ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2018)

Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting

Zijian Guo,

Rutgers University, Piscataway, USA

Hyunseung Kang

University of Wisconsin—Madison, Madison, USA

and T. Tony Cai and Dylan S. Small

University of Pennsylvania, Philadelphia, USA

[Received August 2016. Final revision March 2018]

Summary. A major challenge in instrumental variable (IV) analysis is to find instruments that are valid, or have no direct effect on the outcome and are ignorable. Typically one is unsure whether all of the putative IVs are in fact valid. We propose a general inference procedure in the presence of invalid IVs, called two-stage hard thresholding with voting. The procedure uses two hard thresholding steps to select strong instruments and to generate candidate sets of valid IVs. Voting takes the candidate sets and uses majority and plurality rules to determine the true set of valid IVs. In low dimensions with invalid instruments, our proposal correctly selects valid IVs, consistently estimates the causal effect, produces valid confidence intervals for the causal effect and has oracle optimal width, even if the so-called 50% rule or the majority rule is violated. In high dimensions, we establish nearly identical results without oracle optimality. In simulations, our proposal outperforms traditional and recent methods in the invalid IV literature. We also apply our method to reanalyse the causal effect of education on earnings.

Keywords: Exclusion restriction; High dimensional covariates; Invalid instruments; Majority voting; Plurality voting; Treatment effect

1. Introduction

1.1. Motivation: invalid instruments

Instrumental variable (IV) analysis is a popular method to deduce causal effects in the presence of unmeasured confounding. Informally, an IV analysis requires instruments that

(a) are associated with the exposure (assumption 1),
(b) have no direct pathway to the outcome (assumption 2) and
(c) are not related to unmeasured variables that affect the exposure and the outcome (assumption 3); see Section 2.1 for details.

A major challenge in IV analysis is to find valid instruments, i.e. instruments that satisfy assumptions 2 and 3.

Address for correspondence: Dylan S. Small, Department of Statistics, University of Pennsylvania, Wharton School, 400 Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104, USA.
E-mail: dsmall@wharton.upenn.edu

 1369–7412/18/80000

For example, a long-standing interest in economics is studying the causal effect of education on earnings (Angrist and Krueger, 1991; Card, 1993, 1999). Often, IV analysis is used to deduce this effect and a popular instrument for the analysis is a person's proximity to a college when growing up (Card, 1993, 1999). However, proximity to a college may be related to a person's socio-economic status, high school characteristics and other variables that may affect a person's earnings, thereby invalidating the instrument. Often, covariates, potentially many, are controlled for to make instruments more plausibly valid (Hernán and Robins, 2006; Swanson and Hernán, 2013; Baiocchi *et al.*, 2014; Imbens, 2014). For instance, in the earnings example, socio-economic status, family background and genetic status can be controlled for to alleviate concerns about instrument validity.

But, some IVs may still turn out to be invalid after controlling and subsequent analysis assuming that all the IVs are valid after conditioning can be misleading (Murray, 2006). For instance, in the earnings example with proximity as the instrument, if living close to college had other benefits beyond receiving more education, say by being exposed to many programmes that are 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 can directly affect an individual's earning potential and violate assumption 2 (Card, 1999). This problem is also prevalent in other applications of IVs, most notably in Mendelian randomization (Davey Smith and Ebrahim, 2003, 2004) where the instruments are genetic in nature and some instruments are likely to be invalid because they have pleiotropic effects (Lawlor *et al.*, 2008; Burgess *et al.*, 2015).

This paper tackles the problem of constructing confidence intervals for causal effects when invalid instruments may be present. We consider two major cases. The first case is where the number of covariates and instruments is small and fixed relative to the sample size; this setting is typical in Mendelian randomization studies and many traditional applied settings. The second case is where the number of covariates and/or instruments is growing and may exceed the sample size, which is becoming more prevalent with modern large data sets.

1.2. Prior work and our contributions

In non-IV settings with high dimensional covariates as controls, Zhang and Zhang (2014), Javanmard and Montanari (2014), van de Geer *et al.* (2014), Belloni *et al.* (2014) and Cai and Guo (2017) have provided confidence intervals for a treatment effect. In IV settings with high dimensional covariates (or IVs), Gautier and Tsybakov (2011), Belloni *et al.* (2012), Fan and Liao (2014) and Chernozhukov *et al.* (2015) have provided confidence intervals for a treatment effect, under the assumption that all the IVs are valid after controlling for the said covariates. In invalid IV settings, Kolesár *et al.* (2015) and Bowden *et al.* (2015) have provided inferential methods for treatment effects. However, the method requires that the effects of the instruments on the treatment are orthogonal to their direct effects on the outcome, which is a stringent assumption. Bowden *et al.* (2016), Burgess *et al.* (2016), Kang *et al.* (2016b) and Windmeijer *et al.* (2016) also worked on the invalid IV setting, but without making a stringent orthogonality assumption.

Unfortunately, all of these methods

(a) work only in the low dimensional setting and
(b) rely on the sufficient condition in Han (2008) and Kang *et al.* (2016b),

which is known as the '50% rule' or the 'majority rule' where a majority of the instruments must be valid to establish consistency or inferential guarantees (see Section 2.2 for details); to the best of our knowledge, no method in this literature has established consistency, inferential and oracle

guarantees under a general condition established in theorem 1 of Kang *et al.* (2016b), including the setting where the majority rule is violated; see Section 3.6 for a review and a comparison of the methods in the literature with our proposed method.

Our work makes three major contributions in inferring treatment effects in the presence of possibly invalid instruments. First, we propose a novel two-stage hard thresholding (TSHT) method with voting that works both in low and high dimensional settings. Second, in the low dimensional setting, our method is the first method to be complete; our method relies only on a general condition for identification under invalid instruments to

(a) to select valid IVs correctly,
(b) to estimate the causal effect consistently,
(c) to produce confidence intervals with the desired level of coverage and
(d) to achieve oracle optimality in the sense that it performs as well asymptotically as the oracle procedure that knows which instruments are valid.

In particular, our method can guarantee these properties even when more than 50% of the instruments are invalid as long as a more general condition, which we call the plurality rule, holds; see Section 2.2 for details. Third, in the high dimensional setting, our method achieves the same selection, estimation and inferential guarantees without the oracle optimality.

The outline of the paper is as follows. After describing the model set-up in Section 2, we describe our procedure TSHT with voting in Section 3 and provide theoretical justification for it in Section 4. In Section 5, we investigate the performance of our procedure in a simulation study and compare it with existing methods, in particular the median method with bootstrapping of Bowden *et al.* (2016) and Burgess *et al.* (2016) and the adaptive lasso method of Windmeijer *et al.* (2016). We find that our method and that of Windmeijer *et al.* (2016) are comparable when the 50% rule holds, whereas the median estimator suffers from coverage and optimality issues. However, when the 50% rule fails, our method dominates all these methods. In Section 6, we present an empirical study where we revisit the question of the causal effect of years of schooling on income by using data from the Wisconsin longitudinal study (WLS). We provide conclusions and discussions in Section 7. The code to implement the proposed method along with a running example is available from `https://github.com/hyunseungkang/invalidIV`.

2. Model

To define causal effects and instruments, the potential outcome approach (Neyman, 1923; Rubin, 1974) that is 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 or treatment $d \in \mathbb{R}$ 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 or her observed candidate instrument values $\mathbf{Z}_{i\cdot} \in \mathbb{R}^{p_z}$ and exposure value D_i. We also denote baseline covariates for each individual i as $\mathbf{X}_{i\cdot} \in \mathbb{R}^{p_x}$. In total, n sets of outcome, exposure, instruments and baseline covariates, which are denoted as $(Y_i, D_i, \mathbf{Z}_{i\cdot}, \mathbf{X}_{i\cdot})$, are observed in an independent and identically distributed fashion.

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be an n-dimensional vector of observed outcomes, $\mathbf{D} = (D_1, \ldots, D_n)$ be an n-dimensional vector of observed exposures, $\mathbf{Z}$ be an $n \times p_z$ matrix of instruments, where row i consists of $\mathbf{Z}_{i\cdot}$, and $\mathbf{X}$ be an $n \times p_x$ matrix of covariates where row i consists of $\mathbf{X}_{i\cdot}$. Let $\mathbf{W}$ be an $n \times p = p_z + p_x$ matrix where $\mathbf{W}$ is the result of concatenating the matrices $\mathbf{Z}$ and $\mathbf{X}$ and let $\boldsymbol{\Sigma}^* = \mathbf{E}(\mathbf{W}_{i\cdot}\mathbf{W}_{i\cdot}^{\mathrm{T}})$ be the positive definite covariance matrix of the instrument–covariate matrix.

For any vector $\mathbf{v} \in \mathbb{R}^p$, let 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 $\mathrm{supp}(\mathbf{v}) \subseteq \{1, \ldots, p\}$ denote the support of the vector $\mathbf{v}$, $\mathrm{supp}(\mathbf{v}) = \{j : v_j \neq 0\}$, and $\|\mathbf{v}\|_0$ denote the size of the support of $\mathbf{v}$ or, equivalently, the number of non-zero elements in $\mathbf{v}$. For a set J, let $|J|$ denote its cardinality and J^{C} denote its complement. For an $n \times p$ 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 $\mathbf{M}_{\cdot j}$. For any sets $A \subseteq \{1, \ldots, n\}$ and $B \subseteq \{1, \ldots, p\}$, let $\mathbf{M}_{A,B}$ denote the submatrix that is formed by the rows specified by the set A and the columns specified by the set B. Let $\mathbf{M}^{\mathrm{T}}$ be the transpose of $\mathbf{M}$ and $\|\mathbf{M}\|_\infty$ represent the elementwise matrix sup-norm of matrix $\mathbf{M}$. For a symmetric matrix $\mathbf{M}$, let $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ denote the smallest and largest eigenvalues of $\mathbf{M}$ respectively.

For a sequence of random variables X_n, let $X_n \to^{\mathrm{p}} X$ and $X_n \to^{\mathrm{d}} X$ denote that X_n converges to X in probability and in distribution respectively. For any two sequences a_n and b_n, let $a_n \gg b_n$ denote that $\limsup_{n\to\infty} b_n/a_n = 0$; similarly, let $a_n \ll b_n$ denote $b_n \gg a_n$.

2.1. Models and instrumental variables assumptions

We consider the additive linear, constant effects model of Holland (1988) and extend it to allow for possibly invalid instruments as in Small (2007) and Kang *et al.* (2016b). For two possible values of the exposure d', d and instruments $\mathbf{z}'$ and $\mathbf{z}$, we assume the following potential outcomes model:

$$\begin{aligned} Y_i^{(d',\mathbf{z}')} - Y_i^{(d,\mathbf{z})} &= (\mathbf{z}' - \mathbf{z})^{\mathrm{T}}\boldsymbol{\kappa}^* + (d' - d)\beta^*, \\ \mathbf{E}(Y_i^{(0,\mathbf{0})} | \mathbf{Z}_{i\cdot}, \mathbf{X}_{i\cdot}) &= \mathbf{Z}_{i\cdot}^{\mathrm{T}}\boldsymbol{\eta}^* + \mathbf{X}_{i\cdot}^{\mathrm{T}}\boldsymbol{\phi}^* \end{aligned} \tag{1}$$

where $\boldsymbol{\kappa}^*$, β^*, $\boldsymbol{\eta}^*$ 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 effect of covariates on the baseline potential outcome $Y_i^{(0,\mathbf{0})}$. The parameter $\boldsymbol{\kappa}^*$ represents the violation of the no-direct-effect assumption between the instruments and the outcome. For example, if instruments have a causal effect on an unobserved confounder of the exposure–outcome relationship, this would lead to a direct effect on the outcome and be reflected in $\boldsymbol{\kappa}^*$. The parameter $\boldsymbol{\eta}^*$ represents the presence of unmeasured confounding between the instrument and the outcome. Finally, the model does not assume that the instruments are uncorrelated with each other.

