Received: 14 June 2019 | Accepted: 13 January 2021

DOI: 10.1111/rssb.12449

ROYAL STATISTICAL SOCIETY
Series B Statistical Methodology

ORIGINAL ARTICLE

# The confidence interval method for selecting valid instrumental variables

**Frank Windmeijer**[1,2] | **Xiaoran Liang**[3] | **Fernando P. Hartwig**[2,4] | **Jack Bowden**[2,5]

[1]Department of Statistics and Nuffield College, University of Oxford, Oxford, UK

[2]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK

[3]Department of Economics, University of Bristol, Bristol, UK

[4]Center for Epidemiological Research, University of Pelotas, Pelotas, Brazil

[5]College of Medicine and Health, University of Exeter, Exeter, UK

**Correspondence**
Frank Windmeijer, Department of Statistics, University of Oxford, 24-29 St Giles', Oxford, UK.
Email: frank.windmeijer@stats.ox.ac.uk

**Funding information**
Economic and Social Research Council, Grant/Award Number: ES/P000630/1; Medical Research Council, Grant/Award Number: MC_UU_00011/2

## Abstract

We propose a new method, the confidence interval (CI) method, to select valid instruments from a larger set of potential instruments for instrumental variable (IV) estimation of the causal effect of an exposure on an outcome. Invalid instruments are such that they fail the exclusion conditions and enter the model as explanatory variables. The CI method is based on the CIs of the per instrument causal effects estimates and selects the largest group with all CIs overlapping with each other as the set of valid instruments. Under a plurality rule, we show that the resulting standard IV, or two-stage least squares (2SLS) estimator has oracle properties. This result is the same as for the hard thresholding with voting (HT) method of Guo et al. (*Journal of the Royal Statistical Society : Series B*, 2018, *80*, 793–815). Unlike the HT method, the number of instruments selected as valid by the CI method is guaranteed to be monotonically decreasing for decreasing values of the tuning parameter. For the CI method, we can therefore use a downward testing procedure based on the Sargan (*Econometrica*, 1958, *26*, 393–415) test for overidentifying restrictions and a main advantage of the CI downward testing method is that it selects the model with the largest number of instruments selected as valid that passes the Sargan test.

**KEYWORDS**
causal inference, instrumental variables, invalid instruments

# 1 | INTRODUCTION

Instrumental variables (IV) estimation is a well-established method for determining causal effects of an exposure on an outcome, when this relationship is potentially affected by unobserved confounding. For recent reviews and examples, see Clarke and Windmeijer (2012), Imbens (2014), Kang et al. (2016) and Burgess et al. (2017).

As Guo et al. (2018, p. 793) state, an IV analysis requires instruments that

1. are associated with the exposure (Condition 1),
2. have no direct pathway to the outcome (Condition 2) and
3. are not related to unmeasured variables that affect the exposure and the outcome (Condition 3).

Condition 1 is often referred to as the *relevance* condition and Conditions 2 and 3 as the *exclusion* conditions, see Section 2 for details.

This paper is concerned with violations of the exclusion conditions of the instruments. Following closely the setup of Kang et al. (2016), Windmeijer et al. (2019) and Guo et al. (2018), if an instrument satisfies the exclusion Conditions 2 and 3 it is classified as a valid instrument. If an instrument does not satisfy Condition 2 and/or 3, it is classified as invalid. Use of invalid instruments in an IV analysis leads to inconsistent estimates of the causal effect and it is therefore important to select the set of valid instruments from the set of putative IVs that may include invalid ones.

As an example, Mendelian randomisation is a technique employed in epidemiology to learn about the causal effects of modifiable health exposures on disease. It posits that genetic variants, which are known to be associated with the exposure and hence satisfy Condition 1, additionally satisfy the exclusion conditions and are only associated with the outcome through the exposure. In our Mendelian randomisation application in Section 8, we utilise genetic variants as potential instruments for BMI in order to determine its causal effect on diastolic blood pressure. However, a genetic variant could be an invalid instrument for various reasons, such as linkage disequilibrium and horizontal pleiotropy, see, for example, Lawlor et al. (2008) and von Hinke et al. (2016).

The so-called plurality rule holds if the set of valid instruments forms the largest group, as specified in Section 2. An approach for selecting the valid instruments could then be to follow Andrews (1999) and estimate the causal effect for all $2^{k_z} - \left(k_z + 1\right)$ possible subsets of at least two instruments, where $k_z$ denotes the total number of instruments, and to select the model that minimises an information criterion based on the Sargan (1958) test of overidentifying restrictions. A large value of the Sargan test statistic is an indication that invalid instruments are present. This approach is only feasible with a relatively small number of instruments, unlike in our application where we have 96 putative genetic instruments. We therefore need dimension reduction techniques, even though we are in a setting of a fixed number of instruments $k_z$ with a large sample size $n$, the setting referred to as low dimensional by Guo et al. (2018).

Following the Lasso proposal by Kang et al. (2016), Windmeijer et al. (2019) proposed an adaptive Lasso estimator in combination with a downward testing procedure based on the Sargan test as in Andrews (1999). When the majority rule holds, meaning that more than 50% of the potential instruments are valid, then this approach results in consistent selection of the invalid instruments and oracle properties of the resulting standard IV, or two-stage least squares (2SLS)

estimator. This means that the limiting distribution of the estimator is the same as the oracle estimator, which is the 2SLS estimator when the set of invalid instruments is known. Guo et al. (2018) proposed a two-stage hard thresholding with voting (HT) method that results in consistent selection of the valid instruments and oracle properties of the 2SLS estimator when the weaker plurality rule holds.

In this paper, we develop an alternative method, which we call the confidence interval (CI) method as presented in Section 3. This method simply selects as valid instruments the largest group of instruments where all CIs of the instrument-specific causal effect estimates overlap, with a tuning parameter varying the width of the CIs. Like the Guo et al. (2018) method, we show that the CI method results in consistent selection and oracle properties of the resulting 2SLS estimator when the plurality rule holds. An advantage of the CI method is that the number of instruments selected as valid decreases monotonically for decreasing values of the tuning parameter, which is not the case for the HT method as we discuss in Section 4. For the CI method, we can therefore use a downward testing procedure based on the Sargan test and a main advantage of this CI method is that it selects the model with the largest number of instruments selected as valid that passes the Sargan test.

While initially making the assumptions of conditional homoskedasticity and strong instruments in Section 2 for ease of exposition, we discuss in Section 5 how to adapt the methods to deal with general forms of heteroskedasticity. We further discuss the first-stage thresholding method of Guo et al. (2018) to dealing with weak instruments in Section 6.

We evaluate the two methods in the Monte Carlo exercise in Section 7, for a design very similar to that in Guo et al. (2018). We find that, overall, the CI method has a better finite sample performance than the HT method in this design. In the application in Section 8, we find that the HT method selects too few instruments as invalid, resulting in models that are rejected by the Sargan test. By design, the CI method selects models that pass the Sargan test. It produces results very similar to the adaptive Lasso method which suggests that the majority rule is not violated in this application.

We adopt the following notation. $\boldsymbol{x}$ denotes the vector with elements $x_j$. For a general matrix $\boldsymbol{X}$, $\boldsymbol{X}'$ denotes its transpose. All vectors are taken as column vectors, including $\mathbf{X}_{i.}$, where the row vector $\mathbf{X}_{i.}'$ is the $i$th row of the matrix $\boldsymbol{X}$. For a full column-rank matrix $\mathbf{X}$ with $n$ rows define $\mathbf{P}_X = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$, the projection onto the column space of $\mathbf{X}$, and $\mathbf{M}_X = \mathbf{I}_n - \mathbf{P}_X$, where $\mathbf{I}_n$ is the $n$-dimensional identity matrix. Proofs of Lemma 1 and Theorems 3 and 4 in Section 3 are presented in the Supplementary Appendix A.1.

## 2 | MODEL AND ASSUMPTIONS

Let the observed outcome for observation $i$ be denoted by the scalar $Y_i$, the treatment or exposure by the scalar $D_i$ and the vector of $k_z$ potential instruments by $\mathbf{Z}_{i.}$. The instruments may not all be valid and can have a direct effect on, and/or an indirect association with the outcome, violating Condition 2 and/or 3. We have a sample $\left\{Y_i, D_i, \mathbf{Z}_{i.}'\right\}_{i=1}^{n}$. We follow Kang et al. (2016) and Guo et al. (2018), who, starting from the additive linear, constant effects model of Holland (1988), arrived at the observed data model for the sample given by

$$Y_i = D_i\beta + \mathbf{Z}_{i.}'\boldsymbol{\alpha} + u_i, \tag{1}$$

where $\beta$ is the causal parameter of interest, and with $E\left[u_i|\mathbf{Z}_{i.}\right] = 0$, but $D_i$ might be correlated with $u_i$. The parameter vector $\boldsymbol{\alpha}$ represents the possible violations of the exclusion conditions and can be used to formalise the definition of valid IVs as follows (Guo et al. 2018, p 797).

**Definition 1** If $\alpha_j = 0$, then instrument $j$, $j = 1, \ldots, k_z$, is valid, it satisfies both Conditions 2 and 3. If $\alpha_j \neq 0$, then instrument $j$ is invalid.

We present some graphical representations of the causal model and possible violations of the exclusion conditions in Appendix A.3.

Let $\mathbf{y}$ and $\mathbf{d}$ be the $n$-vectors of $n$ observations on $\{Y_i\}$ and $\{D_i\}$, respectively, and let $\mathbf{Z}$ be the $n \times k_z$ matrix of potential instruments. As an intercept is implicitly present in the model, $\mathbf{y}$, $\mathbf{d}$ and the columns of $\mathbf{Z}$ have all been centered by the subtraction of their means. Other covariates can be partialled out in the same way. Let $\mathbf{Z}_{\mathcal{V}_0}$ and $\mathbf{Z}_{\mathcal{A}_0}$ be the sets of valid and invalid instruments, $\mathcal{V}_0 = \{j: \alpha_j = 0\}$, $\mathcal{A}_0 = \{j: \alpha_j \neq 0\}$, with dimensions $k_{\mathcal{V}_0}$ and $k_{\mathcal{A}_0}$, respectively, and $k_z = k_{\mathcal{V}_0} + k_{\mathcal{A}_0}$. $\mathcal{V} = \{1, \ldots, k_z\}$ denotes the full set and so $\mathcal{A}_0 = \mathcal{V} \backslash \mathcal{V}_0$.

The oracle model is then given by

$$\mathbf{y} = \mathbf{d}\beta + \mathbf{Z}_{\mathcal{A}_0}\boldsymbol{\alpha}_{\mathcal{A}_0} + \mathbf{u}. \tag{2}$$

Let $\widehat{\mathbf{d}} = \mathbf{P}_Z\mathbf{d}$, then the oracle 2SLS estimator for $\beta$ is the OLS estimator in the specification

$$\mathbf{y} = \widehat{\mathbf{d}}\beta + \mathbf{Z}_{\mathcal{A}_0}\boldsymbol{\alpha}_{\mathcal{A}_0} + \xi,$$

where $\xi$ is defined implicitly, and is given by

$$\widehat{\beta}_{or} = \left(\widehat{\mathbf{d}}'\mathbf{M}_{Z_{\mathcal{A}_0}}\widehat{\mathbf{d}}\right)^{-1}\widehat{\mathbf{d}}'\mathbf{M}_{Z_{\mathcal{A}_0}}\mathbf{y}. \tag{3}$$

Under standard assumptions, as detailed below, and as $n \to \infty$,

$$\sqrt{n}\left(\widehat{\beta}_{or} - \beta\right) \xrightarrow{d} N\left(0, \sigma^2_{\beta_{or}}\right), \tag{4}$$

where

$$\begin{aligned}\sigma^2_{\beta_{or}} &= \sigma^2_u\left(\operatorname{plim}\left(\frac{1}{n}\widehat{\mathbf{d}}'\mathbf{M}_{Z_{\mathcal{A}_0}}\widehat{\mathbf{d}}\right)^{-1}\right),\\ &= \sigma^2_u\left(E\left[\mathbf{Z}_{i.}D_i\right]'E\left[\mathbf{Z}_{i.}\mathbf{Z}'_{i.}\right]^{-1}E\left[\mathbf{Z}_{i.}D_i\right] - E\left[\mathbf{Z}_{\mathcal{A}_0,i.}D_i\right]'E\left[\mathbf{Z}_{\mathcal{A}_0,i.}\mathbf{Z}'_{\mathcal{A}_0,i.}\right]^{-1}E\left[\mathbf{Z}_{\mathcal{A}_0,i.}D_i\right]\right)^{-1},\end{aligned} \tag{5}$$

see Appendix A.2 for a derivation.

