specifically single nucleotide polymorphisms (SNPs), that were part of WLS to further control for potential variations between graduates; see Section 2 in Supplementary Materials for details on the non-genetic and genetic covariates. In summary, our data analysis includes 7 non-genetic covariates and 81 genetic covariates.

We used five instruments in our analysis, all derived from past studies of education on earnings (Blundell et al., 2005; Card, 1993; Gary-Bobo et al., 2006). They are (i) total number of sisters, (ii) total number of brothers, (iii) individual's birth order in the family, all from Gary-Bobo et al. (2006), (iv) proximity to college from Card (1993), and (v) teacher's interest in individual's college education from Blundell et al. (2005), all measured in 1957. Although all these IVs have been suggested to be valid with varying explanations as to why they satisfy (A2) and (A3) after controlling for the aforementioned covariates, in practice, we are always uncertain due to the lack of complete socioeconomic knowledge about the effect of these IVs. Our method should provide some protection against this uncertainty compared to traditional methods where they simply assume that all five IVs are valid. Also, the first-stage F-test produces an F-statistic of 90.3 with a p-value less than 10^{-16}, which indicates very strong set of instruments. For more details on the instruments, see Section 2 of the Supplementary Materials.

Table 1 summarizes the results of our data analysis. OLS refers to running a regression of the treatment and the covariates on the outcome and looking at the slope coefficient of the treatment variable. 2SLS refers to running two-stage least squares as described in Section 5 under the operating assumption that all the five instruments are valid; this is the usual and most popular analysis in the IV literature. Finally, we run the Pocedure 1.

Method	Point Estimate	95% Confidence Interval
OLS	0.097	(0.051, 0.143)
2SLS	0.169	(0.029, 0.301)
Our Method	0.062	(0.046, 0.077)

Table 1: Estimates of the Effect of Years of Education on Log Earnings. OLS is ordinary least squares, 2SLS is two-stage least squares, and our method is Procedure 1.

The OLS estimate suggests a positive association between education and earnings, with statistically significant result at $\alpha = 0.05$ level. This agrees with previous literature which suggests a statistically significant positive association between years of education and log earnings Card (1999). However, OLS does not completely control for confounding even after controlling for covariates. 2SLS provides an alternative method of controlling for confounding by using instruments so long as all the instruments satisfy the three core assumptions and the inclusion of covariates helps make these assumptions more plausible. Unfortunately, we notice that the 2SLS estimate in Table 1 is inconsistent with previous studies' estimates among individuals from the U.S. between 1950s to 1970s, which range from 0.06 to 0.13 (see Table 4 in Card (1999)). Our method, which addresses the concern for invalid instruments with 2SLS, provides an estimate of 0.062, which is more consistent with previous studies' estimates of the effect of years of education on earnings.

The data analysis suggests that our method can be a useful tool in IV analysis when there is concern for invalid instruments, even after attempting to mitigate this problem via covariates. Our method provides much more accurate estimates of the returns on education than 2SLS, which naively assumes all the instruments are valid.

7 Discussion

Inspired by large observational data, we present a method to estimate the effect of the treatment on the outcome using instrumental variables where we do not make the assumption that all the instruments are valid. Our approach differs from many instrumental variables approaches in that it allows for possibly invalid instruments even after conditioning on many, possibly high dimensional, covariates. Our approach provides robust confidence intervals in the presence of invalid IVs even after controlling for many covariates. In simulation and in real data settings, our approach provides a more robust analysis than the traditional IV approaches, most notably 2SLS, by providing some protection against possibly invalid instruments.

SUPPLEMENTARY MATERIAL

Title: Supplement to "Confidence Interval for Causal Effects with Possibly Invalid Instruments Even After Controlling for Many Confounders". (.pdf file)

References

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

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

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

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.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Richard Blundell, Lorraine Dearden, and Barbara Sianesi. Evaluating the effect of education on earnings: models, methods and results from the national child development survey. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(3):473–512, 2005.

Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer, 2011.

T Tony Cai and Zijian Guo. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of statistics*, To appear, 2016.

David Card. Using geographic variation in college proximity to estimate the return to schooling. Working Paper 4483, National Bureau of Economic Research, October 1993.

David Card. Chapter 30 - the causal effect of education on earnings. In Orley C. Ashenfelter and David Card, editors, *Handbook of Labor Economics*, volume 3, Part A, pages 1801 – 1863. Elsevier, 1999.

Xu Cheng and Zhipeng Liao. Select the valid and relevant moments: An information-based lasso for gmm with many moments. *Journal of Econometrics*, 186(2):443–464, 2015.

Victor Chernozhukov, Christian Hansen, and Martin Spindler. Post-selection and post-regularization inference in linear models with many controls and instruments. 2015.

Jianqing Fan and Yuan Liao. Endogeneity in high dimensions. *Annals of statistics*, 42(3): 872, 2014.

Robert Gary-Bobo, Nathalie Picard, and Ana Prieto. Birth order and sibship sex composition as instruments in the study of education and earnings. CEPR Discussion Papers 5514, C.E.P.R. Discussion Papers, 2006.

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

Chirok Han. Detecting invalid instruments using l 1-gmm. *Economics Letters*, 101(3):285–287, 2008.

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

Pamela Herd, Deborah Carr, and Carol Roan. Cohort profile: Wisconsin longitudinal study (wls). *International Journal of Epidemiology*, 43(1):34–41, 2014.

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

Paul W. Holland. Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology*, 18(1):449–484, 1988.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

Hyunseung Kang, Anru Zhang, T Tony Cai, and Dylan S Small. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, To appear, 2015.

Michal Kolesár, Raj Chetty, John N. Friedman, Edward L. Glaeser, and Guido W. Imbens. Identification and inference with many invalid instruments. *Journal of Business & Economic Statistics*, 33(4):474–484, 2015.

Jerzy Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472, 1923.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Dylan S. Small. Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association*, 102(479):1049–1058, 2007.

James H Stock and Jonathan H Wright. Gmm with weak identification. *Econometrica*, pages 1055–1096, 2000.

James H. Stock, Jonathan H. Wright, and Motohiro Yogo. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20(4), 2002.

Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 101(2):269–284, 2012.

Sara van de Geer, Peter Bühlmann, Yaacov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina Eldar and Gitta Kutyniok, editors, *Compressed Sensing: Theory and Applications*, pages 210–268. Cambridge University Press, 2012.

Michael W. Weiner, Dallas P. Veitch, Paul S. Aisen, Laurel A. Beckett, Nigel J. Cairns, Jesse Cedarbaum, Michael C. Donohue, Robert C. Green, Danielle Harvey, Clifford R. Jack Jr., William Jagust, John C. Morris, Ronald C. Petersen, Andrew J. Saykin, Leslie Shaw, Paul M. Thompson, Arthur W. Toga, and John Q. Trojanowski. Impact of the alzheimer's disease neuroimaging initiative, 2004 to 2014. *Alzheimer's & Dementia*, 11(7):865 – 884, 2015. ISSN 1552-5260.

Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT press, 2nd ed. edition, 2010.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.