Our model parameters $\boldsymbol{\kappa}^*$ and $\boldsymbol{\eta}^*$ encode a particular case of the definitions of the exclusion restriction (assumption 2) and no unmeasured confounding (assumption 3) in Angrist *et al.* (1996) where we assume an additive, linear and constant treatment effect β^*; see Holland (1988) and its appendix, and section 1.4 of Hernán and Robins (2006) for additional discussions about different formalizations of the IV assumptions 2 and 3. For example, the exclusion restriction (assumption 2) which is typically stated (Angrist *et al.*, 1996) as $Y_i^{(d,z)} = Y_i^{(d,z')}$ for all $z, z' \in \mathbb{R}$ implies that $\boldsymbol{\kappa}^* = 0$. Also, the assumption of no unmeasured confounding of the IV–outcome relationship 3, which is typically stated (Angrist *et al.*, 1996) as $Y_i^{(d,z)}$ and $D_i^{(z)}$ are independent of Z_i for all $d, z \in \mathbb{R}$, implies that $\boldsymbol{\eta}^* = 0$; we note that Angrist *et al.* (1996) considered the instrument to have a non-zero average causal effect on the exposure, and hence the potential outcome notation for the exposure $D_i^{(z)}$.

Let $\boldsymbol{\pi}^* = \boldsymbol{\kappa}^* + \boldsymbol{\eta}^*$, $e_i = Y_i^{(0,\mathbf{0})} - \mathbf{E}(Y_i^{(0,\mathbf{0})} | \mathbf{Z}_{i\cdot}, \mathbf{X}_{i\cdot})$ and $\mathrm{var}(e_i | \mathbf{Z}_{i\cdot}, \mathbf{X}_{i\cdot}) = \sigma^2$. When we combine equation (1) along with the definition of e_i, we have the following observed data model, which is also known as the underidentified single-equation linear model in econometrics (page 83 of Wooldridge (2010)):

$$\begin{aligned} Y_i &= \mathbf{Z}_{i.}^{\mathrm{T}}\boldsymbol{\pi}^* + D_i\beta^* + \mathbf{X}_{i.}^{\mathrm{T}}\boldsymbol{\phi}^* + e_i, \\ \mathbf{E}(e_i|\mathbf{Z}_{i.}, \mathbf{X}_{i.}) &= 0. \end{aligned} \tag{2}$$

The observed model is not a usual regression model because D_i might be correlated with e_i. In particular, the parameter β^* measures the causal effect of changing D on Y rather than an association. Also, the parameter $\boldsymbol{\pi}^*$ in model (2) combines two assumptions: the exclusion restriction (assumption 2) parameterized by $\boldsymbol{\kappa}^*$, and no unmeasured confounding (assumption 3) parameterized by $\boldsymbol{\eta}^*$; in econometrics, the two assumptions are often combined and referred to as instrument exogeneity (Holland, 1988; Imbens and Angrist, 1994; Angrist *et al.*, 1996; Wooldridge, 2010). If both assumptions are satisfied, $\boldsymbol{\kappa}^* = \boldsymbol{\eta}^* = 0$ so $\boldsymbol{\pi}^* = 0$, although the converse is not necessarily true. Nevertheless, under both assumptions, the instruments are said to be valid (Murray, 2006) and $\boldsymbol{\pi}^*$ can be used to define valid IVs and we formalize the definition of valid IVs as follows.

Definition 1. Suppose that we have p_z candidate instruments along with models (1) and (2). We say that instrument $j = 1, \ldots, p_z$ satisfies both assumptions 2 and 3, or is valid, if $\pi_j^* = 0$ and we denote $\mathcal{P}^*$ to be the set of valid instruments.

Definition 1 is closely related to definitions of valid instruments in the literature. If $p_z = 1$, our definition is identical to the definition of a valid instrument in Holland (1988) and is a special case of the definition in Angrist *et al.* (1996) where we assume a model. In particular, the exclusion restriction and no unmeasured confounding in Angrist *et al.* (1996) imply that $\boldsymbol{\phi}^* = \boldsymbol{\psi}^* = 0$ and, consequently, $\pi^* = 0$, which is the definition of a valid IV in definition 1; however, satisfying definition 1 only implies $\boldsymbol{\phi}^* = -\boldsymbol{\psi}^*$, not necessarily $\boldsymbol{\phi}^* = 0$ or $\boldsymbol{\psi}^* = 0$. If $p_z > 1$, our framework is a generalization of these two prior works. In Mendelian randomization, our definition is identical to the definition of a valid instrument using $\boldsymbol{\pi}^*$ (Bowden *et al.*, 2016; Burgess *et al.*, 2016). We note the validity of an instrument j in the context of the set of instruments $\{1, \ldots, p_z\}$ being considered; see Section 2.3 of Kang *et al.* (2016b) for details.

In addition to the model for the outcome, we assume a linear association or observational model between the exposure D_i, the instruments $\mathbf{Z}_{i.}$ and the covariates $\mathbf{X}_{i.}$:

$$\begin{aligned} D_i &= \mathbf{Z}_{i.}^{\mathrm{T}}\boldsymbol{\gamma}^* + \mathbf{X}_{i.}^{\mathrm{T}}\boldsymbol{\psi}^* + \epsilon_{i2}, \\ \mathbf{E}(\epsilon_{i2}|\mathbf{Z}_{i.}, \mathbf{X}_{i.}) &= 0. \end{aligned}$$

Each element γ_j^*, $j = 1, \ldots, L$, is the partial correlation between the jth instrument and D_i. The parameter $\boldsymbol{\psi}^*$ represents the association between the covariates and D_i. Also, unlike models (1) and (2), we do not need a causal model between D_i, $\mathbf{Z}_{i.}$ and $\mathbf{X}_{i.}$; this is because the constant effect assumption that we make in model (1) eliminates the need to assume a causal instrument; see Angrist *et al.* (1996) for details.

On the basis of model (3), we formally define assumption 1, the instruments' relevance to the exposure; this is sometimes referred to as existence of non-redundant instruments in econometrics (Cheng and Liao, 2015).

Definition 2. Suppose that we have p_z candidate instruments along with model (3). We say that instrument $j = 1, \ldots, p_z$ satisfies assumption 1, or is a non-redundant IV, if $\gamma_j^* \neq 0$ and we denote $\mathcal{S}^*$ to be the set of these instruments.

Like definition 1, if $p_z = 1$, definition 2 is a special case of the more general definition of assumption 1 in Angrist *et al.* (1996) and, if $p_z > 1$, our definition is a local version of satisfying assumption 1 in econometrics, which is typically stated as $\boldsymbol{\gamma}^* \neq 0$ (see section 5.2.1 of Wooldridge (2010)). In Mendelian randomization, typically, all p_z instruments are relevant.

Combining definitions 1 and 2, we can formally define the usual three core conditions for instruments, i.e. assumptions 1–3.

Definition 3. Suppose that we have p_z candidate instruments along with models (1)–(3). We say that Z_{ij}, $j=1,\ldots,p_z$, is an instrument if assumptions 1–3 are satisfied, i.e. if $\pi_j^*=0$ and $\gamma_j^*\neq 0$. Let $\mathcal{V}^*=\mathcal{S}^*\cap\mathcal{P}^*$ be the set of instruments.

For the rest of the paper, we define the sparsity level of $\boldsymbol{\pi}^*, \boldsymbol{\phi}^*, \boldsymbol{\gamma}^*$ and $\boldsymbol{\psi}^*$ as $s_{z2}=\|\boldsymbol{\pi}^*\|_0$, $s_{x2}=\|\boldsymbol{\phi}^*\|_0$, $s_{z1}=\|\boldsymbol{\gamma}^*\|_0$ and $s_{x1}=\|\boldsymbol{\psi}^*\|_0$. Let $s=\max\{s_{z2},s_{x2},s_{z1},s_{x1}\}$.

2.2. Identification of model parameters

Identification of the model parameters with invalid instruments has been discussed in several references (Han, 2008; Bowden *et al.*, 2015; Kolesár *et al.*, 2015; Kang *et al.*, 2016b). This section briefly discusses these references to guide the discussion of our inferential method for the treatment effect β^*; because the focus of our paper is inference, we defer additional remarks about identification to section A of the on-line supplementary materials.

We start by rewriting the models of Y and D in equations (2) and (3) in reduced forms, i.e. models of Y and D that are functions of $\mathbf{Z}_{i.}$ and $\mathbf{X}_{i.}$ only:

$$\begin{aligned} D_i &= \mathbf{Z}_{i.}^{\mathrm{T}}\boldsymbol{\gamma}^* + \mathbf{X}_{i.}^{\mathrm{T}}\boldsymbol{\psi}^* + \epsilon_{i2}, \\ \mathbf{E}(\epsilon_{i2}|\mathbf{Z}_{i.},\mathbf{X}_{i.}) &= 0, \end{aligned} \tag{3}$$

$$\begin{aligned} Y_i &= \mathbf{Z}_{i.}^{\mathrm{T}}\boldsymbol{\Gamma}^* + \mathbf{X}_{i.}^{\mathrm{T}}\Psi^* + \epsilon_{i1}, \\ \mathbf{E}(\epsilon_{i1}|\mathbf{Z}_{i.},\mathbf{X}_{i.}) &= 0. \end{aligned} \tag{4}$$

Here, $\boldsymbol{\Gamma}^*=\beta^*\boldsymbol{\gamma}^*+\boldsymbol{\pi}^*$, $\Psi^*=\beta^*\boldsymbol{\psi}^*+\boldsymbol{\phi}^*$ and $\epsilon_{i1}=\beta^*\epsilon_{i2}+e_i$ is the reduced form error term. The term $\boldsymbol{\Gamma}^*$ represents the intent-to-treat effect of the instruments on the outcome and the term $\boldsymbol{\gamma}^*$ represents the association between the instruments and the treatment. The terms ϵ_{i1} and ϵ_{i2} are reduced form errors with covariance matrix $\boldsymbol{\Theta}^*$ where $\Theta_{11}^*=\mathrm{var}(\epsilon_{i1}|\mathbf{Z}_{i.},\mathbf{X}_{i.})$, $\Theta_{22}^*=\mathrm{var}(\epsilon_{i2}|\mathbf{Z}_{i.},\mathbf{X}_{i.})$ and $\Theta_{12}^*=\mathrm{cov}(\epsilon_{i1},\epsilon_{i2}|\mathbf{Z}_{i.},\mathbf{X}_{i.})$. Each reduced form model is the usual regression model with regressors $\mathbf{Z}_{i.}$ and $\mathbf{X}_{i.}$ and outcomes D_i and Y_i and, thus, the parameters of the reduced form models, especially $\boldsymbol{\Gamma}^*$ and $\boldsymbol{\gamma}^*$, can be identified and estimated. Then, the identification of parameters in equations (2) and (3), specifically β^* and $\boldsymbol{\pi}^*$, can be framed as finding conditions that provide a unique, invertible mapping between $\boldsymbol{\Gamma}^*, \boldsymbol{\gamma}^*$ and β^* and $\boldsymbol{\pi}^*$, through the relation $\boldsymbol{\Gamma}^*=\beta^*\boldsymbol{\gamma}^*+\boldsymbol{\pi}^*$. A popular condition is that the majority of the instruments are valid, i.e. the *majority rule or 50% rule*, $|\mathcal{V}^*|>\frac{1}{2}|\mathcal{S}^*|$: Han (2008) and Kang *et al.* (2016b) discussed a special case of the 50% rule where all the instruments are relevant, i.e. $|\mathcal{S}^*|=p_z$. However, as stressed in Kang *et al.* (2016b), the 50% rule is only a sufficient condition to identify the model parameters. A more general condition, which we state in theorem 1, is that the valid instruments form a plurality defined by ratios of $\boldsymbol{\pi}^*$ and $\boldsymbol{\gamma}^*$.