The vector $\widehat{\mathbf{d}} = \mathbf{P}_Z\mathbf{d} = \mathbf{Z}\widehat{\boldsymbol{\gamma}}$ is the linear projection of $\mathbf{d}$ on $\mathbf{Z}$, with $\widehat{\boldsymbol{\gamma}}$ the OLS estimator of $\boldsymbol{\gamma} = E\left[\mathbf{Z}_{i.}\mathbf{Z}'_{i.}\right]^{-1}E\left[\mathbf{Z}_{i.}D_i\right]$ in the linear model specification

$$D_i = \mathbf{Z}'_{i.}\boldsymbol{\gamma} + \varepsilon_{di}, \tag{6}$$

with $E\left[\mathbf{Z}_{i.}\varepsilon_{di}\right] = 0$. We initially assume that all instruments satisfy Condition 1, implying that the $k_z$ elements $\gamma_j$ in $\boldsymbol{\gamma}$, are all different from 0:

**Assumption 1** $\boldsymbol{\gamma} = \left(E\left[\mathbf{Z}_{i.}\mathbf{Z}'_{i.}\right]\right)^{-1}E\left[\mathbf{Z}_{i.}D_i\right]$, $\gamma_j \neq 0$, $j = 1, \ldots, k_z$.

This is the same assumption as in Kang et al. (2016) and Windmeijer et al. (2019). Guo et al. (2018) relaxed this assumption and proposed a first-stage hard thresholding procedure to consistently select only instruments with $\gamma_j \neq 0$. We will discuss this further in Section 6 and apply this first-stage thresholding in our application.

Let $\Gamma = E\left[\mathbf{Z}_{i.}\mathbf{Z}_{i.}'\right]^{-1} E\left[\mathbf{Z}_{i.}Y_i\right]$. As $Y_i = D_i\beta + \mathbf{Z}_{i.}'\boldsymbol{\alpha} + u_i = \mathbf{Z}_{i.}'\boldsymbol{\gamma}\beta + \mathbf{Z}_{i.}'\boldsymbol{\alpha} + u_i + \varepsilon_{di}\beta$, it follows that $\Gamma = \boldsymbol{\gamma}\beta + \boldsymbol{\alpha}$. Then define $\beta_j$ as

$$\beta_j \equiv \frac{\Gamma_j}{\gamma_j} = \beta + \frac{\alpha_j}{\gamma_j}, \tag{7}$$

for $j = 1, \ldots, k_z$. It follows from Definition 1 and Assumption 1 that for valid instruments, $j \in \mathcal{V}_0$, $\beta_j = \beta$. Following Theorem 1 in Kang et al. (2016) and Guo et al. (2018), a necessary and sufficient condition to identify $\beta$ and the $\alpha_j$, given $\Gamma$ and $\boldsymbol{\gamma}$, is that the valid instruments form the largest group, where instruments form a group if they have the same value for $\beta_j$. This is the plurality rule. As in Guo et al. (2018), we maintain the assumption that this condition is satisfied:

**Assumption 2** $\left|\mathcal{V}_0\right| > \max_{g \neq 0}\left|\mathcal{V}_g\right|$, where $\mathcal{V}_g = \left\{j: \frac{\alpha_j}{\gamma_j} = g\right\}$.

For the sample $\left\{Y_i, D_i, \mathbf{Z}_{i.}'\right\}_{i=1}^n$, and models (1) and (6), we further assume that the following standard conditions hold:

**Assumption 3** $E\left[\mathbf{Z}_{i.}\mathbf{Z}_{i.}'\right] = \mathbf{Q}$, with $\mathbf{Q}$ a finite and full rank matrix.

**Assumption 4** Let $\mathbf{w}_i = \left(u_i \ \varepsilon_{di}\right)'$. Then $E\left[\mathbf{w}_i\right] = 0$; $E\left[\mathbf{w}_i\mathbf{w}_i'\right] = \begin{bmatrix} \sigma_u^2 & \sigma_{u\varepsilon_d} \\ \sigma_{u\varepsilon_d} & \sigma_{\varepsilon_d}^2 \end{bmatrix} = \boldsymbol{\Sigma}$. The elements of $\boldsymbol{\Sigma}$ are finite.

**Assumption 5** $\operatorname{plim}\left(n^{-1}\mathbf{Z}'\mathbf{Z}\right) = E\left[\mathbf{Z}_{i.}\mathbf{Z}_{i.}'\right] = \mathbf{Q}$; $\operatorname{plim}\left(n^{-1}\mathbf{Z}'\mathbf{d}\right) = E\left[\mathbf{Z}_{i.}D_i\right]$;
$\operatorname{plim}\left(n^{-1}\mathbf{Z}'\mathbf{u}\right) = E\left[\mathbf{Z}_{i.}u_i\right] = 0$; $\operatorname{plim}\left(n^{-1}\mathbf{Z}'\varepsilon_d\right) = E\left[\mathbf{Z}_{i.}\varepsilon_{di}\right] = 0$;
$\operatorname{plim}\left(n^{-1}\sum_{i=1}^n \mathbf{w}_i\right) = 0$; $\operatorname{plim}\left(n^{-1}\sum_{i=1}^n \mathbf{w}_i\mathbf{w}_i'\right) = \Sigma$.

**Assumption 6** $\frac{1}{\sqrt{n}}\sum_{i=1}^n vec\left(\mathbf{Z}_{i.}\mathbf{w}_i'\right) \to N\left(0, \boldsymbol{\Sigma} \otimes \boldsymbol{Q}\right)$ as $n \to \infty$.

While Assumption 5 holds if the observations are i.i.d., as the moments are assumed to exist, these conditions further hold under various weak dependence assumptions, see Staiger and Stock (1997, p. 560).

Note that conditional homoskedasticity $E\left[\mathbf{w}_i\mathbf{w}_i'|\mathbf{Z}_{i.}\right] = \boldsymbol{\Sigma}$ is implicit in Assumption 6. We make this assumption primarily for ease of exposition and will relax this in Section 5.

The plurality rule, Assumption 2, is the main assumption on the instruments needed to establish oracle properties for the CI method described below and the HT method of Guo et al. (2018). In particular, the values of $\alpha_j$ and $\gamma_j$ can be arbitrary and arbitrarily correlated. The CI and HT methods are robust to any such correlation. Alternatively, the methods of Kolesár et al. (2015) and Bowden, Smith and Burgess (2015) do not make the plurality assumption and can have all instruments invalid. A bias corrected 2SLS estimator is then consistent under the INstrument Strength Independent of Direct Effect (INSIDE) assumption that $Cov\left(\alpha_j, \gamma_j\right) = 0$, together with the requirement that the number of instruments increases with the sample size. Guo et al. (2018) provide a discussion of and comparison to these methods, also including alternative methods proposed by Bowden et al. (2016), Hartwig, Smith and Bowden (2017) and Burgess et al. (2018).

## 3 | THE CONFIDENCE INTERVAL METHOD

From the plurality rule Assumption 2, it follows that consistent instrument selection procedures can be based on consistent and asymptotic normal estimators of the parameters $\beta_j$ as defined in (7). Then groups of instruments are formed by similar estimates $\widehat{\beta}_j$, and, in large samples, the largest group will constitute the group of valid instruments under Assumption 2. While in principle all combinations of instruments could be tested separately, see Andrews (1999), in practice this may not be feasible when there are a large number of instruments. The Guo et al. (2018) method as described further in Section 4 reduces the dimensionality of the problem by essentially performing $k_z\left(k_z-1\right)/2$ pairwise tests of the null $H_0\colon \beta_j=\beta_k$, combined with a voting scheme to group the instruments.

A clear reduction of the dimensionality of the problem is achieved by alternatively considering testing $H_0\colon \beta_j=\delta_g$, for a grid $\delta_g$ spanning the possible values of $\beta$ and selecting as the set of valid instruments the largest set over all values of $\delta_g$ for which a particular value of $\delta_g$ is not rejected. The CI method operationalises this idea without having to consider the grid points $\delta_g$ by grouping together instruments with overlapping CIs.

Let $\widehat{\Gamma}$ and $\widehat{\boldsymbol{\gamma}}$ be the OLS estimators for $\Gamma$ and $\boldsymbol{\gamma}$ in the model specifications

$$\mathbf{y}=\mathbf{Z}\Gamma+\boldsymbol{\varepsilon}_y;\quad \mathbf{d}=\mathbf{Z}\boldsymbol{\gamma}+\boldsymbol{\varepsilon}_d.$$

Under Assumptions 3-6 it follows that

$$\sqrt{n}\left(\begin{pmatrix}\widehat{\Gamma}\\ \widehat{\boldsymbol{\gamma}}\end{pmatrix}-\begin{pmatrix}\Gamma\\ \boldsymbol{\gamma}\end{pmatrix}\right)\xrightarrow{d}N\left(0,\boldsymbol{\Lambda}\right), \tag{8}$$

where $\boldsymbol{\Lambda}=\boldsymbol{\Omega}\otimes\mathbf{Q}^{-1}$, with $\boldsymbol{\Omega}=E\left[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'|\mathbf{Z}_{i.}\right]$, $\boldsymbol{\varepsilon}_i=\left(\varepsilon_{yi},\varepsilon_{di}\right)'$.

Following Guo et al. (2018), let an estimator for $\beta_j$ be

$$\widehat{\beta}_j=\frac{\widehat{\Gamma}_j}{\widehat{\gamma}_j}, \tag{9}$$

then it follows, using the delta method, that $\sqrt{n}\left(\widehat{\beta}_j-\beta_j\right)\xrightarrow{d}N\left(0,\sigma_j^2\right)$, with, denoting $\mathbf{Q}_{jj}^{-1}$ the $j$th diagonal element of $\mathbf{Q}^{-1}$,

$$\sigma_j^2=\frac{\tau_j^2\mathbf{Q}_{jj}^{-1}}{\gamma_j^2};\quad \tau_j^2=\left(1-\beta_j\right)\boldsymbol{\Omega}\begin{pmatrix}1\\ -\beta_j\end{pmatrix}. \tag{10}$$

An estimator for the variance of $\widehat{\beta}_j$ is then given by

$$\widehat{Var}\left(\widehat{\beta}_j\right)=\frac{\widehat{\tau}_j^2\left(\mathbf{Z}'\mathbf{Z}\right)_{jj}^{-1}}{\widehat{\gamma}_j^2};\quad \widehat{\tau}_j^2=\left(1-\widehat{\beta}_j\right)\widehat{\boldsymbol{\Omega}}\begin{pmatrix}1\\ -\widehat{\beta}_j\end{pmatrix}, \tag{11}$$

where $\widehat{\boldsymbol{\Omega}}=\frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{\varepsilon}}_i\widehat{\boldsymbol{\varepsilon}}_i'$, with $\widehat{\boldsymbol{\varepsilon}}_i$ the OLS residual vector $\left(\widehat{\varepsilon}_{yi},\widehat{\varepsilon}_{di}\right)'$. It follows that $n\widehat{Var}\left(\widehat{\beta}_j\right)\xrightarrow{p}\sigma_j^2$.

We show in Appendix A.5 that $\widehat{\beta}_j$ is identical to the 2SLS estimator of $\beta_j$ in the just-identified model

$$\mathbf{y} = \mathbf{d}\beta_j + \mathbf{Z}_{\{-j\}}\boldsymbol{\pi}^{[j]} + \mathbf{u}_j, \tag{12}$$

where $\mathbf{Z}_{\{-j\}} = \mathbf{Z} \setminus \{\mathbf{Z}_j\}$, using $\mathbf{Z}_j$ as the instrument for $\mathbf{d}$. This therefore implies that $\widehat{\beta}_j$ is the IV estimator for $\beta$ based on instrument $\mathbf{Z}_j$ while treating all other instruments as invalid. The variance estimator $\widehat{Var}\left(\widehat{\beta}_j\right)$ as defined in (11) is also the same as the standard 2SLS variance estimator in the just-identified model (12).

The CI method is a fast method that consistently selects the valid instruments. Let $\widehat{v}_j = \sqrt{\widehat{Var}\left(\widehat{\beta}_j\right)}$. Given a value $\psi_n$, define the confidence interval $ci_j\left(\psi_n\right)$ for $\beta_j$ as

$$ci_j\left(\psi_n\right) = \left[\widehat{\beta}_j - \widehat{v}_j\psi_n, \widehat{\beta}_j + \widehat{v}_j\psi_n\right], \tag{13}$$

for $j = 1, \ldots, k_z$. The following lemma gives the conditions on $\psi_n$ under which all CIs within groups $\mathcal{V}_g$ will overlap which each other when $n \to \infty$, whereas none of the CIs in different groups will overlap with each other.