Theorem 1. Suppose that models (2) and (3) hold and $\boldsymbol{\Sigma}^*$ exists and is invertible. Then, given reduced form parameters $\boldsymbol{\gamma}^*$ and $\boldsymbol{\Gamma}^*$, there is a unique β^* and $\boldsymbol{\pi}^*$ if and only if the following *plurality rule* condition holds:

$$|\mathcal{V}^*|>\max_{c\neq 0}|\{j\in\mathcal{S}^*:\pi_j^*/\gamma_j^*=c\}|.$$

The result in theorem 1 provides a blueprint for building a confidence interval for β^*. Specifically, theorem 1 implies that we need an estimate of IVs that satisfy assumption 1, i.e. the set $\mathcal{S}^*$,

and an estimate of IVs that satisfy assumptions 2 and 3, i.e. $\mathcal{P}^*$. Additionally, these estimates must satisfy the plurality rule condition to identify and eventually to construct a confidence interval for β^*. Our method, TSHT with voting, does exactly this. In particular, the first stage of TSHT estimates $\mathcal{S}^*$ and the second stage of TSHT generates many candidate estimates of $\mathcal{P}^*$. The voting step ensures asymptotically that we provide a good estimator of $\mathcal{V}^*$ under the plurality rule condition.

3. Confidence interval estimation via two-stage hard thresholding with voting

3.1. An illustration of two-stage hard thresholding in low dimensional settings

We first illustrate TSHT under the low dimensional setting where $n \gg p_z + p_x$. The low dimensional setting is common in many applications of IVs, such as economics, social sciences and medical sciences, including Mendelian randomization.

As mentioned before, each reduced form model in equations (3) and (4) is the usual regression model with regressors $\mathbf{Z}_{i\cdot}$ and $\mathbf{X}_{i\cdot}$ and outcomes D_i and Y_i respectively. There are consistent and asymptotically normal (CAN) estimators of the regression model parameters in low dimensional settings, for instance estimators based on ordinary least squares (OLS) stated below:

$$
\begin{aligned}
(\hat{\gamma}, \hat{\psi})^{\mathrm{T}} &= (\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{D}, \\
(\hat{\mathbf{\Gamma}}, \hat{\mathbf{\Psi}})^{\mathrm{T}} &= (\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{Y}, \\
\hat{\Theta}_{11} &= \frac{1}{n}\|\mathbf{Y} - \mathbf{Z}\hat{\mathbf{\Gamma}} - \mathbf{X}\hat{\mathbf{\Psi}}\|_2^2, \\
\hat{\Theta}_{22} &= \frac{1}{n}\|\mathbf{D} - \mathbf{Z}\hat{\gamma} - \mathbf{X}\hat{\psi}\|_2^2, \\
\hat{\Theta}_{12} &= \frac{1}{n}(\mathbf{Y} - \mathbf{Z}\hat{\mathbf{\Gamma}} - \mathbf{X}\hat{\mathbf{\Psi}})^{\mathrm{T}}(\mathbf{D} - \mathbf{Z}\hat{\gamma} - \mathbf{X}\hat{\psi}).
\end{aligned}
$$

Let $\hat{\mathbf{U}}$ denote an estimate of $(\mathbf{\Sigma}^*)^{-1}$, the precision matrix of $\mathbf{W}$, i.e. $\hat{\mathbf{U}} = (\mathbf{W}^{\mathrm{T}}\mathbf{W}/n)^{-1}$. Then, $\Theta_{11}\hat{\mathbf{U}}/n$ and $\Theta_{22}\hat{\mathbf{U}}/n$ are the covariance matrices of the OLS estimators $(\hat{\mathbf{\Gamma}}, \hat{\mathbf{\Psi}})$ and $(\hat{\gamma}, \hat{\psi})$ respectively.

The estimators above are the only necessary inputs for TSHT:

(a) CAN estimators of the reduced form coefficients in equations (3) and (4),
(b) a consistent estimator of the error variance matrix $\mathbf{\Theta}^*$ and
(c) the instrument–covariate matrix $\mathbf{W}$, primarily to estimate its precision matrix, $(\mathbf{\Sigma}^*)^{-1}$;

there is an implicit assumption in our notation of $\hat{\gamma}$, $\hat{\psi}$, $\hat{\mathbf{\Gamma}}$ and $\hat{\mathbf{\Psi}}$ where we know which estimates are associated with instruments and covariates and we cannot swap the role of covariates and instruments. Although our discussion was restricted to OLS estimators, any estimator that satisfies the input requirements will work for TSHT. For example, in Section 3.7, we discuss input estimators for TSHT when the data are high dimensional. Finally, we emphasize that no additional choices or inputs are needed for TSHT, such as tuning parameters, beyond the inputs that are stated above.

3.2. First hard thresholding: select strong instrumental variables satisfying assumption 1, $\mathcal{S}^*$

The first thresholding step estimates the set of instruments that satisfy assumption 1, or the set $\mathcal{S}^* = \{j : \gamma_j^* \neq 0\}$ defined in definition 2. To do this, we use one of the inputs for TSHT, the estimator for γ^*: $\hat{\gamma}$. Specifically, if the jth component of $\hat{\gamma}$ exceeds some threshold away from

zero, then j most likely belongs to the set $\mathcal{S}^*$. Estimating $\mathcal{S}^*$ based on this principle is called hard thresholding (Donoho and Johnstone, 1994; Donoho, 1995) and we denote an estimator of $\mathcal{S}^*$ as $\hat{\mathcal{S}}$:

$$\hat{\mathcal{S}} = \left\{ j : |\hat{\gamma}_j| \geqslant \frac{\sqrt{\hat{\Theta}_{22}} \|\mathbf{W}\hat{\mathbf{U}}_{\cdot j}\|_2}{\sqrt{n}} \sqrt{\left[\frac{2.01 \log\{\max(p_z, n)\}}{n} \right]} \right\}. \tag{5}$$

The threshold to declare whether the estimate $\hat{\gamma}_j$ is away from zero consists of two terms. The first term $\sqrt{\hat{\Theta}_{22}} \|\mathbf{W}\hat{\mathbf{U}}_{\cdot j}\|_2/n$ represents the standard error of $\hat{\gamma}_j$. The second term $\sqrt{[2.01 \log\{\max(p_z, n)\}]}$ represents a multiplicity correction for checking whether normally distributed estimators, like $\hat{\gamma}_j$, are away from zero. In particular, the $\sqrt{\{2.01 \log(\cdot)\}}$ part comes from the tail bound of a normal distribution. The $\max(p_z, n)$ part comes from checking multiple $\hat{\gamma}_j$'s distance from zero. Without the multiplier term $\max(p_z, n)$ in equation (5) and if we have many instruments, some estimates $\hat{\gamma}_j$ may exceed the threshold by chance and be part of the set $\hat{\mathcal{S}}$, even though their true γ_j^*s may actually be 0. In practice, $\max(p_z, n)$ is often replaced by p_z or n to improve the finite sample performance of hard thresholding procedures and we explore this numerically in Section 5. But, so long as this term grows with n, like $\max(p_z, n)$ or p_z that grow with n in high dimensional asymptotics, the asymptotic properties of our procedure described in Section 4 hold. Combined are the two terms for the variability of the estimate $\hat{\gamma}_j$ as well as the repeated testing of whether an IV satisfies assumption 1. Finally, the estimator of $\mathcal{S}^*$ does not require selection of tuning parameters, which is in contrast with other variable selection procedures like the lasso (Tibshirani, 1996) which typically uses cross-validation to select the tuning parameters (Hastie *et al.*, 2016); all the components of our threshold in equation (5) are predetermined from the inputs that were provided in Section 3.1.

If external information suggests that all instruments are strongly associated with the exposure, then the first thresholding step may not be necessary and we can simply set $\hat{\mathcal{S}} = \{1, \ldots, p_z\}$. However, when some of these associations may be weak, we recommend running the first thresholding step to improve the finite sample performance of TSHT since the first thresholding should eliminate weak instruments and make TSHT more robust.

3.3. Second hard thresholding: select valid instrumental variables satisfying assumptions 2 and 3, $\mathcal{P}^*$

The second thresholding step estimates the set of instruments that satisfy assumptions 2 and 3, or the set $\mathcal{P}^* = \{j : \pi_j^* \neq 0\}$ defined in definition 1. Unfortunately, unlike the first thresholding step, none of the inputs for TSHT in Section 3.1 directly estimate $\boldsymbol{\pi}^*$, which we can use to estimate $\mathcal{P}^*$ via hard thresholding. Instead, we propose many estimates of $\mathcal{P}^*$ and combine each estimate via voting. We illustrate the estimation of $\mathcal{P}^*$ in this section and the voting in the next section.

To estimate $\mathcal{P}^*$, we take each individually strong IV $j \in \hat{\mathcal{S}}$ and propose a plug-in estimate of $\boldsymbol{\pi}^*$, which is denoted as $\hat{\boldsymbol{\pi}}^{[j]}$, based on the relationship between the model parameters $\boldsymbol{\Gamma}^* = \beta^* \boldsymbol{\gamma}^* + \boldsymbol{\pi}^*$ in equation (4):

$$\hat{\boldsymbol{\pi}}^{[j]} = \hat{\boldsymbol{\Gamma}} - \frac{\hat{\Gamma}_j}{\hat{\gamma}_j} \hat{\boldsymbol{\gamma}}, \qquad j \in \hat{\mathcal{S}}. \tag{6}$$

The terms $\hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{\gamma}}$ in equation (6) are directly from the inputs to TSHT. The term $\hat{\Gamma}_j/\hat{\gamma}_j$ in equation (6) is a Wald-type or ratio estimate of β^* based on instrument j. We also propose an estimate of the variance σ^2 based on this jth strong IV as $\widehat{\sigma^2}^{[j]} = \hat{\Theta}_{11} + (\hat{\beta}^{[j]})^2 \hat{\Theta}_{22} - 2\hat{\beta}^{[j]} \hat{\Theta}_{12}$. In total, we have $|\hat{\mathcal{S}}|$ estimates of $\boldsymbol{\pi}^*$ and σ^2.

For each estimate of $\boldsymbol{\pi}^*$, we estimate the set $\mathcal{P}^*$ similarly to the first hard thresholding step; the only difference is that we select instrument k with $\pi_k^* = 0$ whereas, in the first thresholding step, we select instrument k with $\gamma_k^* \neq 0$. Specifically, for each estimate $\boldsymbol{\pi}^{[j]}$, we threshold each component of the vector $\boldsymbol{\pi}^{[j]}$ below some threshold and we denote the set consisting of these components as $\hat{\mathcal{P}}^{[j]}$:

$$\hat{\mathcal{P}}^{[j]} =: \left\{ k : |\hat{\pi}_k^{[j]}| \leqslant \sqrt{\widehat{\sigma^2}^{[j]}} \frac{\|\mathbf{W}\{\hat{\mathbf{U}}_{.k} - (\hat{\gamma}_k/\hat{\gamma}_j)\hat{\mathbf{U}}_{.j}\}\|_2}{\sqrt{n}} \sqrt{\left[\frac{2.01^2 \log\{\max(p_z, n)\}}{n} \right]} \right\}. \tag{7}$$

Like the first threshold in equation (5), the threshold in equation (7) comprises two terms: the term $\sqrt{\widehat{\sigma^2}^{[j]}}\|\mathbf{W}(\hat{\mathbf{U}}_{.k} - (\hat{\gamma}_k/\hat{\gamma}_j)\hat{\mathbf{U}}_{.j})\|_2/n$ representing the standard error of $\hat{\pi}_k^{[j]}$ and $\sqrt{[2.01^2 \log\{\max(p_z, n\}]}$ representing the multiplicity correction. The constant 2.01^2 is because we are performing (at most) p_z^2 hypothesis testing for all candidate–component combinations. Also, similarly to the first thresholding step, the thresholds in equation (7) are predetermined. In the end, we have $|\hat{\mathcal{S}}|$ estimates of $\mathcal{P}^*$, $\hat{\mathcal{P}}^{[j]}$, $j \in \hat{\mathcal{S}}$.

Combining the two thresholding steps gives estimates of IVs that satisfy all assumptions 1–3, or the set $\mathcal{V}^*$ in definition 3. Specifically, each intersection $\hat{\mathcal{V}}^{[j]} = \hat{\mathcal{S}} \cap \hat{\mathcal{P}}^{[j]}$ is an estimate of $\mathcal{V}^*$ and we have $|\hat{\mathcal{S}}|$ estimates of $\mathcal{V}^*$ (i.e. $\hat{\mathcal{V}}^{[j]}$, $j \in \hat{\mathcal{S}}$). The remaining task is to combine the information from these estimates to produce a single consistent estimate of the set $\mathcal{V}^*$.

3.4. Majority and plurality voting

To explain how we combine several estimates $\hat{\mathcal{V}}^{[j]}$, $j \in \hat{\mathcal{S}}$, to produce a single estimate of $\mathcal{V}^*$, it is helpful to consider a voting analogy where each $j \in \hat{\mathcal{S}}$ is an expert and $\hat{\mathcal{V}}^{[j]}$ is expert j's ballot that contains expert j's opinion about which instruments he or she believes satisfy assumptions 1–3. Because $\hat{\mathcal{V}}^{[j]} \subseteq \hat{\mathcal{S}}$ for any j, all experts must pick instruments from the set $\hat{\mathcal{S}}$ when they cast their ballots. For example, $k \in \hat{\mathcal{V}}^{[j]}$ indicates that expert j voted on instrument k as satisfying assumptions 1–3.

Following the voting analogy, we can tally the number of experts who cast their votes for a particular candidate IV as satisfying assumptions 1–3. Specifically, let $\mathbf{1}(k \in \hat{\mathcal{V}}^{[j]})$ be the indicator function that denotes whether the kth instrument belongs to $\hat{\mathcal{V}}^{[j]}$ and $\mathrm{VM}_k = \Sigma_{j \in \hat{\mathcal{S}}} \mathbf{1}(k \in \hat{\mathcal{V}}^{[j]})$ denote the tally of votes that the kth instrument received from all experts where $k \in \hat{\mathcal{S}}$. For example, $\mathrm{VM}_k = 3$ indicates that three out of $|\hat{\mathcal{S}}|$ total experts have voted instrument k as satisfying assumptions 1–3.

Now, suppose that the kth instrument received votes from a majority of experts, i.e. more than 50% of experts, as satisfying assumptions 1–3, i.e. $\mathrm{VM}_k > \frac{1}{2}|\hat{\mathcal{S}}|$. Let $\hat{\mathcal{V}}_M$ consist of such instruments and we refer to this type of voting as majority voting:

$$\hat{\mathcal{V}}_M = \left\{ k \in \hat{\mathcal{S}} | \mathrm{VM}_k > \tfrac{1}{2}|\hat{\mathcal{S}}| \right\}. \tag{8}$$

Also suppose that $\hat{\mathcal{V}}_M$ is empty and no instrument won support from a majority of the voters. Instead, suppose that a candidate IV k received a plurality of votes to satisfy assumptions 1–3, i.e. $\mathrm{VM}_k = \max_l \mathrm{VM}_l$. Let $\hat{\mathcal{V}}_P$ denote instruments that received a plurality of votes and we refer to this type of voting as plurality voting:

$$\hat{\mathcal{V}}_{\mathrm{P}} = \left\{ k \in \hat{\mathcal{S}} | \mathrm{VM}_k = \max_l \mathrm{VM}_l \right\}. \tag{9}$$

Intuitively, $\hat{\mathcal{V}}_{\mathrm{M}}$ or $\hat{\mathcal{V}}_{\mathrm{P}}$ is a good proxy of $\mathcal{V}^*$ if the 50% rule or plurality rule condition respectively hold. For example, if instrument $k \in \mathcal{V}^*$ and the 50% rule condition held, a majority of experts would vote for k so that $k \in \hat{\mathcal{V}}^{[j]}$ and the total votes for instrument k across experts, VM_k,

would exceed 50% so that $k \in \hat{\mathcal{V}}_{\mathrm{M}}$. Similarly, under the plurality rule condition, there are more experts using valid instruments to inform their ballots $\hat{\mathcal{V}}^{[j]}$ and their ballots contain $k \in \hat{\mathcal{V}}^{[j]}$; in contrast, experts using invalid instruments would not contain k and their ballots would not form a plurality over any instrument l under consideration. Thus, VM_k would be the largest among all instruments under consideration and $k \in \hat{\mathcal{V}}_{\mathrm{P}}$.

A single, robust estimate of $\mathcal{V}^*$ under any of the two conditions is the union of the two sets $\hat{\mathcal{V}} = \hat{\mathcal{V}}_{\mathrm{M}} \cup \hat{\mathcal{V}}_{\mathrm{P}}$. Technically speaking, because the plurality rule condition is both sufficient and necessary, the union can consist of only the set $\hat{\mathcal{V}}_{\mathrm{P}}$. However, we find that, in simulation studies and in practice, taking the union of the two sets provides robustness in finite samples.

3.5. Point estimate, standard error and confidence interval

Once we estimated the set of instruments that satisfy assumptions 1–3, i.e. $\hat{\mathcal{V}}$, estimation and inference of β^* are straightforward in the low dimensional setting. In particular, we can use two-stage least squares (TSLS) with $\hat{\mathcal{V}}$ as the set of IVs that satisfy assumptions 1–3 and obtain a point estimate for β^*, which we denote as $\hat{\beta}_{\mathrm{L}}$

$$\hat{\beta}_{\mathrm{L}} = \frac{\hat{\gamma}_{\hat{\mathcal{V}}}^{\mathrm{T}} \hat{\mathbf{A}} \hat{\mathbf{\Gamma}}_{\hat{\mathcal{V}}}}{\hat{\gamma}_{\hat{\mathcal{V}}}^{\mathrm{T}} \hat{\mathbf{A}} \hat{\gamma}_{\hat{\mathcal{V}}}}, \qquad \hat{\mathbf{A}} = \hat{\mathbf{\Sigma}}_{\hat{\mathcal{V}},\hat{\mathcal{V}}} - \hat{\mathbf{\Sigma}}_{\hat{\mathcal{V}},\hat{\mathcal{V}}^{\mathrm{c}}} \hat{\mathbf{\Sigma}}_{\hat{\mathcal{V}}^{\mathrm{c}},\hat{\mathcal{V}}^{\mathrm{c}}}^{-1} \hat{\mathbf{\Sigma}}_{\hat{\mathcal{V}}^{\mathrm{c}},\hat{\mathcal{V}}}. \tag{10}$$

The $\hat{\mathbf{A}}$ is a weighting matrix for the estimates $\hat{\gamma}$ and $\hat{\mathbf{\Gamma}}$, which, among other things, comprises $\hat{\mathbf{\Sigma}} = \mathbf{W}^{\mathrm{T}}\mathbf{W}/n$, the inverse of the estimated precision matrix of $\mathbf{W}$ that we used in the inputs for TSHT. The estimated variance of $\hat{\beta}_{\mathrm{L}}$ is

$$\widehat{\mathrm{var}}_{\mathrm{L}} = \frac{\hat{\gamma}_{\hat{\mathcal{V}}}^{\mathrm{T}} \hat{\mathbf{A}} (\hat{\mathbf{\Sigma}}^{-1})_{\hat{\mathcal{V}},\hat{\mathcal{V}}} \hat{\mathbf{A}} \hat{\gamma}_{\hat{\mathcal{V}}}}{(\hat{\gamma}_{\hat{\mathcal{V}}}^{\mathrm{T}} \hat{\mathbf{A}} \hat{\gamma}_{\hat{\mathcal{V}}})^2} (\hat{\Theta}_{11} + \hat{\beta}_{\mathrm{L}}^2 \hat{\Theta}_{22} - 2\hat{\beta}_{\mathrm{L}} \hat{\Theta}_{12}) \tag{11}$$

which simplifies to

$$\widehat{\mathrm{var}}_{\mathrm{L}} = \frac{\hat{\Theta}_{11} + \hat{\beta}_{\mathrm{L}}^2 \hat{\Theta}_{22} - 2\hat{\beta}_{\mathrm{L}} \hat{\Theta}_{12}}{\hat{\gamma}_{\hat{\mathcal{V}}}^{\mathrm{T}} \hat{\mathbf{A}} \hat{\gamma}_{\hat{\mathcal{V}}}}.$$

Finally, for any α where $0 < \alpha < 1$, the $1 - \alpha$ confidence interval for β^* is

$$(\hat{\beta}_{\mathrm{L}} - z_{1-\alpha/2}\sqrt{(\widehat{\mathrm{var}}_{\mathrm{L}}/n)},\ \hat{\beta}_{\mathrm{L}} + z_{1-\alpha/2}\sqrt{(\widehat{\mathrm{var}}_{\mathrm{L}}/n)}), \tag{12}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution.

In Section 4.1, we show that the $\hat{\beta}_{\mathrm{L}}$ achieves optimal performance in the sense that $\hat{\beta}_{\mathrm{L}}$ converges to an asymptotic normal distribution that is identical to the asymptotic normal distribution of the TSLS estimator for β^* that knows which IVs satisfy assumptions 1–3, i.e. $\mathcal{V}^*$.

3.6. Comparison with other methods

We make some remarks about our method and the methods that have been proposed in the literature on invalid IVs. The work by Windmeijer *et al.* (2016) is the methodologically most similar to our method in that it also estimates $\mathcal{V}^*$ and uses the estimate of $\mathcal{V}^*$ to obtain an oracle optimal point estimate and confidence interval of β^* like we do in Section 3.5. To estimate $\mathcal{V}^*$, Windmeijer *et al.* (2016) utilized the adaptive lasso with a median estimator of Han (2008) and Bowden *et al.* (2016) as the initial estimator; the tuning parameter in the adaptive lasso is chosen by cross-validation. In contrast, TSHT with voting utilizes hard thresholding

steps to estimate $\mathcal{V}^*$ where our 'tuning' parameters, i.e. the thresholds, are predetermined and theoretically motivated. On the basis of numerical results in Section 5.2, we suspect that, in low dimensional settings, their method and TSHT with voting are asymptotically equivalent when the 50% rule condition holds.

Another inferential method in the invalid IV literature is bootstrapping the median estimator (Bowden *et al*., 2016; Burgess *et al*., 2016). The key idea is to go directly after the target estimand β^* with the median estimator mentioned above and to bootstrap the estimate with sufficient statistics. Their works are under the two-sample designs with summary data where the errors in the reduced form models are independent of each other. In contrast, TSHT with voting and the method of Windmeijer *et al*. (2016) focus on the one-sample design with individual level data and correlated error terms. Also, neither method utilizes the bootstrap to generate inferential quantities.

We argue that TSHT with voting is a major improvement from the methods of Windmeijer *et al*. (2016), Bowden *et al*. (2016) and Burgess *et al*. (2016) for the following three reasons. First, all three methods rely on the 50% rule condition because their estimators rely on the median estimator, which is consistent for β^* only under the 50% rule condition. In contrast, TSHT with voting does not rely on an initial consistent estimator and our inferential guarantees are possible under the more general plurality rule condition. Second, the median methods of Bowden *et al*. (2016) and Burgess *et al*. (2016) may not be oracle optimal in the sense that it may not be as efficient as the oracle estimator that knows, *a priori*, which instruments are invalid. The method of Windmeijer *et al*. (2016) is oracle optimal in low dimensional settings, but only when the majority rule holds. In contrast, TSHT with voting is oracle optimal in low dimensional settings under the more general plurality rule condition; see Section 4.1. Third, there are no theoretical guarantees that the bootstrap approach to inference for the median method will always generate a confidence interval that will cover the true parameter with probability $1-\alpha$, although it does perform well in large numerical studies under two-sample designs (Burgess *et al*., 2016). Similarly, the theoretical properties of the method of Windmeijer *et al*. (2016) are under the assumption that the tuning parameter is not chosen via cross-validation, despite the fact that Windmeijer *et al*. (2016) utilized cross-validation when they used their method in simulations and in a real data example. In contrast, TSHT with voting uses predetermined thresholding values both in theory and in numerical studies and has theoretical guarantees on inference; see Section 4. Finally, work by Kang *et al*. (2016a) does not rely on the 50% rule to obtain inferential quantities, but it is conservative and works only in low dimensional settings.

The work by Kang *et al*. (2016b), which is the precursor of this paper, also proposes a joint estimator of β^* and $\boldsymbol{\pi}^*$ called sisVIVE. The estimator of Kang *et al*. (2016b) is based on the lasso that minimizes the sum of squared errors from model (2) with respect to an l_1-penalty on $\boldsymbol{\pi}^*$. The tuning parameter of the lasso is chosen via cross-validation. A nice feature of sisVIVE is that it is a one-step method to estimate β^*. In contrast, TSHT requires two thresholding steps plus voting to estimate β^*. Unfortunately, sisVIVE requires more stringent conditions for consistency than the identification 50% rule condition. Also, like the method of Windmeijer *et al*. (2016), the theory behind consistency is developed under the assumption that the tuning parameter is not chosen via cross-validation. More importantly, sisVIVE did not resolve the issue of confidence interval construction.