*Lemma 1* *Let the groups $\mathcal{V}_g$ be as defined in Assumption 2 and the confidence intervals $ci_j\left(\psi_n\right)$, $j = 1, \ldots, k_z$, as defined in (13). Then, under Assumptions 1 and 3-6, for $n \to \infty$, $\psi_n \to \infty$, $\psi_n = o\left(n^{1/2}\right)$, and $\forall g$, all confidence intervals $ci_j\left(\psi_n\right)$ within a group, $j \in \mathcal{V}_g$, will overlap with each other, whereas none of the confidence intervals in different groups, $ci_j\left(\psi_n\right)$, $ci_{j'}\left(\psi_n\right)$, $j \in \mathcal{V}_g$, $j' \in \mathcal{V}_{g'}$, will overlap with each other.*

We can use the results of Lemma 1 to obtain a selection rule that consistently selects the valid instruments as valid, with the resulting 2SLS estimator having oracle properties. For any value $\psi_n$, classify the instruments in groups $\widehat{\mathcal{V}}_t^{over}\left(\psi_n\right)$, for $t = 1, \ldots, T\left(\psi_n\right)$, with $1 \leq T\left(\psi_n\right) \leq k_z$. For members $j \in \widehat{\mathcal{V}}_t^{over}\left(\psi_n\right)$, all $ci_j\left(\psi_n\right)$ overlap with each other. Only the largest of such groups are considered, and not their subdivisions. If, for example, all $k_z$ CIs overlap with each other, then $T\left(\psi_n\right) = 1$. It is clear from this definition that instruments can be members of multiple groups, and a group can be a singleton. For any value $\psi_n$, we then select as the group of valid instruments the largest group, denoted $\widehat{\mathcal{V}}_n$, defined as

$$\widehat{\mathcal{V}}_n := \left\{ \widehat{\mathcal{V}}_m\left(\psi_n\right) : \left|\widehat{\mathcal{V}}_m\left(\psi_n\right)\right| = \max_{t=1,\ldots,T(\psi_n)} \left|\widehat{\mathcal{V}}_t^{over}\left(\psi_n\right)\right| \right\}. \tag{14}$$

Note that for any value of $\psi_n$, there may be multiple groups with the largest number of overlapping CIs. If that is the case, at this point we simply randomly select one of these in order to have a single set of instruments for each $\psi_n$. We will discuss selection using the Sargan test in Section 3.1.

The next theorem states the conditions under which the selection $\widehat{\mathcal{V}}_n$ is consistent, which follows directly from the results of Lemma 1, as $\mathcal{V}_0$ is the largest group by Assumption 2.

*Theorem 1* *Let the $\widehat{\beta}_j$ be defined as in (9) and their confidence intervals as in (13). Let $\widehat{\mathcal{V}}_n$ be one of the largest groups of instruments for which all confidence intervals overlap with each other as defined in (14). For $\psi_n \to \infty$, $\psi_n = o\left(n^{1/2}\right)$, and under Assumptions 1-6 it follows that*

$$\lim_{n\to\infty} P\left(\widehat{\mathcal{V}}_n = \mathcal{V}_0\right) = 1.$$

The next theorem states the oracle properties of the 2SLS estimator based on selecting $\mathbf{Z}_{\widehat{\mathcal{V}}_n}$ as the valid instruments and thus $\mathbf{Z}_{\widehat{\mathcal{A}}_n} = \mathbf{Z} \setminus \left\{\mathbf{Z}_{\widehat{\mathcal{V}}_n}\right\}$ as the set of invalid instruments. This result follows directly from Theorem 2 in Guo et al. (2018).

*Theorem 2* *Let* $\mathbf{Z}_{\widehat{\mathcal{A}}_n} = \mathbf{Z} \setminus \left\{\mathbf{Z}_{\widehat{\mathcal{V}}_n}\right\}$ *and let* $\widehat{\beta}_{\widehat{\mathcal{A}}_n}$ *be the 2SLS estimator of* $\beta$*, given by*

$$\widehat{\beta}_{\widehat{\mathcal{A}}_n} = \left(\widehat{\mathbf{d}}' \mathbf{M}_{Z_{\widehat{\mathcal{A}}_n}} \widehat{\mathbf{d}}\right)^{-1} \widehat{\mathbf{d}}' \mathbf{M}_{Z_{\widehat{\mathcal{A}}_n}} \mathbf{y}.$$

Then under the conditions of Theorem 1, it follows that

$$\sqrt{n}\left(\widehat{\beta}_{\widehat{\mathcal{A}}_n} - \beta\right) \xrightarrow{d} N\left(0, \sigma^2_{or}\right).$$

For any value $\psi_n$, the sets of overlapping CIs can easily and rapidly be obtained as follows.

*Algorithm 1* *Denote the lower and upper endpoints of* $ci_j(\psi_n)$ *as defined in (13) by* $cil_j(\psi_n)$ *and* $ciu_j(\psi_n)$*. Order the confidence intervals in ascending order of the lower endpoints, and use the notation* $cil_{[j]}(\psi_n)$ *and* $ciu_{[j]}(\psi_n)$ *for the ordered intervals. For* $j = 2, \ldots, k_z$*, let* $no_{[j]}(\psi_n) = \sum_{k=1}^{j-1} 1\left(ciu_{[k]}(\psi_n) > cil_{[j]}(\psi_n)\right)$*. Then the largest set(s) of overlapping intervals are those associated with the maximum value of* $no_{[j]}(\psi_n)$*.*

For the sequences $\psi_n \to \infty, \psi_n = o\left(n^{1/2}\right)$, it follows from the results of Lemma 1 and Theorem 1 that $\widehat{\mathcal{V}}_n$ as defined in (14) converges to the unique set $\mathcal{V}_0$. It is therefore immaterial for consistent selection and oracle properties how we choose the set $\widehat{\mathcal{V}}_n$ for those values of $\psi_n$ where there are multiple groups with the largest number of overlapping CIs. We can extend the range of sequences $\psi_n$ if we choose in that case the group with the minimum value of the Sargan test as we show next.

## 3.1 | Sargan test

For the oracle model (2),

$$\mathbf{y} = \mathbf{d}\beta + \mathbf{Z}_{\mathcal{A}_0}\boldsymbol{\alpha}_{\mathcal{A}_0} + \mathbf{u} = \mathbf{X}_{\mathcal{A}_0}\boldsymbol{\theta}_{\mathcal{A}_0} + \mathbf{u},$$

with $\mathbf{X}_{\mathcal{A}_0} = \left[\mathbf{d}\ \mathbf{Z}_{\mathcal{A}_0}\right]$ and $\boldsymbol{\theta}_{\mathcal{A}_0} = \left(\beta\ \boldsymbol{\alpha}'_{\mathcal{A}_0}\right)'$, the Sargan (1958) test statistic is given by

$$S\left(\widehat{\boldsymbol{\theta}}_{\mathcal{A}_0}\right) = \frac{\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{A}_0})'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{A}_0})}{\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{A}_0})'\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{A}_0})/n}, \tag{15}$$

where $\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{A}_0}) = \mathbf{y} - \mathbf{X}_{\mathcal{A}_0}\widehat{\boldsymbol{\theta}}_{\mathcal{A}_0}$, with $\widehat{\boldsymbol{\theta}}_{\mathcal{A}_0}$ the 2SLS estimator of $\boldsymbol{\theta}_{\mathcal{A}_0}$.

As $E\left[\mathbf{Z}_i u_i\right] = 0$, and for $k_{\mathcal{A}_0} < k_z$, it follows under Assumptions 1 and 3-6 that $\sqrt{n}\left(\widehat{\theta}_{\mathcal{A}_0} - \theta_{\mathcal{A}_0}\right) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{\Sigma}_0\right)$, with $\mathbf{\Sigma}_0 = \sigma_u^2 \text{plim}\left(\mathbf{X}'_{\mathcal{A}_0}\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}_{\mathcal{A}_0}/n\right)^{-1}$, and $S\left(\widehat{\theta}_{\mathcal{A}_0}\right) \xrightarrow{d} \chi^2_{k_z - k_{\mathcal{A}_0} - 1}$. For any other selection $\mathbf{Z}_{\mathcal{A}} \neq \mathbf{Z}_{\mathcal{A}_0}$ with $k_{\mathcal{A}} \leq k_{\mathcal{A}_0}$, we have that $S\left(\widehat{\theta}_{\mathcal{A}}\right) = O_p(n)$.

The results of the CI selection method can be linked to the behaviour of the Sargan test statistic as it follows from the results of Theorems 1 and 2 that, under the conditions of Theorem 1, $S\left(\widehat{\theta}_{\widehat{\mathcal{A}}_n}\right) \xrightarrow{d} \chi^2_{k_z - k_{\mathcal{A}_0} - 1}$.

We can now allow for a wider range of values of the sequence $\psi_n$ if we select from the groups with the largest number of overlapping CIs the one with the minimum value of the Sargan test statistic. Let $M(\psi_n)$ denote the number of groups with the largest number of overlapping CIs, the collection of these groups denoted by $\left\{\widehat{\mathcal{V}}^{\max}_{m'}(\psi_n)\right\}, m' = 1, \ldots, M(\psi_n)$.

Then define $\widehat{\mathcal{V}}^{sar}_n$ as

$$\widehat{\mathcal{V}}^{sar}_n := \left\{ \widehat{\mathcal{V}}_m(\psi_n) : \left|\widehat{\mathcal{V}}_m(\psi_n)\right| = \max_{t=1,\ldots,T(\psi_n)} \left|\widehat{\mathcal{V}}^{over}_t(\psi_n)\right|, \; S\left(\widehat{\theta}_{\widehat{\mathcal{A}}_m(\psi_n)}\right) = \min_{m'=1,\ldots,M(\psi_n)} S\left(\widehat{\theta}_{\widehat{\mathcal{A}}^{\max}_{m'}(\psi_n)}\right) \right\}, \tag{16}$$

where $\widehat{\mathcal{A}}_m(\psi_n) = \mathcal{V} \setminus \widehat{\mathcal{V}}_m(\psi_n)$ and $\widehat{\mathcal{A}}^{\max}_{m'}(\psi_n) = \mathcal{V} \setminus \widehat{\mathcal{V}}^{\max}_{m'}(\psi_n), m' = 1, \ldots, M(\psi_n)$.

The next theorem gives the conditions for consistent selection and oracle properties when selecting $\widehat{\mathcal{V}}^{Sar}_n$ as the set of valid instruments.