Finally, since this paper was submitted for review, Hartwig *et al*. (2017) proposed confidence intervals for β^* under the plurality rule condition, referred to as the zero modal pleiotropy assumption. Their method fits a kernel smoothing density on the distribution of different estimates of β^* from each instrument and takes the mode of the fitted density as the estimate of

β^*. Inference is achieved by running a bootstrap. Although the method is simple to understand, the method suffers from

(a) choosing a good bandwidth parameter for the kernel smoothing density estimator,
(b) a potential lack of oracle optimality and
(c) no theoretical guarantee on inference.

In contrast, TSHT uses predetermined thresholds that lead to oracle optimal and valid inference for the parameter β^*.

3.7. High dimensional setting

TSHT with voting can accommodate settings where we have high dimensional covariates and/or instruments. The modifications that we must make are the estimation of the reduced form model parameters in equations (3) and (4), the weighting matrix A in equation (10) and the formula for the standard error; the rest of the procedure is identical.

Specifically, instead of using OLS estimators in Section 3.1, we must resort to estimators that can handle the case when $n \ll p$ and are CAN so that the input requirements for TSHT are met. There are many estimators in high dimensions that meet this criterion, such as the debiased lasso or its variants laid out in Zhang and Zhang (2014), Javanmard and Montanari (2014), van de Geer *et al*. (2014) and Cai and Guo (2017). For completeness, we present one estimator in high dimensional regression that is CAN: the debiased square-root lasso estimator (Belloni *et al*., 2011; Javanmard and Montanari, 2014); see the references cited for additional details on CAN estimators in high dimensions. First, the square-root lasso estimator (Belloni *et al*., 2011) estimates high dimensional reduced form model parameters in equations (3) and (4) based on the following optimization problems:

$$\{\tilde{\mathbf{\Gamma}}, \tilde{\mathbf{\Psi}}\} = \underset{\mathbf{\Gamma}\in\mathbb{R}^{p_z}, \mathbf{\Psi}\in\mathbb{R}^{p_x}}{\arg\min} \frac{\|\mathbf{Y} - \mathbf{Z}\mathbf{\Gamma} - \mathbf{X}\mathbf{\Psi}\|_2}{\sqrt{n}} + \frac{\sqrt{\{2.01\log(p)\}}}{n}\left(\sum_{j=1}^{p_z} \|\mathbf{Z}_{\cdot j}\|_2|\Gamma_j| + \sum_{j=1}^{p_x} \|\mathbf{X}_{\cdot j}\|_2|\Psi_j|\right),$$

$$\{\tilde{\gamma}, \tilde{\psi}\} = \underset{\mathbf{\Gamma}\in\mathbb{R}^{p_z}, \mathbf{\Psi}\in\mathbb{R}^{p_x}}{\arg\min} \frac{\|\mathbf{D} - \mathbf{Z}\gamma - \mathbf{X}\psi\|_2}{\sqrt{n}} + \frac{\sqrt{\{2.01\log(p)\}}}{n}\left(\sum_{j=1}^{p_z} \|\mathbf{Z}_{\cdot j}\|_2|\gamma_j| + \sum_{j=1}^{p_x} \|\mathbf{X}_{\cdot j}\|_2|\psi_j|\right).$$

Also, the corresponding estimates of the variances Θ_{11}^*, Θ_{22}^* and Θ_{12}^* from the square-root lasso are

$$\hat{\Theta}_{11} = \frac{1}{n}\|\mathbf{Y} - \mathbf{Z}\tilde{\mathbf{\Gamma}} - \mathbf{X}\tilde{\mathbf{\Psi}}\|_2^2,$$
$$\hat{\Theta}_{22} = \frac{1}{n}\|\mathbf{D} - \mathbf{Z}\tilde{\gamma} - \mathbf{X}\tilde{\psi}\|_2^2,$$
$$\hat{\Theta}_{12} = \frac{1}{n}(\mathbf{Y} - \mathbf{Z}\tilde{\mathbf{\Gamma}} - \mathbf{X}\tilde{\mathbf{\Psi}})^{\mathrm{T}}(\mathbf{D} - \mathbf{Z}\tilde{\gamma} - \mathbf{X}\tilde{\psi}).$$

Unfortunately, the square-root lasso estimator is biased because of the penalty term and Javanmard and Montanari (2014) proposed a way to debias the square-root lasso estimator and to turn it into CAN estimators. Specifically, Javanmard and Montanari (2014) proposed p_z optimization problems where the solution to each p_z optimization problem, which is denoted as $\hat{\mathbf{U}}_{\cdot j} \in \mathbb{R}^p$, $j = 1, \ldots, p_z$, is

$$\hat{\mathbf{U}}_{\cdot j} = \underset{\mathbf{u}\in\mathbb{R}^p}{\arg\min} \frac{1}{n}\|\mathbf{W}\mathbf{u}\|_2^2 \qquad \text{subject to } \|\hat{\mathbf{\Sigma}}\mathbf{u} - \mathbf{I}_{\cdot j}\|_\infty \leqslant 12M_1^2\sqrt{\left\{\frac{\log(p)}{n}\right\}} \tag{13}$$

with $\hat{\mathbf{\Sigma}} = (1/n)\mathbf{W}^{\mathrm{T}}\mathbf{W}$. Here, $\mathbf{I}_{\cdot j}$ denotes the jth column of the identity matrix $\mathbf{I}$ and M_1 denotes the largest eigenvalue of $\mathbf{\Sigma}^*$. Let $\hat{\mathbf{U}}$ denote the matrix concatenation of the p_z solutions to the optimization problem. Then, the debiased estimates of $\tilde{\mathbf{\Gamma}}$ and $\tilde{\gamma}$, which are denoted as $\hat{\mathbf{\Gamma}}$ and $\hat{\gamma}$, are

$$\begin{aligned}\hat{\mathbf{\Gamma}} &= \tilde{\mathbf{\Gamma}} + \frac{1}{n}\hat{\mathbf{U}}\mathbf{W}^{\mathrm{T}}(\mathbf{Y} - \mathbf{Z}\tilde{\mathbf{\Gamma}} - \mathbf{X}\tilde{\mathbf{\Psi}}),\\ \hat{\gamma} &= \tilde{\gamma} + \frac{1}{n}\hat{\mathbf{U}}\mathbf{W}^{\mathrm{T}}(\mathbf{D} - \mathbf{Z}\tilde{\gamma} - \mathbf{X}\tilde{\psi}).\end{aligned} \tag{14}$$

We now have obtained all the ingredients for TSHT in the high dimensional setting:

(a) the CAN estimators of $\mathbf{\Gamma}^*$ and γ^*, $\hat{\mathbf{\Gamma}}$ and $\hat{\gamma}$ respectively, based on the debiased square-root lasso;
(b) consistent estimators of the error variances $\Theta^*_{11}, \Theta^*_{22}$ and Θ^*_{12}, $\hat{\Theta}_{11}, \hat{\Theta}_{22}$ and $\hat{\Theta}_{12}$ respectively, from the square-root lasso;
(c) an estimate of the precision matrix of $\mathbf{W}$, $\hat{\mathbf{U}}$ from the debiasing procedure.

Running TSHT with these inputs will estimate the set of valid instruments $\hat{\mathcal{V}}$ in high dimensional settings.

For point estimation of β^* in high dimensions, we simply replace $\hat{\mathbf{A}}$ in equation (10) with the identity matrix

$$\hat{\beta} = \frac{\hat{\gamma}_{\hat{\mathcal{V}}}^{\mathrm{T}}\hat{\mathbf{\Gamma}}_{\hat{\mathcal{V}}}}{\hat{\gamma}_{\hat{\mathcal{V}}}^{\mathrm{T}}\hat{\gamma}_{\hat{\mathcal{V}}}}. \tag{15}$$

The variance estimate of the estimator in equation (15) uses a high dimensional estimate of the precision matrix in equation (11), i.e.

$$\widehat{\mathrm{var}} = \frac{\hat{\gamma}_{\hat{\mathcal{V}}}^{\mathrm{T}}(\hat{\mathbf{U}}_{\cdot,\hat{\mathcal{V}}})^{\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\mathbf{W}/n)\hat{\mathbf{U}}_{\cdot,\hat{\mathcal{V}}}\hat{\gamma}_{\hat{\mathcal{V}}}}{(\hat{\gamma}_{\hat{\mathcal{V}}}^{\mathrm{T}}\hat{\gamma}_{\hat{\mathcal{V}}})^2}(\hat{\Theta}_{11} + \hat{\beta}^2\hat{\Theta}_{22} - 2\hat{\beta}\hat{\Theta}_{12}). \tag{16}$$

Given the point estimate and the variance, the confidence interval for β^* follows the usual form

$$(\hat{\beta} - z_{1-\alpha/2}\sqrt{(\widehat{\mathrm{var}}/n)},\ \hat{\beta} + z_{1-\alpha/2}\sqrt{(\widehat{\mathrm{var}}/n)}). \tag{17}$$

4. Theoretical results

In this section, we state the asymptotic properties of TSHT with voting. In Section 4.1, we consider the low dimensional setting where p_x and p_z are fixed. In Section 4.2, we consider the general case when p_x and/or p_z are allowed to grow and exceed sample size n.

4.1. Invalid instrumental variables in low dimensional setting

First, we prove that the estimated set $\hat{\mathcal{V}}$ is an asymptotically consistent estimator of the true set $\mathcal{V}^*$ in the low dimensional setting where p_x and p_z are fixed.

Lemma 1. Under the plurality rule assumption, $\lim_{n\to\infty}\mathbf{P}(\hat{\mathcal{V}} = \mathcal{V}^*) = 1$.

Lemma 1 confirms our intuition in Section 3.4 that the voting process correctly generates the set of instruments that are relevant and valid. In fact, a useful feature of our method is that it provably and correctly selects the IVs that satisfy assumptions 1–3, which is something that is

not possible with prior methods that target only β^*, e.g. the median method. Next, theorem 2 states that the confidence interval that was outlined in Section 3.5 has the desired coverage and optimal length in the low dimensional settings with fixed p_x and p_z.

Theorem 2. Suppose that the plurality rule assumption holds. Then, as $n \to \infty$, we have

$$\sqrt{n}(\hat{\beta}_{\mathrm{L}} - \beta^*) \xrightarrow{\mathrm{d}} N\left\{0, \frac{\sigma^2}{\gamma^{*\mathrm{T}}_{\mathcal{V}^*}(\boldsymbol{\Sigma}^*_{\mathcal{V}^*\mathcal{V}^*} - \boldsymbol{\Sigma}^*_{\mathcal{V}^*(\mathcal{V}^*)^{\mathrm{C}}}\boldsymbol{\Sigma}^{*-1}_{(\mathcal{V}^*)^{\mathrm{C}}(\mathcal{V}^*)^{\mathrm{C}}}\boldsymbol{\Sigma}^*_{(\mathcal{V}^*)^{\mathrm{C}}\mathcal{V}^*})\gamma^*_{\mathcal{V}^*}}\right\}. \tag{18}$$

Consequently, the confidence interval that is given in equation (12) has asymptotic coverage probability $1-\alpha$, i.e.

$$\mathbf{P}\left\{\beta^* \in (\hat{\beta}_{\mathrm{L}} - z_{1-\alpha/2}\sqrt{(\widehat{\mathrm{var}}_{\mathrm{L}}/n)},\ \hat{\beta}_{\mathrm{L}} + z_{1-\alpha/2}\sqrt{(\widehat{\mathrm{var}}_{\mathrm{L}}/n)})\right\} \to 1-\alpha. \tag{19}$$

We note that the proposed estimator $\hat{\beta}_{\mathrm{L}}$ has the same asymptotic variance as the oracle TSLS estimator with prior knowledge of $\mathcal{V}^*$, which is efficient under the homoscedastic variance assumption; see theorem 5.2 in Wooldridge (2010) for details. Consequently, our confidence interval in equation (12) asymptotically performs like the oracle TSLS confidence interval and is of optimal length. But, unlike TSLS, we achieve this oracle performance without prior knowledge of $\mathcal{V}^*$. We remind readers that the previous estimators that were proposed by Bowden *et al.* (2015, 2016), Burgess *et al.* (2016) and Kang *et al.* (2016a) do not achieve oracle performance and TSLS-like efficiency whereas the estimator that was proposed by Windmeijer *et al.* (2016) does achieve this, but only when the 50% rule condition holds.