*Theorem 3* *Let the $\widehat{\beta}_j$ be defined as in (9) and their confidence intervals as in (13). Let $\widehat{\mathcal{V}}^{sar}_n$ be as defined in (16) and $\widehat{\mathcal{A}}^{sar}_n = \mathcal{V} \setminus \widehat{\mathcal{V}}^{sar}_n$. For $k_{\mathcal{V}_0} < k_z$, let $c_n = O(1) > 0$ be such that when $n \to \infty$, $\psi_n \to \infty$, for $\frac{\psi_n}{\sqrt{n}} \leq c_n$, $\max_{t=1,\ldots,T(\psi_n)} \left|\widehat{\mathcal{V}}^{over}_t(\psi_n)\right| \to k_{\mathcal{V}_0}$ and for $\frac{\psi_n}{\sqrt{n}} > c_n$, $\max_{t=1,\ldots,T(\psi_n)} \left|\widehat{\mathcal{V}}^{over}_t(\psi_n)\right| \to K$, with $K \geq k_{\mathcal{V}_0} + 1$. Then for $n \to \infty$, $\psi_n \to \infty$, $k_{\mathcal{V}_0} = k_z$, or $k_{\mathcal{V}_0} < k_z$ and $\frac{\psi_n}{\sqrt{n}} \leq c_n$, and under Assumptions 1-6 it follows that*

$$\lim_{n\to\infty} P\left(\widehat{\mathcal{V}}^{sar}_n = \mathcal{V}_0\right) = 1$$

and

$$\sqrt{n}\left(\widehat{\beta}_{\widehat{\mathcal{A}}^{sar}_n} - \beta\right) \xrightarrow{d} N\left(0, \sigma^2_{or}\right).$$

## 3.2 | Downward testing procedure

From the results of Theorem 3, we can devise a downward testing procedure as in Andrews (1999), reducing the dimension of the problem by evaluating only the models selecting the sets with the largest number of overlapping CIs as valid instruments. The Andrews (1999) downward testing procedure uses the Sargan test statistic as a selection device for the consistent selection of the valid instruments. It starts with the model that selects all $k_z$ instruments as valid. If the Sargan test rejects this model, then the procedure next evaluates the $k_z$ models with $k_z - 1$ instruments selected as valid, treating each instrument in turn as invalid. If the minimum of the $k_z$ Sargan test statistics does not reject the null, then the associated model is selected as the valid model. If the minimum

rejects the null, then all $\binom{k_z}{2}$ models with $k_z - 2$ instruments selected as valid are evaluated. This gets repeated until a model with $k_z - k_{\mathcal{A}} - 1$ degrees of freedom has a Sargan test result that does not reject the null hypothesis. Denote the minimum of the $\binom{k_z}{k_{\mathcal{A}}}$ Sargan statistics of all possible models with $k_{\mathcal{A}}$ instruments selected as invalid by $S_{\min}\left(k_{\mathcal{A}}\right)$. Let

$$\widehat{\mathcal{A}}_{ns} := \left\{ \mathcal{A},\ k_{\mathcal{A}} = \min\left(0,1,\ldots,k_z-2\right) : S\left(\widehat{\theta}_{\mathcal{A}}\right) = S_{\min}\left(k_{\mathcal{A}}\right) < \zeta_{n,k_z-k_{\mathcal{A}}-1} \right\}.$$

Then if the critical values $\zeta_{n,k_z-k_{\mathcal{A}}-1}$ of the $\chi^2_{k_z-k_{\mathcal{A}}-1}$ distribution satisfy

$$\zeta_{n,k_z-k_{\mathcal{A}}-1} \to \infty \text{ for } n \to \infty, \text{and } \zeta_{n,k_z-k_{\mathcal{A}}-1} = o\left(n\right), \tag{17}$$

it follows from the results in Andrews (1999), that, under Assumptions 1–6,
$\lim_{n\to\infty} P\left(\widehat{\mathcal{A}}_{ns} = \mathcal{A}_0\right) = 1$, or equivalently, $\lim_{n\to\infty} P\left(\widehat{\mathcal{V}}_{ns} = \mathcal{V}_0\right) = 1$, with $\widehat{\mathcal{V}}_{ns} = \mathcal{V} \setminus \widehat{\mathcal{A}}_{ns}$.

In order to use the CI method to reduce the dimension of the downward testing procedure, consider the set of breakpoints

$$\psi^*_{j,r} = \frac{\left|\widehat{\beta}_j - \widehat{\beta}_r\right|}{\widehat{v}_j + \widehat{v}_r}, \tag{18}$$

for $j = 1,\ldots,k_z-1$, $r = j+1,\ldots,k_z$. From Algorithm 1 it follows that for $\psi_n \leq \psi^*_{j,r}$, $ci_j\left(\psi_n\right)$ and $ci_r\left(\psi_n\right)$ do not overlap, whereas they do when $\psi_n > \psi^*_{j,r}$. Let $\psi^*_{[k_z-1]} = \max_{j,r}\left(\psi^*_{j,r}\right)$. For $\psi_n > \psi^*_{[k_z-1]}$ all $k_z$ confidence intervals overlap. At $\psi_n = \psi^*_{[k_z-1]}$, the number of overlapping CIs in the largest groups drops by one to $k_z - 1$, and there will be two groups, denoted as before as $\left\{\widehat{\mathcal{V}}^{\max}_{m'}\left(\psi^*_{[k_z-1]}\right)\right\}$, $m' = 1,\ldots,M\left(\psi^*_{[k_z-1]}\right)$, with $M\left(\psi^*_{[k_z-1]}\right) = 2$. The next breakpoint where the size of the largest groups of overlapping CIs is equal to $k_z - 2$ is the minimum of the maximum of the breakpoints (18) in the two largest groups of size $k_z - 1$. Denote these maximum group-specific breakpoints by $\psi^*_{m',[k_z-2]} = \max_{\{j,r\}\in\widehat{\mathcal{V}}^{\max}_{m'}\left(\psi^*_{[k_z-1]}\right)}\left(\psi^*_{j,r}\right)$, for $m' = 1,2$, and the minimum by $\psi^*_{[k_z-2]} = \min_{m'}\left(\psi^*_{m',[k_z-2]}\right)$. Note that for $\psi^*_{[k_z-2]} < \psi_n \leq \psi^*_{[k_z-1]}$, the maximum group size remains $k_z - 1$. Then at $\psi_n = \psi^*_{[k_z-2]}$, there will be $2 \leq M\left(\psi^*_{[k_z-2]}\right) \leq 3$ groups with the maximum $k_z - 2$ overlapping CIs, and the next breakpoint where the size of the largest groups of overlapping CIs is equal to $k_z - 3$ is again determined by the minimum of the maxima of the breakpoints (18) in these groups. Repeating this, we get the $k_z - 2$ breakpoints

$$\psi^*_{[2]} < \psi^*_{[3]} < \ldots < \psi^*_{[k_z-1]}, \tag{19}$$

with $\psi^*_{[s]} = \min_{m'}\left(\psi^*_{m',[s]}\right)$, $\psi^*_{m',[s]} = \max_{\{j,r\}\in\widehat{\mathcal{V}}^{\max}_{m'}\left(\psi^*_{[s+1]}\right)}\left(\psi^*_{j,r}\right)$, and at each breakpoint we have $2 \leq M\left(\psi^*_{[s]}\right) \leq k_z - s + 1 = k_{\mathcal{A}} + 1$ groups with the maximum $s$ overlapping CIs.

Combining the results of Theorem 3 with the downward testing procedure of Andrews (1999) we get the following consistent selection and oracle properties.

*Theorem 4* *Let the breakpoints* $\left\{\psi^*_{[s]}\right\}_{s=2}^{k_z-1}$ *be as defined in (19) and let* $\psi^*_{[k_z]} = \psi^*_{[k_z-1]} + \delta$, *for a constant* $\delta > 0$, *so that the model with all* $k_z$ *instruments selected as valid is included. Let*

$$\widehat{\mathcal{V}}_n^{dts} := \left\{ \widehat{\mathcal{V}}_n^{sar}\left(\psi^*_n\right); \psi^*_n = \max_{s=2,\ldots k_z}\left(\psi^*_{[s]}\right) : S\left(\widehat{\theta}_{\widehat{\mathcal{A}}_n^{sar}\left(\psi^*_{[s]}\right)}\right) < \zeta_{n,s-1} \right\},$$

where $\widehat{\mathcal{A}}_n^{sar}\left(\psi^*_{[s]}\right) = \mathcal{V} \setminus \widehat{\mathcal{V}}_n^{sar}\left(\psi^*_{[s]}\right)$, with $\widehat{\mathcal{V}}_n^{sar}\left(\psi^*_{[s]}\right)$ defined in (16), and where $\zeta_{n,s-1}$ satisfy the conditions stated in (17). Let $\widehat{\mathcal{A}}_n^{dts} = \mathcal{V} \setminus \widehat{\mathcal{V}}_n^{dts}$. Then under the conditions of Theorem 3, it follows that

$$\lim_{n\to\infty} P\left(\widehat{\mathcal{V}}_n^{dts} = \mathcal{V}_0\right) = 1$$

and

$$\sqrt{n}\left(\widehat{\beta}_{\widehat{\mathcal{A}}_n^{dts}} - \beta\right) \xrightarrow{d} N\left(0, \sigma^2_{or}\right).$$

It follows from Theorem 4 that $\psi^*_n = O_p\left(n^{1/2}\right)$, as $\psi^*_n$ is asymptotically equivalent to $\psi^*_{\left[k_{\mathcal{V}_0}\right]}$ and $\psi^*_{\left[k_{\mathcal{V}_0}\right]}/\sqrt{n}$ is asymptotically equivalent to $c_n$ as specified in Theorem 3, see for details the proof in Appendix A.1.

Following a result in Pötscher (1983), Andrews (1999) shows that (17) holds if the $p$-value of the Sargan test satisfies $p_n \to 0$ and $\log(p_n) = o(n)$. Therefore, instead of choosing values $\zeta_{n,s-1}$ for each $s$, we can choose a single sequence $p_n$ for consistent selection. Windmeijer et al. (2019) choose as threshold $p$-value for the Sargan test $0.1/\log(n)$, following the suggestion of Belloni et al. (2012) and which satisfies the conditions for consistent model selection and oracle properties of the resulting 2SLS estimator.

With this strategy, there is a maximum of $k_z\left(k_z-1\right)/2$ models to be evaluated. Together with the use of Algorithm 1, which has a computational cost of $O\left(k_z \log\left(k_z\right)\right)$, at at most $k_z - 2$ breakpoints, the computational cost of this downward testing algorithm is of the order $O\left(k_z^2 \log\left(k_z\right)\right)$. We give a stepwise description of the full downward testing algorithm in Appendix A.4, together with an illustration using a single generated data set.[1]

Under the plurality Assumption 2, the CI downward testing procedure will consistently select the set of valid instruments. In any application it may well be the case that multiple sets of maximum size are found for which the Sargan test statistics do not reject the null. The method of Andrews (1999) is then to select the model with the minimum value of the Sargan test statistics for these models with the same degrees of freedom, which is replicated by $\widehat{\mathcal{V}}_n^{dts}$. In practice, however, a researcher should acknowledge the fact that there are multiple such models, which could be an indication of a violation of Assumption 2, and investigate their results, which could lead

[1]This method is available in the R-package CIIV, https://github.com/xlbristol/CIIV. Appendix A.9 further discusses how the method can be applied with multisample (e.g. GWAS) summary data under the assumption that the instruments are independent.

to additional insights on the possible pathways from instruments to exposure and from exposure to outcomes.

While the CI method achieves dimension reduction by ignoring the covariances between the estimators $\widehat{\beta}_j$ when constructing the sets with overlapping CIs, by using the downward Sargan based testing procedure the selected model is the one with the largest number of instruments with overlapping CIs for which the joint null hypothesis is not rejected, incorporating the full covariance structure.

## 4 | HARD THRESHOLDING METHOD

Consider next pairwise testing of the null hypotheses $H_0: \beta_j = \beta_k, j = 1, \ldots, k_z - 1; k = j+1, \ldots, k_z$. These are equivalent to $H_0: \frac{\Gamma_j}{\gamma_j} = \frac{\Gamma_k}{\gamma_k}$ and a reformulation is given by $H_0: \Gamma_k - \frac{\Gamma_j}{\gamma_j}\gamma_k = \pi_k^{[j]} = 0$. Guo et al. (2018) use the latter as the basis for their pairwise testing using Wald test statistics. Unlike the score test, the Wald test is not invariant to the reformulation of a non-linear restriction, see for example Davidson and MacKinnon (2004, pp. 422-424), and while the Wald tests for $H_0: \beta_j = \beta_k$ are symmetric, this is not the case for $H_0: \pi_k^{[j]} = 0$. As we discuss below in Section 4.3, the score test here is the same as the Sargan test for overidentifying restrictions when $\mathbf{Z}_{.j}$ and $\mathbf{Z}_{.k}$ are the excluded instruments.

An estimator for $\pi_k^{[j]}$ is given by

$$\widehat{\pi}_k^{[j]} = \widehat{\Gamma}_k - \frac{\widehat{\Gamma}_j}{\widehat{\gamma}_j}\widehat{\gamma}_k. \tag{20}$$

It follows from the delta method that $\sqrt{n}\left(\widehat{\pi}_k^{[j]} - \pi_k^{[j]}\right) \xrightarrow{d} N\left(0, \sigma^2_{\pi_k^{[j]}}\right)$, with $\sigma^2_{\pi_k^{[j]}} = \tau_j^2\left(\mathbf{Q}_{kk}^{-1} - 2\left(\frac{\gamma_k}{\gamma_j}\right)\mathbf{Q}_{kj}^{-1} + \left(\frac{\gamma_k}{\gamma_j}\right)^2\mathbf{Q}_{jj}^{-1}\right)$, where $\tau_j^2$ is as defined in (10). An estimator for the variance of $\widehat{\pi}_k^{[j]}$ is therefore given by

$$\widehat{Var}\left(\widehat{\pi}_k^{[j]}\right) = \widehat{\tau}_j^2\left((\mathbf{Z}'\mathbf{Z})_{kk}^{-1} - 2\left(\frac{\widehat{\gamma}_k}{\widehat{\gamma}_j}\right)(\mathbf{Z}'\mathbf{Z})_{kj}^{-1} + \left(\frac{\widehat{\gamma}_k}{\widehat{\gamma}_j}\right)^2(\mathbf{Z}'\mathbf{Z})_{jj}^{-1}\right), \tag{21}$$

where $\widehat{\tau}_j^2$ is as defined in (11), with $n\widehat{Var}\left(\widehat{\pi}_k^{[j]}\right) \xrightarrow{p} \sigma^2_{\pi_k^{[j]}}$.