4.2. Invalid instrumental variables in high dimensional setting

We now consider the asymptotic properties of TSHT with voting under the general case when p_z and/or p_x are allowed to grow, potentially exceeding sample size n. As noted in Section 3.2, to be in alignment with the traditional high dimensional literature where p_z and/or p_x are always larger than n and growing faster than n, we simplify TSHT by replacing the thresholds in equations (5) and (7) from $\log(\max\{p_z, n\})$ to $\log(p_z)$.

We first introduce the regularity assumptions that are used in high dimensional statistics (Bickel *et al.*, 2009; Bühlmann and van de Geer, 2011; Cai and Guo, 2017).

Assumption 4 (coherence). The matrix $\boldsymbol{\Sigma}^*$ satisfies $1/M_1 \leqslant \lambda_{\min}(\boldsymbol{\Sigma}^*) \leqslant \lambda_{\max}(\boldsymbol{\Sigma}^*) \leqslant M_1$ for some constant $M_1 \geqslant 1$ and has bounded sub-Gaussian norm.

Assumption 5 (normality). The error terms in equations (3) and (4) follow a bivariate normal distribution.

Assumption 6 (global IV strength). The IVs are globally strong with $\sqrt{\{(\gamma^*_{\mathcal{V}^*})^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathcal{V}^*,\mathcal{V}^*}\gamma^*_{\mathcal{V}^*}\}} \gg s_{z1}\log(p)/\sqrt{n}$, where $\mathcal{V}^*$ is the set of valid IVs defined in definition 3.

Assumption 4 makes sure that the spectrum of the design matrix $\mathbf{W}$ is well behaved as $p \to \infty$. Assumption 5 is made out of simplicity, similarly to the normal error assumption that is made in the work on inference in the weak IV literature (e.g. section 2 of Moreira (2003) and section 2.2.1 of Andrews *et al.* (2007)) and in high dimensional linear models (e.g. theorem 2.5 in Javanmard and Montanari (2014) and theorem 2.1 in van de Geer *et al.* (2014)). Finally, assumption 6 bounds the global strength of the instruments, measured by the weighted l_2-norm of $\gamma^*_{\mathcal{V}^*}$, away from zero. Assumption 6 is commonly assumed in the traditional IV literature under the guise of a concentration parameter, which is a measure of instrument strength and is the weighted l_2-norm of $\gamma^*_{\mathcal{V}^*}$ (Stock *et al.*, 2002; Wooldridge, 2010), and in the high dimensional IV literature

Belloni *et al*. (2012) and Chernozhukov *et al*. (2015). Sections B and C in the on-line supplementary materials provide additional discussions and show that, if the IVs are valid, then regularity assumptions 4–6 are sufficient to construct valid confidence intervals in high dimensions.

When IVs are invalid, we need to make two additional assumptions that are not part of the high dimensional statistics or IVs literature and may be of theoretical interest in future work.

Assumption 7 (individual IV strength). For IVs in $\mathcal{S}^*$, $\delta_{\min} = \min_{j\in\mathcal{S}^*} |\gamma_j^*| \gg \sqrt{\{\log(p_z)/n\}}$.

Assumption 8 (separated levels of violation). For the pair $j, k \in \mathcal{S}^*$ with $\pi_j^*/\gamma_j^* \neq \pi_k^*/\gamma_k^*$,

$$\left|\frac{\pi_j^*}{\gamma_j^*} - \frac{\pi_k^*}{\gamma_k^*}\right| \geqslant \frac{12(1+\max_{j\in\mathcal{S}^*}|\Gamma_j^*/\gamma_j^*|)}{\delta_{\min}} \sqrt{\left\{\frac{M_1 \log(p_z)}{\lambda_{\min}(\boldsymbol{\Theta}^*)n}\right\}}. \tag{20}$$

Assumption 8 bounds individual IV strength away from zero so that all IVs in selected $\hat{\mathcal{S}}$ are strong. Without this condition, an individually weak IV with small $\hat{\gamma}_j$ may be included in the first thresholding step and subsequently cause trouble to the second thresholding step in equation (7) that uses $\hat{\gamma}_j$ in the denominator to construct a candidate estimate of $\boldsymbol{\pi}^*$ and $\mathcal{P}^*$. In the literature, assumption 8 is similar to the 'beta-min'-assumption in high dimensional linear regression without IVs, with the exception that this condition is not imposed on our inferential quantity of interest, β^*. Also, assumption 8 is different from assumption 6 in that assumption 6 requires only the global IV strength to be bounded away from zero. Next, assumption 9 requires that the difference between different levels of ratios π_j^*/γ_j^* is sufficiently large. Without this assumption, it would be difficult to distinguish subsets of instruments with different π_j^*/γ_j^*-values from the data and to identify the set of valid IVs based on the plurality rule. For example, consider instruments k and j with $\pi_k^*/\gamma_k^* \neq \pi_j^*/\gamma_j^*$. If equation (20) is satisfied, then $k \notin \hat{\mathcal{P}}^{[j]}$ with high probability because π_k^*/γ_k^* and π_j^*/γ_j^* are far apart from each other. In contrast, if equation (20) does not hold, then $\hat{\mathcal{P}}^{[j]}$ might contain instrument k by chance because π_k^*/γ_k^* and π_j^*/γ_j^* are close to each other.

Lemma 2 shows that, with assumptions 4–6 and 8 and 9 and the plurality rule condition, TSHT with voting produces a consistent estimator of the set of valid instruments in the high dimensional setting.

Lemma 2. Suppose that $s\sqrt{s_{z1}}\log(p)/\sqrt{n} \to 0$ and assumptions 4–6, and 8 and 9 and the plurality rule condition are satisfied. With probability larger than $1 - c\{p^{-c} + \exp(-cn)\}$ for some $c > 0$, $\hat{\mathcal{V}} = \mathcal{V}^*$.

Next, the following theorem shows that $\hat{\beta}$ is a consistent and asymptotic normal estimator of β^*.

Theorem 3. Under the same assumption as lemma 2, we have

$$\sqrt{n}(\hat{\beta} - \beta^*) = T^{\beta^*} + \Delta^{\beta^*} \tag{21}$$

where $T^{\beta^*}|\mathbf{W} \sim N(0, \text{var})$ and $\text{var} = \sigma^2 \boldsymbol{\gamma}_{\mathcal{V}^*}^{\mathrm{T}} (\hat{\mathbf{U}}_{\cdot,\mathcal{V}^*})^{\mathrm{T}} \mathbf{W}^{\mathrm{T}} \mathbf{W} \hat{\mathbf{U}}_{\cdot,\mathcal{V}^*} \boldsymbol{\gamma}_{\mathcal{V}^*} / n(\boldsymbol{\gamma}_{\mathcal{V}^*}^{\mathrm{T}} \boldsymbol{\gamma}_{\mathcal{V}^*})^2$. As $s\sqrt{s_{z1}}\log(p)/\sqrt{n} \to 0$, $\Delta^{\beta^*}/\sqrt{\text{var}} \to^{\mathrm{p}} 0$ and the confidence interval that is given in equation (17) has asymptotic coverage probability of $1 - \alpha$, i.e.

$$\mathbf{P}\{\beta^* \in (\hat{\beta} - z_{1-\alpha/2}\sqrt{(\widehat{\text{var}}/n)},\ \hat{\beta} + z_{1-\alpha/2}\sqrt{(\widehat{\text{var}}/n)})\} \to 1 - \alpha. \tag{22}$$

5. Simulation

5.1. Set-up: low dimensional setting

In addition to the theoretical analysis of our method in Section 4, we also conduct a simulation study to investigate the numerical performance of our method. The design of the simulation study follows closely that of Windmeijer *et al.* (2016) where we use models (2) and (3) in Section 2.1. Specifically,

(a) there are $p_z = 7$ or $p_z = 10$ instruments,
(b) there are no covariates,
(c) the instruments are generated from a multivariate normal distribution with mean 0 and identity covariance,
(d) the treatment effect is fixed to be $\beta^* = 1$ and
(e) the errors have variance 1 and covariance 0.25.

Similarly to Windmeijer *et al.* (2016), we vary

(a) the sample size n,
(b) the strength of the IV by manipulating $\boldsymbol{\gamma}^* = (1, \ldots, 1)C_\gamma$ with different values of C_γ and
(c) the degree of violations of assumptions 2 and 3 by manipulating $\boldsymbol{\pi}^*$.

With respect to the last variation, if $p_z = 7$, we set $\boldsymbol{\pi}^* = (1, 1, 0.5, 0.5, 0, 0, 0)C_\pi$ where C_π is a constant that we vary to change the magnitude of $\boldsymbol{\pi}^*$. If $p_z = 10$, we set $\boldsymbol{\pi}^* = (1, 1, 1, 0, \ldots, 0)C_\pi$. The first setting mimics the case where the 50% rule condition holds, similarly to Windmeijer *et al.* (2016), whereas the second setting mimics the case where the 50% rule fails but the plurality rule condition holds.

Under this data-generating mechanism, we compare our procedure TSHT with voting with

(a) the naive TSLS that assumes that all the instruments satisfy assumptions 1–3,
(b) the oracle TSLS that knows, *a priori*, which instruments satisfy assumptions 1–3,
(c) the method of Windmeijer *et al.* (2016) that uses the adaptive lasso tuned via cross-validation and the initial median estimator and
(d) the unweighted median estimator of Bowden *et al.* (2016) and Burgess *et al.* (2016) with bootstrapped confidence intervals by using the R package `MendelianRandomization` (Yavorska and Burgess, 2017) under default settings.

For (a), we implement TSLS so that it mimics most practitioners' use of TSLS by simply assuming that all the instruments $\mathbf{Z}$ are valid. For (b), we have the oracle TSLS where an oracle provides us with the true set of valid IVs, which will not occur in practice. Because TSLS is not robust against weak instruments, we purposely set our C_γ to correspond to strong IV regimes. Finally, for (c) and (d), see Section 3.6 for discussions of the methods. Our simulations are repeated 500 times and we measure the median absolute error, the empirical coverage proportion and the average length of the confidence interval computed across simulations.