Guo et al. (2018) consider the test statistics[2]

$$t_k^{[j]} = \frac{\widehat{\pi}_k^{[j]}}{\widehat{v}_{\pi_k^{[j]}}} \tag{22}$$

[2]We provide detail of the correspondence between the specification in Guo et al. (2018) and our notation in Appendix A.6.

for $k,j = 1,\ldots,k_z$, $k \neq j$, where $\widehat{v}_{\pi_k^{[j]}} = \sqrt{\widehat{Var}\left(\widehat{\pi}_k^{[j]}\right)}$. Let $\widehat{\sigma}_{\pi_k^{[j]}} = \sqrt{n}\widehat{v}_{\pi_k^{[j]}}$. It follows that under the null, $H_0: \pi_k^{[j]} = 0$, $t_k^{[j]} \xrightarrow{d} N(0,1)$. Hence, for the sequence $\psi_n \to \infty$, $\psi_n = o\left(n^{1/2}\right)$, when $\pi_k^{[j]} = 0$,

$$\lim_{n\to\infty} P\left(\left|t_k^{[j]}\right| \leq \psi_n\right) = 1, \tag{23}$$

and when $\pi_k^{[j]} \neq 0$,

$$\lim_{n\to\infty} P\left(\left|t_k^{[j]}\right| \leq \psi_n\right) = \lim_{n\to\infty} P\left(\left|\frac{\sqrt{n}\left(\widehat{\pi}_k^{[j]} - \pi_k^{[j]}\right)}{\widehat{\sigma}_{\pi l}^{[j]}} + \frac{\sqrt{n}\pi_k^{[j]}}{\widehat{\sigma}_{\pi k}^{[j]}}\right| \leq \psi_n\right) = 0. \tag{24}$$

Guo et al. (2018) then define the set $\widehat{\mathcal{V}}_n^{[j]}$ as

$$\widehat{\mathcal{V}}_n^{[j]} = \left\{k: \left|t_k^{[j]}\right| \leq \psi_n\right\}. \tag{25}$$

These are the instruments $k = 1,\ldots,k_z$, for which $H_0: \pi_k^{[j]} = 0$ is not rejected using critical value, or threshold, $\psi_n$. Note that instrument $j$ is always contained in $\widehat{\mathcal{V}}_n^{[j]}$. It follows that, for $\psi_n \to \infty$, $\psi_n = o\left(n^{1/2}\right)$, if $\beta_k = \beta_j$, $\lim_{n\to\infty} P\left(k \in \widehat{\mathcal{V}}_n^{[j]}\right) = 1$ and if $\beta_k \neq \beta_j$, $\lim_{n\to\infty} P\left(k \in \widehat{\mathcal{V}}_n^{[j]}\right) = 0$.

As these are not joint, but only pairwise comparisons, Guo et al. (2018) propose a majority and plurality voting scheme to consistently obtain the set of valid instruments. In their terminology, $\widehat{\mathcal{V}}_n^{[j]}$ is expert $j$'s ballot that contains expert $j$'s opinion about which instruments are valid. The number of votes an instrument $k$ gets is given by

$$VM_k = \sum_{j=1}^{k_z} 1\left(k \in \widehat{\mathcal{V}}_n^{[j]}\right).$$

The majority rule then selects an instrument as valid if it gets a vote from more than 50% of the experts. The group of instruments selected as valid is then given by

$$\widehat{\mathcal{V}}_M = \left\{k: VM_k > \frac{k_z}{2}\right\}. \tag{26}$$

If none of the instruments gets a majority vote, the plurality rule is applied, which defines the set of instruments selected as valid by

$$\widehat{\mathcal{V}}_P = \left\{k: VM_k = \max_l VM_l\right\}. \tag{27}$$

Let $\widehat{\mathcal{V}}_n^{HT} = \widehat{\mathcal{V}}_M \cup \widehat{\mathcal{V}}_P$, then Guo et al. (2018, pp. 13-14) show that under Assumptions 1-6 it follows that

$$\lim_{n\to\infty} P\left(\widehat{\mathcal{V}}_n^{HT} = \mathcal{V}_0\right) = 1$$

and

$$\sqrt{n}\left(\widehat{\beta}_n^{HT} - \beta\right) \xrightarrow{d} N\left(0, \sigma_{or}^2\right),$$

where $\widehat{\beta}_n^{HT} = \left(\widehat{\mathbf{d}}' \mathbf{M}_{Z_{\widehat{\mathcal{A}}_n^{HT}}} \widehat{\mathbf{d}}\right)^{-1} \widehat{\mathbf{d}}' \mathbf{M}_{Z_{\widehat{\mathcal{A}}_n^{HT}}} \mathbf{y}, \mathbf{Z}_{\widehat{\mathcal{A}}_n^{HT}} = \mathbf{Z} \setminus \left\{ \mathbf{Z}_{\widehat{\mathcal{V}}_n^{HT}} \right\}$.

## 4.1 | Choice of tuning parameter

Guo et al. (2018) do not treat $\psi_n$ as a classical tuning parameter and they do not specify the rate, $\psi_n \to \infty$, $\psi_n = o\left(n^{1/2}\right)$, as obtained for results (23) and (24) above. They set $\psi_n = \sqrt{2.01^2 \log\left(\max\left(k_z, n\right)\right)}$ which in the setting here with fixed $k_z$ and $n > k_z$ would lead to $\psi_n = \sqrt{2.01^2 \log\left(n\right)}$. The motivation seems to be from the fact that there are $k_z\left(k_z - 1\right)$ statistics $t_k^{[j]}$. If they were all independent $N(0,1)$ distributed random variables, then it follows that for an increasing number of instruments $k_z$,

$$\lim_{k_z \to \infty} P\left(\max_{k,j}\left(\left|t_k^{[j]}\right|\right) > \sqrt{2\log\left(k_z\left(k_z - 1\right)\right)}\right) = 0, \tag{28}$$

see Donoho and Johnstone (1994). For the $k_z$ fixed case considered here, if the $t_k^{[j]}$ were independent $N(0,1)$ distributed random variables, we have that

$$E\left[\max_{k,j}\left(t_k^{[j]}\right)\right] < \sqrt{2\log\left(k_z\left(k_z - 1\right)\right)}. \tag{29}$$

It is unclear how the result in (29) translates into an optimal choice $\psi_n$ as a function of $n$, even if the $t_k^{[j]}$ were independently distributed, which they are clearly not. We find in the Monte Carlo experiments below that the value of $\psi_n = \sqrt{2.01^2 \log\left(n\right)}$ can be much too large, resulting in selecting a large group of instruments as valid that includes invalid instruments. Guo et al. (2018, p. 800) state that in practice, the $\max\left(k_z, n\right)$ is often replaced by $k_z$ or $n$ to improve the finite sample performance. In the R-routine TSHT.R, Kang (2018), the default threshold parameter for the low dimensional setting is set equal to $\psi = \sqrt{2.01^2 \log\left(k_z\right)}$, in line with the results (28) and (29) above. In principle this choice of $\psi$ does not lead to consistent selection for fixed $k_z$ and $n \to \infty$. In their Monte Carlo simulations, Guo et al. (2018) instead set $\psi = \sqrt{2.01 \log\left(k_z\right)}$ We will use these latter two values to evaluate the performance of the hard thresholding method in the simulations and application below.

## 4.2 | Voting

The Guo et al. (2018) method achieves dimension reduction by pairwise testing of $H_0: \pi_k^{[j]} = 0$ and the voting mechanism. A weakness of the voting scheme is that it does not have a mechanism to choose between sets of instruments when there are ties, and the number of instruments selected as valid is not guaranteed to be monotonically decreasing for decreasing values of $\psi_n$. Consider the example as depicted in Table 1. There are five potential instruments. In the left

**TABLE 1** Examples of voting

| $\psi_1$ | | | | | | | $\psi_2 < \psi_1$ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $k \backslash j$ | **1** | **2** | **3** | **4** | **5** | $VM_k$ | $k \backslash j$ | **1** | **2** | **3** | **4** | **5** | $VM_k$ |
| 1 | × | × | — | — | — | 2 | 1 | × | × | — | — | — | 2 |
| 2 | × | × | × | — | — | 3 | 2 | × | × | — | — | — | 2 |
| 3 | — | × | × | × | — | 3 | 3 | — | — | × | × | — | 2 |
| 4 | — | — | × | × | — | 2 | 4 | — | — | × | × | — | 2 |
| 5 | — | — | — | — | × | 1 | 5 | — | — | — | — | × | 1 |

panel of the table, for a value $\psi_1$ for the tuning parameter, instruments 2 and 3 both get three votes, including the votes for themselves, whereas instruments 1 and 2 get two votes and instrument 5 only one vote. Hence, $\widehat{\mathcal{V}}_{n,1}^{HT} = \{2, 3\}$ and the number of instruments selected as valid is equal to 2. Next consider the right panel, with $\psi_2 < \psi_1$, and the situation is such that $\psi_2 \leq \left|t_3^{[2]}\right| \leq \psi_1$ and $\psi_2 \leq \left|t_2^{[3]}\right| \leq \psi_1$, but $\left|t_k^{[j]}\right| \leq \psi_2$ for $k, j \in \{1, 2\}$ and $k, j \in \{3, 4\}$. Now instruments 1, 2, 3 and 4 all get two votes. Application of the plurality rule (27) then leads to selecting these four instruments all as valid, $\widehat{\mathcal{V}}_{n,2}^{HT} = \{1, 2, 3, 4\}$, and so the number of valid instruments selected here increases for a decreasing value of $\psi$. Because of this, the Andrews (1999) Sargan test-based downward testing procedure cannot be applied in general to the HT method.

As is clear from Table 1, the voting mechanism can select the instruments in non-overlapping groups all as valid. One way to overcome the problem of ties in the voting matrix is to find the maximal cliques, but as this problem is np complete, Karp (1972), this negates the dimension reduction properties of the voting scheme. This problem is circumvented in the CI method, which keeps track of the groupings and selects the group of instruments with the smallest value of the Sargan test in case of ties.

Further note that for the HT method the number of instruments selected as valid can be both larger and smaller than the number of votes, as the examples in Table 1 show. With the asymmetric $t_j^{[k]}$, it could also be the case that only one instrument is selected as valid. This would happen, for example, if the left panel was changed with $\left|t_2^{[3]}\right| > \psi_1$, but $\left|t_3^{[2]}\right| \leq \psi_1$, in which case only instrument 2 is selected as valid with three votes.

## 4.3 | Relationship with the Sargan test

Proposition A1 in Appendix A.5 shows that $t_k^{[j]}$ as defined in (22) can equivalently be specified as

$$t_k^{[j]} = \frac{\widehat{\pi}_{k,2sls}^{[j]}}{\sqrt{\widehat{Var}\left(\widehat{\pi}_{k,2sls}^{[j]}\right)}},$$

after 2SLS estimation of the parameters in the just-identified model (12)

$$\mathbf{y} = \mathbf{d}\beta_j + \mathbf{Z}_{\{-j\}}\pi^{[j]} + \mathbf{u}_j,$$

with $\mathbf{Z}_{\{-j\}} = \mathbf{Z} \setminus \{\mathbf{Z}_{.j}\}$, using $\mathbf{Z}_{.j}$ as the instrument for $\mathbf{d}$, and using the notation $\widehat{\pi}^{[j]}_{2sls} = \left(\widehat{\pi}^{[j]}_{k,2sls}\right)_{k \neq j}$. Instead of the $t$, or Wald test, one could perform a score test for the null $H_0: \pi^{[j]}_k = 0$, with the only difference that the variance is estimated under the null. This score test is the same as the Sargan test of overidentifying restrictions in the model

$$\mathbf{y} = \mathbf{d}\beta_{jk} + \mathbf{Z}_{\{-jk\}}\pi^{[jk]} + \mathbf{u}_{jk}, \tag{30}$$

where $\mathbf{Z}_{\{-jk\}} = \mathbf{Z} \setminus \{\mathbf{Z}_{.j}, \mathbf{Z}_{.k}\}$, using both $\mathbf{Z}_{.j}$ and $\mathbf{Z}_{.k}$ as instruments for $\mathbf{d}$, see Newey and West (1987) and the discussion in Appendix A.5. Denoting this Sargan statistic by $S_{jk}$, then under the null $H_0: E\left[\mathbf{Z}_{i.}u_{jk,i}\right] = 0$, and under Assumptions 1 and 3-6, $S_{jk} \xrightarrow{d} \chi^2_1$.

Unlike the $t^{[j]}_k$, for which $t^{[j]}_k \neq t^{[k]}_j$, the $S_{jk}$ are symmetric, $S_{jk} = S_{kj}$, an invariance feature of the score test which is invariant to specifying the null as $H_0: \frac{\Gamma_k}{\gamma_k} - \frac{\Gamma_j}{\gamma_j} = 0$ or $H_0: \Gamma_k - \frac{\Gamma_j}{\gamma_j}\gamma_k = 0$. There are therefore $k_z\left(k_z - 1\right)/2$ statistics $S_{jk}$ and, instead of the selection rule $\widehat{\mathcal{V}}^{[j]}_n = \left\{k: \left|t^{[j]}_k\right| \leq \psi_n\right\}$, we can use the asymptotically equivalent rule $\widehat{\mathcal{V}}^{[j]}_n = \left\{k: \sqrt{S_{jk}} \leq \psi_n\right\}$.

## 5 | ROBUSTNESS TO HETEROSKEDASTICITY

Both the CI and hard thresholding procedures can be adapted to be robust to heteroskedasticity, clustering and/or serial correlation. Consider for example conditional heteroskedasticity of the general form $E\left[\mathbf{w}_i\mathbf{w}_i'|\mathbf{Z}_{i.}\right] = \mathbf{\Sigma}\left(\mathbf{Z}_{i.}\right)$ and $E\left[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'|\mathbf{Z}_{i.}\right] = \mathbf{\Lambda}\left(\mathbf{Z}_{i.}\right)$, with the functions $\mathbf{\Sigma}\left(\mathbf{Z}_{i.}\right)$ and $\mathbf{\Lambda}\left(\mathbf{Z}_{i.}\right)$ unknown. Let $\widehat{\boldsymbol{\eta}} = \left(\widehat{\Gamma}'\ \widehat{\boldsymbol{\gamma}}'\right)'$, then a robust estimator of $Var\left(\widehat{\boldsymbol{\eta}}\right)$ is given by

$$\widehat{Var}_r\left(\widehat{\boldsymbol{\eta}}\right) = \left(\mathbf{I}_2 \otimes \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\right)\left(\sum_{i=1}^{n}\left(\widehat{\boldsymbol{\varepsilon}}_i\widehat{\boldsymbol{\varepsilon}}_i' \otimes \mathbf{Z}_{i.}\mathbf{Z}_{i.}'\right)\right)\left(\mathbf{I}_2 \otimes \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\right),$$

and straightforward application of the delta method results in robust variance estimators $\widehat{Var}_r\left(\widehat{\beta}_j\right)$ and $\widehat{Var}_r\left(\widehat{\pi}^{[j]}_k\right)$.

For the CI method, instead of using the Sargan test for selection, a robust score test needs to be used, like the two-step Hansen $J$-test, (Hansen, 1982). For the oracle model (2),

$$\mathbf{y} = \mathbf{d}\beta + \mathbf{Z}_{\mathcal{A}_0}\boldsymbol{\alpha}_{\mathcal{A}_0} + \mathbf{u} = \mathbf{X}_{\mathcal{A}_0}\boldsymbol{\theta}_{\mathcal{A}_0} + \mathbf{u},$$

the two-step GMM estimator is given by

$$\widehat{\boldsymbol{\theta}}_{\mathcal{A}_0,2} = \left(\mathbf{X}'_{\mathcal{A}_0}\mathbf{Z}\mathbf{W}_n^{-1}\left(\widehat{\boldsymbol{\theta}}_{\mathcal{A}_0,1}\right)\mathbf{Z}'\mathbf{X}_{\mathcal{A}_0}\right)^{-1}\mathbf{X}'_{\mathcal{A}_0}\mathbf{Z}\mathbf{W}_n^{-1}\left(\widehat{\boldsymbol{\theta}}_{\mathcal{A}_0,1}\right)\mathbf{Z}'\mathbf{y},$$

where $\widehat{\boldsymbol{\theta}}_{\mathcal{A}_0,1}$ is an initial one-step estimator, for example the 2SLS estimator, and

$$\mathbf{W}_n\left(\widehat{\boldsymbol{\theta}}_{\mathcal{A}_0,1}\right) = \sum_{i=1}^{n}\left(Y_i - \mathbf{X}'_{\mathcal{A}_0,i.}\widehat{\boldsymbol{\theta}}_{\mathcal{A}_0,1}\right)^2\mathbf{Z}_{i.}\mathbf{Z}_{i.}'.$$

Let $\hat{\mathbf{u}}_2 = \mathbf{y} - \mathbf{X}_{\mathcal{A}_0}\hat{\boldsymbol{\theta}}_{\mathcal{A}_0,2}$ then the Hansen $J$-test statistic is given by

$$J\left(\hat{\boldsymbol{\theta}}_{\mathcal{A}_0,2}, \hat{\boldsymbol{\theta}}_{\mathcal{A}_0,1}\right) = \hat{\mathbf{u}}_2'\mathbf{Z}\mathbf{W}_n^{-1}\left(\hat{\boldsymbol{\theta}}_{\mathcal{A}_0,1}\right)\mathbf{Z}'\hat{\mathbf{u}}_2.$$

As $E\left[\mathbf{Z}_{i\cdot}u_i\right] = 0$, $J\left(\hat{\boldsymbol{\theta}}_{\mathcal{A}_0,2}, \hat{\boldsymbol{\theta}}_{\mathcal{A}_0,1}\right) \xrightarrow{d} \chi^2_{k_z - k_{\mathcal{A}_0} - 1}$, thus generalising the result for the Sargan test under conditional homoskedasticity to the case of general heteroskedasticity.

As the oracle estimator, we can then specify the 2SLS estimator with robust standard errors, or the efficient two-step GMM estimator.

## 6 | WEAK INSTRUMENTS

The relevance Assumption 1 states that $\gamma_j \neq 0$ for all $j = 1, \ldots, k_z$. In our application we use 96 single nucleotide polymorphisms (SNPs) as potential instruments for BMI to investigate its effect on blood pressure. These SNPs have been found to be associated with BMI in independent genome wide association studies (GWAS), see Locke et al. (2015). While the assumption is therefore very likely to be valid, it may well be the case that in our sample individual instruments are weak in the sense that they only explain a small amount of the variation of the exposure.

The presence of many weak instruments leads to bias in the 2SLS estimator. This many weak instrument bias is much less for the Limited Information Maximum Likelihood (LIML) and Continuously Updated GMM (CU-GMM) estimators, see Davies et al. (2015) and the references therein. Analogous to the problem of heteroskedasticity discussed in the previous section, to counter a potential many weak instruments bias problem of the 2SLS estimator, the CI and HT methods can estimate the parameters by LIML or CU-GMM, with the CI method adjusting the Sargan or Hansen test statistic accordingly.

For the selection of valid instruments, a very weak invalid instrument could often be classified as a valid instrument in the CI method due to its large standard error, and can change the selection in the HT method by giving votes to a large number of instruments. In order to overcome the selection problem with weak instruments, Guo et al. (2018) proposed a first-stage hard thresholding for $H_0: \gamma_j = 0$ and to classify instruments as uninformative and treated as invalid if

$$\left|t_{\gamma_j}\right| = \left|\frac{\hat{\gamma}_j}{\sqrt{\widehat{Var}\left(\hat{\gamma}_j\right)}}\right| < \omega_n, \tag{31}$$

with $\omega_n = \sqrt{2.01\log\left\{\max\left(k_z, n\right)\right\}}$, and where $\widehat{Var}\left(\hat{\gamma}_j\right)$ can be a robust variance estimator in case of heteroskedasticity. As with the setting of $\psi_n$ discussed in Section 4.1, the threshold parameter is set to $\omega_n = \sqrt{2.01\log\left(k_z\right)}$ in the R routine TSHT.R (Kang, 2018), also for the low-dimensional, fixed $k_z$ case, and we will apply this first-stage thresholding in our application.

A potential problem with this first-stage thresholding is that, as the instruments are not a priori considered to be valid, there is a chance that invalid instruments are more likely to cross the threshold. This may occur for instruments of the type $Z_2$ as displayed in Figure A1 in Appendix A.3. As $Z_2$ affects the unmeasured confounders that in turn affect the exposure, the $Z_2$-exposure relationship itself is confounded and could result in a stronger observed effect of the instrument

on the exposure than it otherwise would have been, and a larger chance of crossing the first-stage threshold.

## 7 | SOME MONTE CARLO RESULTS

In order to illustrate how the CI and HT methods utilise the available information, following the discussion in Sections 3 and 4, we consider a design similar to that in Guo et al. (2018; Table 2) who considered a setting with a small number of potential instruments, $k_z = 7$, in their design where the majority rule is violated, but the plurality rule holds. We consider here such setting but with a larger number of potential instruments, $k_z = 21$. We present a replication of their $k_z = 7$ design in Appendix A.7.

The data are generated from

**TABLE 2** Estimation results, $k_z = 21$

| | Mae | Coverage | CI length | $\lvert\widehat{\mathcal{A}}_n\rvert$ | $p_{or}$ | $p_{allinv}$ |
|---|---|---|---|---|---|---|
| $n = 500$ | | | | | | |
| 2SLS or | 0.017 | 0.943 | 0.093 | 12.000 | 1.000 | 1.000 |
| 2SLS | 0.423 | 0.000 | 0.088 | 0.000 | 0.000 | 0.000 |
| $HT_{4k_z}$ | 0.321 | 0.000 | 0.083 | 1.982 | 0.000 | 0.000 |
| $HT_{2k_z}$ | 0.330 | 0.000 | 0.091 | 6.901 | 0.000 | 0.000 |
| $CI_{sar}$ | 0.032 | 0.639 | 0.097 | 10.661 | 0.098 | 0.106 |
| $n = 1000$ | | | | | | |
| 2SLS or | 0.011 | 0.949 | 0.066 | 12.000 | 1.000 | 1.000 |
| 2SLS | 0.423 | 0.000 | 0.062 | 0.000 | 0.000 | 0.000 |
| $HT_{4k_z}$ | 0.325 | 0.000 | 0.065 | 6.822 | 0.000 | 0.000 |
| $HT_{2k_z}$ | 0.305 | 0.088 | 0.222 | 17.102 | 0.001 | 0.137 |
| $CI_{sar}$ | 0.014 | 0.889 | 0.066 | 11.599 | 0.538 | 0.561 |
| $n = 2000$ | | | | | | |
| 2SLS or | 0.008 | 0.949 | 0.047 | 12.000 | 1.000 | 1.000 |
| 2SLS | 0.424 | 0.000 | 0.044 | 0.000 | 0.000 | 0.000 |
| $HT_{4k_z}$ | 0.320 | 0.176 | 0.208 | 18.421 | 0.018 | 0.277 |
| $HT_{2k_z}$ | 0.012 | 0.836 | 0.088 | 13.681 | 0.585 | 0.911 |
| $CI_{sar}$ | 0.008 | 0.943 | 0.047 | 12.008 | 0.978 | 0.992 |
| $n = 5000$ | | | | | | |
| 2SLS or | 0.005 | 0.950 | 0.030 | 12.000 | 1.000 | 1.000 |
| 2SLS | 0.424 | 0.000 | 0.028 | 0.000 | 0.000 | 0.000 |
| $HT_{4k_z}$ | 0.005 | 0.947 | 0.030 | 12.031 | 0.984 | 1.000 |
| $HT_{2k_z}$ | 0.006 | 0.951 | 0.035 | 12.687 | 0.749 | 1.000 |
| $CI_{sar}$ | 0.005 | 0.946 | 0.030 | 12.012 | 0.989 | 1.000 |

*Notes*: Results from 10,000 MC replications; median absolute error; 95% CI coverage and length; number of instruments selected as invalid; frequency of selecting oracle model; frequency of selecting all invalid instruments as invalid.

$$D_i = \mathbf{Z}_{i.}'\boldsymbol{\gamma} + \varepsilon_{di}$$
$$Y_i = D_i\beta + \mathbf{Z}_{i.}'\boldsymbol{\alpha} + u_i,$$

where

$$\begin{pmatrix} u_i \\ \varepsilon_{di} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right);$$

$$\mathbf{Z}_{i.} \sim N\left(0, \boldsymbol{\Sigma}_z\right);$$

with $\beta = 1; k_z = 21; \rho = 0.25; k_{\mathcal{A}_0} = 12, \boldsymbol{\alpha} = c_a\left(\boldsymbol{\iota}_6', 0.5\boldsymbol{\iota}_6', \mathbf{0}_9'\right)'$ and $\boldsymbol{\gamma} = c_\gamma \times \boldsymbol{\iota}_{k_z}$, where $\boldsymbol{\iota}_r$ is an $r$-vector of ones, and $\mathbf{0}_r$ is an $r$-vector of zeros. There are therefore 3 groups of instruments, $\mathcal{V}_{c_\alpha/c_\gamma} = \{1, 2, \ldots, 6\}$, $\mathcal{V}_{0.5c_\alpha/c_\gamma} = \{7, 8, \ldots, 12\}$ and $\mathcal{V}_0 = \{13, 14, \ldots, 21\}$. $\mathcal{V}_0$ is the largest group and so the plurality rule holds, but not the majority rule. The elements of $\boldsymbol{\Sigma}_z$ are given by $\Sigma_{z,jk} = \rho_z^{|j-k|}$. We set $\rho_z = 0.5$ and $c_\alpha = c_\gamma = 0.4$. As in Guo et al. (2018), in this setting all instruments are strong, and the first-stage thresholding is omitted. Note that this simple design represents invalid instruments with a direct effect on the outcome of the type $Z_1$ as displayed in Figure A1 in Appendix A.3.