5.2. Low dimensional setting

We first present the setting where the 50% rule holds. Specifically, following Windmeijer *et al.* (2016), Table 1 shows the cases where we have 10 IVs with $s_{z2} = 3$, $n = (500, 1000, 2000, 5000, 10000)$, $C_\gamma = (0.2, 0.6, 1)$ and $C_\pi = 0.2$. For reference, with $n = 500, 2000, 5000$ and $C_\gamma = 0.2$, the expected concentration parameter is $7nC_\gamma^2$, or, for each n, 140, 560 and 1400 respectively. Because the 50% rule condition holds, TSHT, the method of Windmeijer *et al.* (2016) and the median method should do well. Indeed, between TSHT and the method of Windmeijer *et al.* (2016), there is little difference in terms of median absolute error, coverage and length of the

Table 1. Comparison of methods when the 50% rule holds†

n	C_γ	*Results for the following methods:*														
		TSHT			*Adaptive lasso*			*Median*			*Naive TSLS*			*Oracle TSLS*		
500	0.2	0.09	0.72	0.32	0.09	0.72	0.32	0.13	0.17	0.09	0.30	0.03	0.28	0.08	0.96	0.44
500	0.6	0.02	0.84	0.11	0.03	0.81	0.11	0.05	0.31	0.06	0.10	0.01	0.09	0.03	0.96	0.15
500	1.0	0.02	0.83	0.07	0.02	0.78	0.07	0.03	0.71	0.07	0.06	0.02	0.06	0.01	0.95	0.09
1000	0.2	0.04	0.93	0.24	0.04	0.91	0.23	0.09	0.09	0.05	0.30	0.00	0.20	0.05	0.94	0.31
1000	0.6	0.01	0.95	0.08	0.01	0.93	0.08	0.03	0.27	0.03	0.10	0.00	0.07	0.02	0.96	0.10
1000	1.0	0.01	0.94	0.05	0.01	0.94	0.05	0.02	0.49	0.04	0.06	0.00	0.04	0.01	0.96	0.06
2000	0.2	0.03	0.93	0.17	0.03	0.93	0.17	0.07	0.08	0.02	0.30	0.00	0.14	0.04	0.94	0.22
2000	0.6	0.01	0.96	0.06	0.01	0.96	0.06	0.02	0.16	0.02	0.10	0.00	0.05	0.01	0.96	0.07
2000	1.0	0.01	0.95	0.03	0.01	0.95	0.03	0.01	0.33	0.02	0.06	0.00	0.03	0.01	0.97	0.04
5000	0.2	0.02	0.96	0.11	0.02	0.95	0.10	0.04	0.06	0.01	0.30	0.00	0.09	0.03	0.95	0.14
5000	0.6	0.01	0.96	0.04	0.01	0.96	0.03	0.01	0.12	0.01	0.10	0.00	0.03	0.01	0.95	0.05
5000	1.0	0.00	0.94	0.02	0.00	0.95	0.02	0.01	0.25	0.01	0.06	0.00	0.02	0.00	0.95	0.03
10000	0.2	0.01	0.97	0.08	0.01	0.97	0.07	0.03	0.04	0.00	0.30	0.00	0.06	0.02	0.95	0.10
10000	0.6	0.00	0.96	0.03	0.00	0.97	0.02	0.01	0.06	0.00	0.10	0.00	0.02	0.01	0.94	0.03
10000	1.0	0.00	0.94	0.02	0.00	0.95	0.01	0.01	0.16	0.00	0.06	0.00	0.01	0.00	0.95	0.02

†Adaptive lasso stands for the method of Windmeijer *et al.* (2016) using the median estimator and tuning with cross-validation. For each setting and method, say TSHT under $n = 500$ and $C_\gamma = 0.2$, the corresponding row of numbers (0.09, 0.72, 0.32) represents the median absolute error, the empirical coverage and the average length of the confidence interval.

confidence interval. Both of the methods struggle with low sample size at $n = 500$ but, once $n \geqslant 2000$, the two methods perform as well as the oracle. The median method does well with respect to median absolute error, but not as well as TSHT or the method of Windmeijer *et al.* (2016) and is not near oracle level performance. Also, we note that the confidence interval based on bootstrapping the median estimator is not a wise strategy in the one-sample setting. However, this is expected since the median estimator and the R package that implements it assume a two-sample setting with independent samples. Finally, the naive TSLS consistently has the worst performance across all simulation settings. For example, naive TSLS performs worse than TSHT even at $n = 500$.

Next, we present the setting where the 50% rule is violated, i.e. with $p_z = 7$ IVs where only three satisfy assumptions 1–3. The other parameters of the simulation remain the same, i.e. $n = (500, 1000, 2000, 5000, 10000)$, $C_\gamma = (0.2, 0.6, 1)$ and $C_\pi = 0.2$. We drop the median method from our comparison because of its poor performance in Table 1.

As expected, in Table 2, the adaptive lasso approach of Windmeijer *et al.* (2016) performs as badly as naive TSLS since the adaptive lasso depends on the 50% rule condition for consistency. In contrast, TSHT, which relies on a more general identifying plurality rule condition, has low error along with much better coverage than the adaptive lasso. Also, TSHT requires more samples to achieve the desired level of coverage when the data are generated under the identifying plurality rule condition than the 50% rule condition.

Overall, the simulation study shows that TSHT with voting performs no worse than the competing approaches in the literature. When the 50% rule condition holds, TSHT performs as well as the method that was proposed by Windmeijer *et al.* (2016). But, when the 50% rule condition fails to hold, but the plurality rule condition holds, TSHT performs much better with respect to absolute error, coverage and length of confidence intervals.

Table 2. Comparison of methods when the 50% rule is violated but the plurality rule holds†

n	C_γ	*Results for the following methods:*											
		TSHT			*Adaptive lasso*			*Naive TSLS*			*Oracle TSLS*		
500	0.2	0.37	0.17	0.38	0.35	0.18	0.39	0.41	0.00	0.33	0.09	0.97	0.51
500	0.6	0.11	0.24	0.13	0.13	0.17	0.13	0.14	0.00	0.11	0.03	0.93	0.17
500	1.0	0.07	0.21	0.08	0.08	0.18	0.08	0.09	0.00	0.07	0.02	0.94	0.10
1000	0.2	0.37	0.17	0.36	0.37	0.10	0.33	0.42	0.00	0.24	0.07	0.96	0.36
1000	0.6	0.09	0.32	0.13	0.12	0.12	0.11	0.14	0.00	0.08	0.02	0.96	0.12
1000	1.0	0.06	0.24	0.07	0.07	0.11	0.07	0.09	0.00	0.05	0.01	0.93	0.07
2000	0.2	0.19	0.45	0.32	0.44	0.01	0.27	0.42	0.00	0.17	0.05	0.95	0.25
2000	0.6	0.04	0.62	0.10	0.15	0.02	0.09	0.14	0.00	0.06	0.01	0.94	0.08
2000	1.0	0.03	0.55	0.06	0.09	0.02	0.05	0.09	0.00	0.03	0.01	0.94	0.05
5000	0.2	0.04	0.90	0.19	0.49	0.00	0.19	0.42	0.00	0.11	0.03	0.95	0.16
5000	0.6	0.01	0.91	0.06	0.17	0.00	0.06	0.14	0.00	0.03	0.01	0.94	0.05
5000	1.0	0.01	0.91	0.04	0.10	0.00	0.04	0.09	0.00	0.02	0.01	0.94	0.03
10000	0.2	0.02	0.92	0.13	0.50	0.00	0.14	0.43	0.00	0.07	0.02	0.96	0.11
10000	0.6	0.01	0.92	0.04	0.17	0.00	0.04	0.14	0.00	0.02	0.01	0.95	0.04
10000	1.0	0.00	0.94	0.03	0.10	0.00	0.03	0.09	0.00	0.01	0.00	0.94	0.02

†Adaptive lasso stands for the method of Windmeijer *et al.* (2016) using the median estimator and tuning with cross-validation. For each setting and method, say TSHT under $n = 500$ and $C_\gamma = 0.2$, the corresponding row of numbers (0.37, 0.17, 0.38) represents the median absolute error, the empirical coverage and the average length of the confidence interval.

5.3. High dimensional setting

In this section, we present simulations in high dimensions. We use the same data-generating models as before, except that we have $p_z = 100$ instruments with the first $s_{z1} = 7$ being relevant and the first $s_{z2} = 5$ being valid. We also have $p_x = 150$ covariates with $s_{x2} = s_{x1} = 10$. We refer to this case as the high dimensional instruments and covariates setting. We also consider $p_z = 9$ and $p_x = 150$, which we refer to as the low dimensional instruments and high dimensional covariates setting. The only difference between these two settings is the dimension of IVs. However, from a theoretical standpoint, both settings are considered high dimensional.

Both the instruments and the covariates $\mathbf{W}_{i\cdot}$ are generated from a multivariate normal with mean 0 and covariance $\Sigma^*_{ij} = 0.5^{|i-j|}$ for $1 \leqslant i, j \leqslant p_x + p_z$. The other parameters for the models are $\beta^* = 1$, $\phi^* = (0.6, 0.7, 0.8, \ldots, 1.5, 0, 0, \ldots, 0) \in \mathbb{R}^{150}$, $\psi^* = (1.1, 1.2, 1.3, \ldots, 2.0, 0, 0, \ldots, 0) \in \mathbb{R}^{150}$, $\mathrm{var}(\epsilon_{i1}) = \mathrm{var}(\epsilon_{i2}) = 1.5$ and $\mathrm{cov}(\epsilon_{i1}, \epsilon_{i2}) = 0.75$. We vary

(a) the sample size n,
(b) the strength of IV via $\boldsymbol{\gamma}^*$ and
(c) the degree of violations of assumptions 2 and 3 via $\boldsymbol{\pi}^*$.

For the sample size, we let $n = (200, 300, 1000, 2500)$. For the IV strength, we set $\boldsymbol{\gamma}^* = C_\gamma(1, 1, 1, 1, 1, 1, 0, 0, \ldots, 0)$ with $C_\gamma = 0.5$. For violations of assumptions 2 and 3, we set $\boldsymbol{\pi}^* = (0, 0, 0, 0, 0, 1, 1, 0, 0, \ldots, 0)C_\pi$ where C_π is a constant that we vary to change the magnitude of $\boldsymbol{\pi}^*$.

We compare TSHT with the oracle TSLS method where the oracle uses only the relevant and valid instruments, i.e. knows the seven relevant instruments, of which the first five are valid. We do not include the naive TSLS method because it is not feasible in high dimensions. We also do not include other methods because they were not designed with high dimensionality in mind. The high dimensional instruments and covariate setting is presented in Table 3 whereas the low dimensional instruments and high dimensional covariates setting is presented in Table 4.

Table 3. Performance of TSHT in high dimensional instruments and covariates with $p_X = 150$ and $p_Z = 100$†

n	C_π	*Results for TSHT*			*Results for oracle*		
200	0.25	0.162	0.162	0.202	0.038	0.956	0.219
200	0.50	0.129	0.448	0.232	0.036	0.962	0.218
200	1.00	0.056	0.876	0.259	0.036	0.956	0.221
300	0.25	0.155	0.080	0.164	0.033	0.952	0.179
300	0.50	0.093	0.516	0.197	0.029	0.952	0.177
300	1.00	0.041	0.906	0.209	0.029	0.946	0.176
1000	0.25	0.136	0.062	0.094	0.016	0.936	0.096
1000	0.50	0.020	0.942	0.119	0.016	0.936	0.095
1000	1.00	0.020	0.958	0.120	0.016	0.964	0.096
2500	0.25	0.015	0.802	0.068	0.011	0.946	0.060
2500	0.50	0.012	0.956	0.069	0.011	0.948	0.060
2500	1.00	0.011	0.954	0.069	0.010	0.942	0.060

†For each setting and method, say TSHT under $n = 200$ and $C_\pi = 0.25$, the row of numbers (0.162, 0.162, 0.202) represents the median absolute error, the empirical coverage and average length of the confidence interval.

To mimic the low dimensional results, Table 3 presents the result for $C_\pi = (0.25, 0.5, 1)$. In both settings, for $n = 200$, our TSHT method does not achieve the desired level of coverage, although coverage improves dramatically once the violation of assumptions 2 and 3 becomes bigger, i.e. when $C_\pi = 1$. When $n \geqslant 300$ and if the violations of assumptions 2 and 3 are substantial, TSHT achieves the desired level of coverage with absolute error and length of the confidence interval that are comparable with those of the oracle.

6. Application: causal effect of years of education on annual earnings

To demonstrate our method in a real setting, we analyse the causal effect of years of education on yearly earnings, which has been studied extensively in economics by using IV methods (Angrist and Krueger, 1991; Card, 1993, 1999). The data come from the WLS, which is 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) prepared by the WLS (see WLS documentation and Hauser (2005) for details) and, for the treatment, we use the total years of education, all from the 1974 survey. The median total earnings is \$9200 with a 25% quartile of \$1000 and a 75% quartile of \$15320 in 1974 dollars. The mean time of total education is 13.7 years with a standard deviation of 2.3 years.

We incorporate many covariates, including sex, graduate's home town population, educational attainment of the graduates' parents, graduates' family income, relative income in graduates' home town, graduates' high school denomination and 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, that were part of the WLS to control further for potential variations between graduates; see section D in the on-line supplementary materials for details on the non-genetic and genetic covariates. In summary, our data analysis includes seven non-genetic covariates and 81 genetic covariates. We used five instruments in our analysis, all

Table 4. Performance of TSHT in low dimension instruments ($p_Z = 9$) and high dimension covariates ($p_X = 150$)†

n	C_π	*Results for TSHT*			*Results for oracle*		
200	0.25	0.169	0.196	0.214	0.037	0.928	0.221
200	0.50	0.167	0.362	0.240	0.039	0.926	0.221
200	1.00	0.057	0.852	0.276	0.041	0.942	0.222
300	0.25	0.155	0.094	0.170	0.031	0.938	0.178
300	0.50	0.123	0.426	0.198	0.031	0.956	0.177
300	1.00	0.043	0.916	0.222	0.030	0.960	0.177
1000	0.25	0.133	0.076	0.090	0.015	0.944	0.095
1000	0.50	0.019	0.962	0.113	0.016	0.954	0.096
1000	1.00	0.020	0.958	0.113	0.016	0.950	0.095
2500	0.25	0.012	0.860	0.067	0.009	0.948	0.060
2500	0.50	0.012	0.952	0.068	0.010	0.950	0.060
2500	1.00	0.012	0.958	0.068	0.011	0.944	0.060

†For each setting and method, say TSHT under $n = 200$ and $C_\pi = 0.25$, the row of numbers (0.169, 0.196, 0.214) represents the median absolute error, the empirical coverage and the average length of the confidence interval.

derived from past studies of education on earnings (Card, 1993; Blundell *et al.*, 2005; Gary-Bobo *et al.*, 2006). They are

(a) total number of sisters,
(b) total number of brothers,
(c) individuals, birth order in the family, all from Gary-Bobo *et al.* (2006),
(d) proximity to college from Card (1993) and
(e) teachers' 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 why they satisfy assumptions 2 and 3 after controlling for the aforementioned covariates, in practice, we are always uncertain because of the lack of complete socio-economic knowledge about the effect of these IVs. Our method should provide some protection against this uncertainty compared with 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 a very strong set of instruments. For more details on the instruments, see section D of the on-line supplementary materials.

When we use OLS where we run a regression of the treatment and the covariates on the outcome and looking at the slope coefficient of the treatment variable, we find the effect estimate to be 0.097 (95% confidence interval 0.051, 0.143). 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. TSLS provides an alternative method of controlling for confounding by using instruments so long as all the five instruments satisfy the three core assumptions and the inclusion of covariates helps to make these assumptions more plausible. The TSLS estimate is 0.169 (95% confidence interval 0.029, 0.301), which is inconsistent with previous studies' estimates among individuals from the USA between the 1950s and the 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 TSLS, provides an estimate of 0.062 (95% confidence interval 0.046, 0.077), 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 more accurate estimates of the returns on education than does TSLS, which naively assumes that all the instruments are valid.

7. Conclusion and discussion

We present a method to estimate the effect of the treatment on the outcome by using IVs where we do not make the assumption that all the instruments are valid. Our approach is based on the novel TSHT procedure with majority and plurality voting. We theoretically show that our approach succeeds in selecting valid IVs in the presence of possibly invalid IVs even when the 50% rule is violated and produces robust confidence intervals. In simulation and in real data settings, our approach provides a more robust analysis than the traditional IV approaches or recent methods in the invalid IV literature by providing some protection against possibly invalid instruments and reaches oracle performance around $n \geqslant 2000$. Overall, we believe that our method can be a valuable tool for researchers in Mendelian randomization and IVs whenever there are concerns for invalid IVs, which is often the case in practice.

Finally, our theoretical analysis for the case of invalid IVs in high dimensions require assumptions 8 and 9. We believe that assumption 9 is probably necessary for the invalid IV problem in high dimensions because of the model selection literature by Leeb and Pötscher (2005) who pointed out that 'in general no model selector can be uniformly consistent for the most parsimonious true model' and hence that the post-model-selection inference is generally non-uniform. Consequently, the set of competing models must be 'well separated' such that we can consistently select a correct model. Assumption 9 serves as this 'well-separated' condition in our invalid IV problem. Although some recent work in high dimensional inference (Zhang and Zhang, 2014; Javanmard and Montanari, 2014; van de Geer *et al.*, 2014; Chernozhukov *et al.*, 2015; Cai and Guo, 2017) does not make this well-separated assumption, our invalid IV problem is different from the prior work because a single invalid IV declared as valid can ruin inference whereas the said prior works assume that the moment conditions are known perfectly. Advanced methods may weaken assumption 9 and we leave this as a direction for further research.

Acknowledgements

The research of Hyunseung Kang was supported in part by National Science Foundation grant DMS-1502437. The research of T. Tony Cai was supported in part by National Science Foundation grants DMS-1208982 and DMS-1403708, and National Institutes of Health grant R01 CA127334. The research of Dylan S. Small was supported in part by National Science Foundation grant SES-1260782.

References

Andrews, D. W. K., Moreira, M. J. and Stock, J. H. (2007) Performance of conditional Wald tests in IV regression with weak instruments. *J. Econmetr.*, **139**, 116–132.

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996) Identification of causal effects using instrumental variables. *J. Am. Statist. Ass.*, **91**, 444–455.

Angrist, J. D. and Krueger, A. B. (1991) Does compulsory school attendance affect schooling and earnings? *Q. J. Econ.*, **106**, 979–1014.

Baiocchi, M., Cheng, J. and Small, D. S. (2014) Instrumental variable methods for causal inference. *Statist. Med.*, **33**, 2297–2340.

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

Belloni, A., Chernozhukov, V. and Hansen, C. (2014) Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.*, **81**, 608–650.
Belloni, A., Chernozhukov, V. and Wang, L. (2011) Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, **98**, 791–806.
Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009) Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, **37**, 1705–1732.
Blundell, R., Dearden, L. and Sianesi, B. (2005) Evaluating the effect of education on earnings: models, methods and results from the National Child Development Survey. *J. R. Statist. Soc.* A, **168**, 473–512.
Bowden, J., Davey Smith, G. and Burgess, S. (2015) Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *Int. J. Epidem.*, **44**, 512–525.
Bowden, J., Davey Smith, G., Haycock, P. C. and Burgess, S. (2016) Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidem.*, **40**, 304–314.
Bühlmann, P. and van de Geer, S. (2011) *Statistics for High-dimensional Data: Methods, Theory and Applications*. Berlin: Springer.
Burgess, S., Bowden, J., Dudbridge, F. and Thompson, S. G. (2016) Robust instrumental variable methods using multiple candidate instruments with application to Mendelian randomization. *arXiv Preprint*. University of Cambridge, Cambridge.
Burgess, S., Timpson, N. J., Ebrahim, S. and Davey Smith, G. (2015) Mendelian randomization: where are we now and where are we going? *Int. J. Epidem.*, **44**, 379–388.
Cai, T. T. and Guo, Z. (2017) Confidence intervals for high-dimensional linear regression: minimax rates and adaptivity. *Ann. Statist.*, **45**, 615–646.
Card, D. (1993) Using geographic variation in college proximity to estimate the return to schooling. *Working Paper 4483*. National Bureau of Economic Research, Cambridge.
Card, D. (1999) The causal effect of education on earnings. In *Handbook of Labor Economics* (eds O. C. Ashenfelter and D. Card), vol. 3, part A, ch. 30, pp. 1801–1863. New York: Elsevier.
Cheng, X. and Liao, Z. (2015) Select the valid and relevant moments: an information-based lasso for GMM with many moments. *J. Econmetr.*, **186**, 443–464.
Chernozhukov, V., Hansen, C. and Spindler, M. (2015) Post-selection and post-regularization inference in linear models with many controls and instruments. *Am. Econ. Rev.*, **105**, 486–490.
Davey Smith, G. and Ebrahim, S. (2003) Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidem.*, **32**, 1–22.
Davey Smith, G. and Ebrahim, S. (2004) Mendelian randomization: prospects, potentials, and limitations. *Int. J. Epidem.*, **33**, 30–42.
Donoho, D. L. (1995) De-noising by soft-thresholding. *IEEE Trans. Inform. Theory*, **41**, 613–627.
Donoho, D. L. and Johnstone, J. M. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.
Fan, J. and Liao, Y. (2014) Endogeneity in high dimensions. *Ann. Statist.*, **42**, 872–917.
Gary-Bobo, R., Picard, N. and Prieto, A. (2006) Birth order and sibship sex composition as instruments in the study of education and earnings. *Discussion Paper 5514*. Centre for Economic and Policy Research, London.
Gautier, E. and Tsybakov, A. B. (2011) High-dimensional instrumental variables regression and confidence sets. *Preprint arXiv:1105.2454*. Center for Research in Economics and Statistics, Malakoff.
van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014) On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, **42**, 1166–1202.
Han, C. (2008) Detecting invalid instruments using l 1-gmm. *Econ. Lett.*, **101**, 285–287.
Hartwig, F. P., Davey Smith, G. and Bowden, J. (2017) Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int. J. Epidem.*, **46**, 1985–1998.
Hastie, T., Tibshirani, R. and Friedman, J. (2016) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 3rd edn. New York: Springer.
Hauser, R. M. (2005) Survey response in the long run: the Wisconsin longitudinal study. *Fld Meth.*, **17**, 3–29.
Hernán, M. A. and Robins, J. M. (2006) Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, **17**, 360–372.
Holland, P. W. (1988) Causal inference, path analysis, and recursive structural equations models. *Sociol. Methodol.*, **18**, 449–484.
Imbens, G. W. (2014) Instrumental variables: an econometrician's perspective. *Statist. Sci.*, **29**, 323–358.
Imbens, G. W. and Angrist, J. D. (1994) Identification and estimation of local average treatment effects. *Econometrica*, **62**, 467–475.
Javanmard, A. and Montanari, A. (2014) Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.*, **15**, 2869–2909.
Kang, H., Cai, T. T. and Small, D. S. (2016a) A simple and robust confidence interval for causal effects with possibly invalid instruments. *Preprint arXiv:1504.03718*. Department of Statistics, University of Wisconsin—Madison, Madison.
Kang, H., Zhang, A., Cai, T. T. and Small, D. S. (2016b) Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *J. Am. Statist. Ass.*, **111**, 132–144.
Kolesár, M., Chetty, R., Friedman, J. N., Glaeser, E. L. and Imbens, G. W. (2015) Identification and inference with many invalid instruments. *J. Bus. Econ. Statist.*, **33**, 474–484.

Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N. and Davey Smith, G. (2008) Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statist. Med.*, **27**, 1133–1163.
Leeb, H. and Pötscher, B. M. (2005) Model selection and inference: facts and fiction. *Econmetr. Theory*, **21**, 21–59.
Moreira, M. J. (2003) A conditional likelihood ratio test for structural models. *Econometrica*, **71**, 1027–1048.
Murray, M. P. (2006) Avoiding invalid instruments and coping with weak instruments. *J. Econ. Perspect.*, **20**, 111–132.
Neyman, J. (1923) On the application of probability theory to agricultural experiments: Essay on principles, section 9. *Statist. Sci.*, **5**, 465–472.
Roetker, N. S., Yonker, J. A., Lee, C., Chang, V., Basson, J. J., Roan, C. L., Hauser, T. S., Hauser, R. M. and Atwood, C. S. (2012) Multigene interactions and the prediction of depression in the Wisconsin Longitudinal Study. *Br. Med. J. Open*, **2**, article e000944.
Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, **66**, 688–701.
Small, D. S. (2007) Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *J. Am. Statist. Ass.*, **102**, 1049–1058.
Stock, J. H., Wright, J. H. and Yogo, M. (2002) A survey of weak instruments and weak identification in generalized method of moments. *J. Bus. Econ. Statist.*, **20**, 518–529.
Swanson, S. A. and Hernán, M. A. (2013) Commentary: How to report instrumental variables analyses (suggestions welcome). *Epidemiology*, **24**, 370–374.
Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.
Windmeijer, F., Farbmacher, H., Davies, N. and Davey Smith, G. (2016) On the use of the lasso for instrumental variables estimation with some invalid instruments. *Technical Report*. Department of Economics, University of Bristol, Bristol.
Wooldridge, J. M. (2010) *Econometric Analysis of Cross Section and Panel Data*, 2nd edn. Cambridge: MIT Press.
Yavorska, O. O. and Burgess, S. (2017) Mendelianrandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int. J. Epidem.*, **46**, 1734–1739.
Zhang, C.-H. and Zhang, S. S. (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Statist. Soc.* B, **76**, 217–242.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary materials: identification conditions'.