Before evaluating estimation results using the downward testing CI method and the HT method as described above, Figure 1 shows the frequency of selection of the oracle model for the HT and CI methods, the latter on the basis of $\widehat{\mathcal{V}}_n^{sar}(\psi)$ as defined in (16), for 10,000 Monte Carlo replications, as a function of values $\psi = (0.15, 0.20, \ldots, 6.95, 7)$ and for a sample size of $n = 2000$ . It is clear that the CI method utilises the available information better in this case and obtains a maximum frequency of selecting the oracle model of 0.98 at $\psi = 2.60$, whereas the maximum frequency for the HT method is only 0.60 at $\psi = 2.40$.

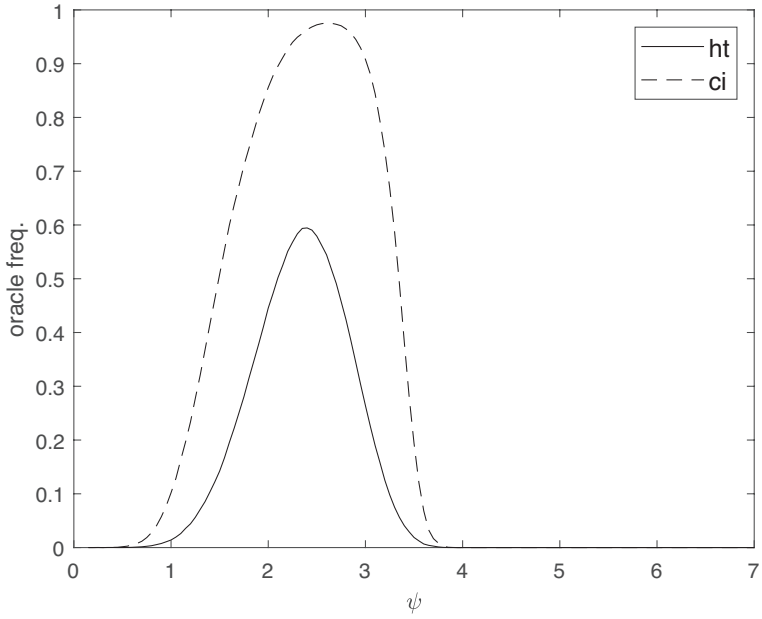


**FIGURE 1** Frequency of selecting oracle model as a function of $\psi$. $n = 2000$, $k_z = 21$, $k_{\mathcal{A}_0} = 12$, $c_\alpha = c_\gamma = 0.4$

Figure 2 shows the average total number of instruments selected as invalid, $\left|\widehat{\mathcal{A}}_n\right|$, and the average number of invalid instruments selected as invalid as a function of $\psi$. While both methods can correctly select the 12 invalid instruments as invalid for a range of values of $\psi$, the CI method can do so without also selecting valid instruments as invalid. In contrast, the HT method selects on average additional valid instruments as invalid, resulting in the difference in the frequencies of selecting the oracle model. At $\psi = 2.40$, the HT method selects on average 11.94 invalid instruments correctly as invalid, but selects on average a total of 13.52 instruments as invalid. At $\psi = 2.60$, the CI method selects on average 11.99 invalid instruments correctly as invalid, and selects on average a total of 12.01 instruments as invalid, hence the much higher frequency of selecting the oracle model for the CI method.

As is clear from Figure 2, the number of selected instruments as invalid is not monotonically increasing for decreasing values of the threshold $\psi$ for the HT method, as discussed in Section 4.2, whereas it is for the CI method.

The proposed threshold value for the HT method, $\psi_n = \sqrt{2.01^2 \log(n)} = 5.54$ is clearly too large a value in this design. The alternative choice is $\psi = \sqrt{2.01^2 \log(k_z)} = 3.51$. As shown in Figure 1, the probability of selecting the oracle model at this value is equal to only 0.018. Figure 2 shows that the average number of correctly selected invalid instruments at this value of $\psi$ is 10.93, and quite a few valid instruments are selected as invalid, with the average total number of instruments selected as invalid equal to 18.42. Guo et al. (2018) used the value of $\psi = \sqrt{2.01 \log(k_z)}$ in their Monte Carlo simulations, which in this case is equal to $\psi = 2.47$, very close to the optimal value of $\psi = 2.40$ for the maximum frequency of oracle selection. Here the probability of selecting the oracle model is equal to 0.59, on average correctly selecting 11.91 invalid instruments as invalid, and selecting on average a total number of 13.68 instruments as invalid.

Table 2 shows estimation results for the downward testing CI method and the HT method for this design for different values of the sample size $n = 500, 1000, 2000, 5000$, for 10,000 Monte

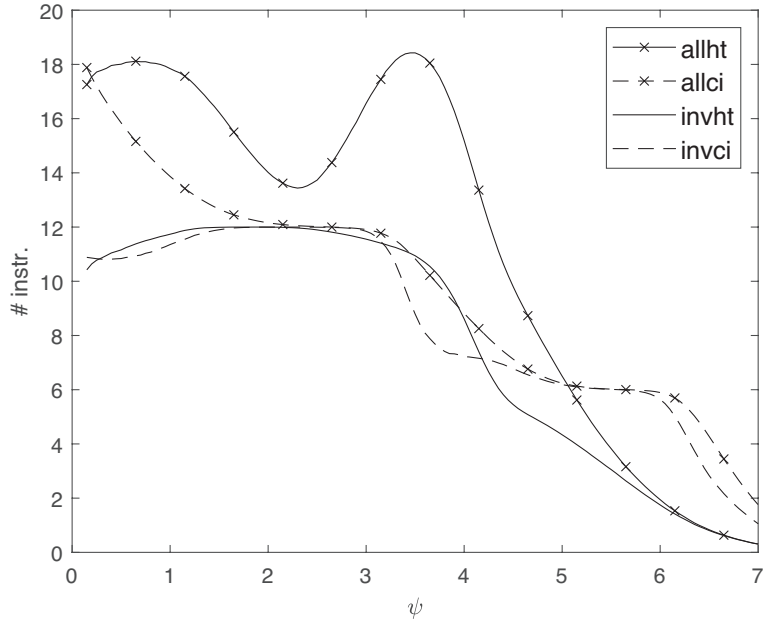


**FIGURE 2** Average total number of instruments selected as invalid (all) and number of invalid instruments selected as invalid (inv) as a function of $\psi$. $n = 2000$, $k_z = 21$, $k_{\mathcal{A}_0} = 12$, $c_\alpha = c_\gamma = 0.4$

Carlo replications. As in Guo et al. (2018), we present the median absolute error (mae), the coverage probability of the 95% CI for $\beta$ and the average length of the confidence intervals. In addition, we present the average number of instruments selected as invalid, $\left|\widehat{\mathcal{A}}_n\right|$, the frequency of selecting the oracle model, $p_{or}$, and the frequency of selecting all invalid instruments as invalid, $p_{allinv}$. The 95% CI is given by $\left[\widehat{\beta}_{\widehat{\mathcal{A}}_n} - 1.96\widehat{v}_{\widehat{\beta}_{\widehat{\mathcal{A}}_n}}, \widehat{\beta}_{\widehat{\mathcal{A}}_n} + 1.96\widehat{v}_{\widehat{\beta}_{\widehat{\mathcal{A}}_n}}\right]$, with $\widehat{v}_{\widehat{\beta}_{\widehat{\mathcal{A}}_n}} = \sqrt{\widehat{Var}\left(\widehat{\beta}_{\widehat{\mathcal{A}}_n}\right)}$, the 2SLS standard error.

Results are presented for the HT method, using $\psi = \sqrt{2.01^2 \log\left(k_z\right)} = 3.51$ and $\psi = \sqrt{2.01 \log\left(k_z\right)} = 2.47$ as threshold values, denoted $\mathrm{HT}_{4k_z}$ and $\mathrm{HT}_{2k_z}$, respectively, and for the CI method using the downward testing procedure based on the Sargan test threshold $p$-value of $0.1/\log(n)$ as described in Section 3.2 and denoted $\mathrm{CI}_{sar}$. Also given are the estimation results for the oracle 2SLS estimator (2SLS or) and the naive 2SLS estimator (2SLS) that treats all instruments as valid.

The $\mathrm{CI}_{sar}$ estimator is better behaved than the HT estimators, especially at the smaller sample sizes $n = 500$ and $n = 1000$, with the $\mathrm{CI}_{sar}$ estimator having a much smaller mae and much better coverage probability than either HT estimator. For example, at $n = 1000$, the mae for $\mathrm{CI}_{sar}$ is very similar to that of oracle 2SLS, 0.014 vs 0.011, and the coverage probability is 0.89, with the average length of the CI being the same as that of the oracle estimator and equal to 0.066. In contrast, the mae for $\mathrm{HT}_{2k_z}$ at $n = 1000$ is equal to 0.31. Its coverage probability is only 0.088, and the average length of the CI is large and equal to 0.22. The latter is due to the fact that too many instruments get selected as invalid, the average $\left|\widehat{\mathcal{A}}_n\right|$ being 17.10, compared to 11.60 for $\mathrm{CI}_{sar}$. In terms of mae and coverage probability $\mathrm{HT}_{2k_z}$ is better behaved than $\mathrm{HT}_{4k_z}$ for $n = 1000$ and $n = 2000$. Although all three estimators are close to oracle 2SLS at $n = 5000$, and select all invalid instruments correctly as invalid, the $\mathrm{HT}_{4k_z}$ is now better behaved overall than $\mathrm{HT}_{2k_z}$ as $\mathrm{HT}_{2k_z}$ still selects on average too many instruments as invalid, 12.69, versus 12.03 and 12.01 for $\mathrm{HT}_{4k_z}$ and $\mathrm{CI}_{sar}$, respectively. This is as expected, as the threshold parameter needs to increase with the sample size for consistent selection in this fixed $k_z$ setup.

The results for the $k_z = 7$ case as presented in Appendix A.7 show again a better performance of the $\mathrm{CI}_{sar}$ estimator in terms of mae and coverage probability compared to the HT estimators, although the differences are overall smaller due to the smaller number of instruments.

The CI method, as it ignores covariances for the grouping of instruments, is well suited to low instrument correlation settings as in Mendelian randomisation, but it clearly does also very well in the instrument correlation setting as specified above. The HT method may well have better finite sample properties in different settings, but a main advantage of the CI downward testing method is that it selects the model with the largest number of instruments selected as valid that passes the Sargan test. In contrast, the HT method may select models that do get rejected by the Sargan test, as we find in the application presented next.

## 8 | APPLICATION: THE EFFECT OF BMI ON BLOOD PRESSURE

We use data on 105,276 individuals from the UK Biobank and investigate the effect of BMI on diastolic blood pressure, DBP. See for further details Windmeijer et al. (2019). We use 96 SNPs as potential instruments for BMI as identified in independent GWAS studies, see Locke et al. (2015). Because of skewness, we log-transformed both BMI and DBP. The linear model specification includes age, age$^2$ and sex, together with 15 principal components of the genetic relatedness

matrix as additional explanatory variables. Because of the log-transformation, the interpretation of the causal parameter of interest $\beta$ is that of an elasticity, that is an increase of BMI by 1% changes DPB by $\beta$%.

Table 3 presents the estimation results. R code for the estimation procedure is available at https://github.com/xlbristol/CIIV. We present here the results based on the assumption of conditional homoskedasticity. Robust methods as discussed in Section 5 produce virtually identical results. The first set of results is based on the full set of instruments, not performing a first-stage thresholding, or in other words setting $\omega_n = 0$ in (31). The OLS estimate of the causal parameter is equal to 0.206 (SE 0.002), whereas the 2SLS estimate treating all 96 instruments as valid is much smaller at 0.087 (SE 0.016). The Sargan test, however, rejects the null that all the instruments are valid with a *p*-value of 2.05e-19.

The $\text{HT}_{4k_z}$ method does not select any instruments as invalid, whereas $\text{HT}_{2k_z}$ selects three instruments as invalid. The $\text{HT}_{2k_z}$ estimate is equal to 0.104 (SE 0.016), slightly larger that the 2SLS estimate, but the Sargan test still has a very small *p*-value of 3.11e-11, rejecting this model.

Using a threshold *p*-value of $0.1/\log(n) = 0.0086$ for the downward testing $\text{CI}_{sar}$ procedure results in a selection of 13 instruments as invalid. The $\text{CI}_{sar}$ estimate is 0.140 (SE 0.019), indicating a downward bias of the 2SLS estimator when treating all instruments as valid. The *p*-value of the Sargan test in the resulting model is equal to 0.011.

Further presented are the estimation results of the post-adaptive Lasso estimator of Windmeijer et al. (2019), also using a downward Sargan *p*-value based testing procedure. This method selects 11 instruments as invalid, resulting in an estimate of 0.163 (SE 0.018) and a *p*-value of the Sargan test of 0.013. This method has oracle properties if more than 50% of the instruments are valid, an assumption that does not appear to be invalid given the estimation results of the $\text{CI}_{sar}$ method. It is more efficient in this case than the $\text{CI}_{sar}$ method as it finds a model with a larger group of valid instruments that passes the Sargan test.

**TABLE 3** Estimation results, the effect of *ln*(*BMI*) on *ln*(*DBP*)

| | **Estimate** | **SE** | $\lvert\widehat{A}_n\rvert$ | ***p*-value Sargan test** |
|---|---|---|---|---|
| $\omega_n = 0, k_z = 96$ | | | | |
| OLS | 0.206 | 0.002 | | |
| 2SLS | 0.087 | 0.016 | 0 | 2.05e-19 |
| $\text{HT}_{4k_z}$ | 0.087 | 0.016 | 0 | 2.05e-19 |
| $\text{HT}_{2k_z}$ | 0.104 | 0.016 | 3 | 3.11e-11 |
| $\text{CI}_{sar}$ | 0.140 | 0.019 | 13 | 0.011 |
| $\text{Post-ALasso}_{sar}$ | 0.163 | 0.018 | 11 | 0.013 |
| $\omega_n = 3.03, k_z = 62$ | | | | |
| OLS | 0.206 | 0.002 | | |
| 2SLS | 0.086 | 0.016 | 0 | 2.80e-19 |
| $\text{HT}_{4k_z}$ | 0.098 | 0.016 | 1 | 5.29e-14 |
| $\text{HT}_{2k_z}$ | 0.104 | 0.017 | 2 | 1.90e-11 |
| $\text{CI}_{sar}$ | 0.174 | 0.020 | 9 | 0.014 |
| $\text{Post}-\text{ALasso}_{sar}$ | 0.174 | 0.020 | 9 | 0.014 |

*Notes*: Sample size $n = 105,276$.

Of the selected invalid instruments, the CI and Lasso methods have eight in common. In particular, the Lasso method is able to select as invalid two instruments that are very weak with large values of $\left|\widehat{\beta}_j\right|$ and $\text{se}\left(\widehat{\beta}_j\right)$. The CI method is not able to classify these as invalid, as discussed in Section 6. We can therefore apply the first-stage thresholding in order to exclude these instruments from consideration.

The second set of results presented in Table 3 performs a first-stage thresholding using the Guo et al. (2018) recommended value of $\omega_n = \sqrt{2.01\log\left(k_z\right)} = 3.03$. A total of 34 instruments do not pass this threshold. They are treated as invalid and included in the model as explanatory variables. The OLS and naive 2SLS estimators are virtually unchanged. The $\text{HT}_{4k_z}$ estimator selects one additional instrument as invalid, with the *p*-value of the Sargan test of the resulting model equal to 5.29e-14, clearly rejecting the model. The $\text{HT}_{2k_z}$ procedure selects two instruments as invalid and the model is also rejected by the Sargan test. Interestingly, the $\text{CI}_{sar}$ and post-adaptive Lasso procedures result in the same model selection with the same nine instruments selected as invalid. The resulting estimate is equal to 0.174 (SE 0.020), again showing that the naive 2SLS estimator of the effect of $\log(BMI)$ on $\log(DBP)$ is downward biased. This result is quite close to the OLS result, indicating that there is much less unobserved confounding in this relationship than suggested by the naive 2SLS estimator. The 9 instruments selected as invalid for $\omega_n = 3.03$ are a subset of the 13 instruments selected for $\omega_n = 0$ for $\text{CI}_{sar}$. For the Lasso procedure, eight of the nine instruments selected as invalid for $\omega_n = 3.03$ were also selected as invalid for $\omega_n = 0$.

Figure A4 in Appendix A.8 displays the CIs for the $\omega_n = 3.03, k_z = 62$ case at the selected final breakpoint $\psi_n^* = 2.35$. Only one of the instruments selected as invalid has a positive estimate for the causal effect, whereas the other eight have negative estimates, resulting in a larger estimate of the causal effect when these instruments are treated as invalid.

In order to compare the results to those found by Zhao et al. (2019), we also performed the analysis on the untransformed BMI and DPB variables. The results for OLS in this case are 0.559 (0.0062), for 2SLS, 0.248 (0.0452), and for $\text{CI}_{sar}$, 0.568 (0.0565), with 13 instruments found to be invalid. For the pre-selected $k_z = 62$ case, the results for 2SLS are 0.244 (0.0469), and for $\text{CI}_{sar}$, 0.494 (0.0557), with nine instruments found to be invalid. In the latter case, these invalid instruments are identical to the ones found above, but this is not the case when $k_z = 96$. Again these results suggest that the original OLS results suffer much less from unobserved confounding bias than the naive 2SLS estimator suggests. These results are similar to those found in the two-sample summary data analysis of Zhao et al. (2019), who found profile score, RAPS, IVW and weighted median estimates of 0.601 (0.054), 0.402 (0.106), 0.514 (0.102) and 0.472 (0.176), respectively in their analysis with 160 SNPs as potential instruments.

## 9 | CONCLUSION AND DISCUSSION

We have shown that the CI method for selecting the set of valid instruments from a putative set of instruments that may include invalid ones for an instrumental variables analysis is a viable alternative to the hard thresholding method and the adaptive Lasso method when the plurality rule holds. The methods developed for selecting invalid instruments thus far have only considered a single endogenous treatment variable. Recent analyses have considered models with multiple treatments, see for example Sanderson et al. (2019) for an examination of multivariable Mendelian randomisation. An extension of the instrument selection methods for multiple treatment models is not straightforward. When the majority rule applies, the adaptive Lasso method

can be utilised by constructing an initial consistent median-of-medians estimator, see Liang and Windmeijer (2020). For the HT and CI methods, such an extension is the subject of future research.

## ACKNOWLEDGEMENTS

Jack Bowden acknowledges support from the Medical Research Council, MC_UU_00011/2, and Xiaoran Liang from the Economic and Social Research Council, ES/P000630/1. The authors thank two referees, an associate editor and the editors, Aurore Delaigle and Simon Wood for their useful comments, which helped to improve the paper.

## ORCID

*Frank Windmeijer* http://orcid.org/0000-0002-4232-2783

## REFERENCES

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

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

Bowden, J., Smith, G. D. & Burgess, S. (2015) Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *International Journal of Epidemiology*, 44, 512–525.

Bowden, J., Smith, G. D., Haycock, P. C. & Burgess, S. (2016) Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*, 40, 304–314.

Burgess, S., Bowden, J., Dudbridge, F. & Thompson, S. G. (2018) Assessing the effectiveness of robust instrumental variable methods using multiple candidate instruments with application to mendelian randomization. arxiv :1606.03729.

Burgess, S., Small, D. S. & Thompson, S. G. (2017) A review of instrumental variable estimators for mendelian randomization. *Statistical Methods in Medical Research*, 26, 2333–2355.

Clarke, P. S. & Windmeijer, F. (2012) Instrumental variable estimators for binary outcomes. *Journal of the American Statistical Association*, 107, 1638–1652.

Davidson, R. & MacKinnon, J. G. (2004) *Econometric theory and methods*. New York: Oxford University Press.

Davies, N. M., von Hinke Kessler, S., Scholder, H., Farbmacher, S., Burgess, F.Windmeijer & Smith, G. D. (2015) The many weak instruments problem and Mendelian randomization. *Statistics in Medicine*, 34, 454–468.

Donoho, D. L. & Johnstone, I. M. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81, 425–455.

Guo, Z., Kang, H., Cai, T. T. & Small, D. S. (2018) Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society: Series B*, 80, 793–815.

Hansen, L. P. (1982) Large sample properties of generalized method of moments estimators. *Econometrica*, 50, 1029–1054.

Hartwig, F. P., Smith, G. D. & Bowden, J. (2017) Robust inference in summary data mendelian randomization via the zero modal pleiotropy assumption. *International Journal of Epidemiology*, 46, 1985–1998.

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

Imbens, G. W. (2014) Instrumental variables: an econometrician's perspective. *Statistical Science*, 29, 323–358.

Kang, H. (2018) TSHT.R. https://github.com/hyunseungkang/invalidiv.

Kang, H., Zhang, A., Cai, T. T. & Small, D. S. (2016) Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, 111, 132–144.

Karp, R. M. (1972) Reducibility among combinatorial problems. In: Miller, R. E., Thatcher, J. W. & Bohlinger, J. D. (Eds.) *Complexity of Computer Computations*. US: Springer, pp. 85–103.

Kolesár, M., Chetty, R., Friedman, J., Glaeser, E. & Imbens, G. W. (2015) Identification and inference with many invalid instruments. *Journal of Business & Economic Statistics*, 33, 474–484.

Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N. & Smith, G. D. (2008) Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27, 1133–1163.

Liang, X. & Windmeijer, F. (2020) *The adaptive lasso method for selecting valid instrumental variables in linear models with two endogenous variables*. Mimeo: University of Bristol.

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H. & Day, F. R. (2015) Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518, 197–206.

Newey, W. K. & West, K. D. (1987) Hypothesis testing with efficient method of moments estimation. *International Economic Review*, 28, 777–787.

Pötscher, B. M. (1983) Order estimation in ARMA-models by lagrangian multiplier tests. *The Annals of Statistics*, 11, 872–885.

Sanderson, E., Smith, G. D., Windmeijer, F. & Bowden, J. (2019) An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. *International Journal of Epidemiology*, 48, 713–727.

Sargan, J. D. (1958) The estimation of economic relationships using instrumental variables. *Econometrica*, 26, 393–415.

Staiger, D. & Stock, J. H. (1997) Instrumental variables regression with weak instruments. *Econometrica*, 65, 557–586.

von Hinke, S., Smith, G. D., Lawlor, D. A., Propper, C. & Windmeijer, F. (2016) Genetic markers as instrumental variables. *Journal of Health Economics*, 45, 131–148.

Windmeijer, F., Farbmacher, H., Davies, N. & Smith, G. D. (2019) On the use of the lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association*, 114, 1339–1350.

Zhao, Q., Wang, J., Hemani, G., Bowden, J. & Small, D. S. (2019) Statistical inference in two-sample summary-data mendelian randomization using robust adjusted profile score. arxiv:1801-09652.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Windmeijer F, Liang X, Hartwig FP, Bowden J. The confidence interval method for selecting valid instrumental variables. *J R Stat Soc Series B*. 2021;00:1–25. https://doi.org/10.1111/rssb.12449