The Confidence Interval Method for Selecting Valid Instrumental Variables

Frank Windmeijer
Xiaoran Liang
Fernando P. Hartwig
Jack Bowden

Discussion Paper 19 / 715

17 June 2019

Department of Economics
University of Bristol
Priory Road Complex
Bristol BS8 1TU
United Kingdom

The Confidence Interval Method for Selecting Valid Instrumental Variables*

Frank Windmeijer[a,b,†], Xiaoran Liang[a],
Fernando P. Hartwig[b,c] and Jack Bowden[b]

[a]Dept of Economics, University of Bristol, UK

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

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

June, 2019

Abstract

We propose a new method, the confidence interval (CI) method, to select valid instruments from a set of potential instruments that may contain invalid ones, for instrumental variables estimation of the causal effect of an exposure on an outcome. Invalid instruments are such that they fail the exclusion restriction and enter the model as explanatory variables. The CI method is based on the confidence intervals of the per instrument causal effects estimates. Each instrument specific causal effect estimate is obtained whilst treating all other instruments as invalid. The CI method selects the largest group with all confidence intervals overlapping with each other as the set of valid instruments. Under a plurality rule, we show that the resulting IV, or two-stage least squares (2SLS) estimator has oracle properties, meaning that it has the same limiting distribution as the oracle 2SLS estimator with the set of invalid instruments known. This result is the same as for the hard thresholding with voting (HT) method of Guo et al. (2018). 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, which determines the width of the confidence intervals. For the CI method, we can therefore use a downward testing procedure based on the Sargan test for overidentifying restrictions. We find in a simulation design similar to that of Guo et al. (2018) better properties for the CI method based estimation and inference than for the HT method and in an application of the effect of BMI on blood pressure that the CI method is better able to detect invalid instruments.

Keywords: Causal inference; Instrumental variables; Invalid instruments

*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.

†*Address for correspondence*: Frank Windmeijer, Department of Economics, University of Bristol, Priory Road, Bristol BS8 1TU, UK. Email:f.windmeijer@bristol.ac.uk

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). An IV needs to be associated with the exposure, the so-called "relevance" condition, but only associated with the outcome indirectly through its association with the exposure. The latter condition is referred to as the "exclusion" condition. This paper is concerned with violations of the exclusion condition of the instruments, following closely the setup of Kang et al. (2016), Windmeijer et al. (2018) and Guo et al. (2018).

The exclusion condition is violated, and an instrument is deemed invalid, if it has a direct effect on the outcome, or an indirect association with the outcome through unobserved confounders, over and above the effect of the exposure on the outcome. Use of an invalid instrument will lead to inconsistency of the IV, or two-stage least squares (2SLS) estimator.

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 (i.e. relevant), additionally satisfy the exclusion restriction of only being associated with the outcome through the exposure. In our Mendelian randomisation application, see also Windmeijer et al. (2018), we have 96 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). It is therefore important to have methods that can determine whether instruments are invalid under as weak a set of assumptions as possible.

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 possible subsets of instruments and to select the model that minimises an information criterion based on the Sargan test of overidentifying restrictions, Sargan (1958). However, this approach

is only feasible with a relatively small number of instruments, unlike in our application. 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. (2018) proposed an adaptive Lasso estimator in combination with a downward testing procedure based on the Sargan test of overidentifying restrictions as in Andrews (1999). When the so-called 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 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.

Relaxing the majority rule, 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 resulting 2SLS estimator when the weaker plurality rule holds. Under the plurality rule more than 50% of the instruments can be invalid.

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 confidence intervals of the instrument specific causal effect estimates overlap. The instrument specific estimates are obtained whilst treating all other instruments as invalid. 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 determines the length of the confidence intervals. As with the adaptive Lasso method, we can therefore use the Sargan test based downward testing procedure. Whilst the CI method achieves dimension reduction by ignoring the covariances between the instrument specific estimates, use of the Sargan test guarantees that the model selection is based on the joint distribution, taking into account the full covariance structure.

In Mendelian randomisation applications the genetic variants have in general a low correlation with each other and thus the CI method is very well suited for that setting.

However, all results are derived for general correlations between the (valid and invalid) instruments, and the instruments are correlated in the Monte Carlo exercise in Section 7. But allowing for a general correlation structure of the instruments does limit the types of invalid instruments that the methods can deal with. In particular, instruments that are invalid because they are affected by unmeasured confounders that also affect the outcome cannot be dealt with under general correlation structures, as explained in Section 2. Instruments that are invalid because they affect the outcome, either directly or through unobserved confounders, can be consistently selected by the HT and CI methods under general correlation structures of the instruments.

We discuss the HT method in Section 4. We show that this method is a collection of pairwise tests determining whether the individual instrument specific causal effects estimates are statistically different from each other, given a threshold value. If they are found to be not statistically different, then they give a vote to each other and the instruments with the largest number of votes are selected as valid. As we discuss in Section 4.2 and show in the Monte Carlo exercise in Section 7, the number of valid instruments selected by the HT method is not guaranteed to be monotonically decreasing in decreasing values of the tuning parameter, and therefore the downward testing procedure does not apply in general. Guo et al. (2018) only consider values of their tuning parameters, or thresholds, that are predetermined and theoretically motivated. These are specific functions of the sample size n or number of instruments k_z, which we discuss in Section 4.1 and evaluate in the Monte Carlo Section 7.

Whilst 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 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 the $k_z = 7$ design of Guo et al. (2018), but we consider a larger number of potential instruments, $k_z = 21$. We find that overall, the CI method performs better 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. The CI method produces results very similar to the adaptive Lasso method which indicates that the majority rule is not violated in this application.

2 Model and Assumptions

We follow Kang et al. (2016) and Guo et al. (2018), who considered the following potential outcomes model. For $i = 1, ..., n$, let $Y_i^{(d,\mathbf{z})}$ be the potential outcome if the individual i were to have exposure d and instrument values $\mathbf{z}$. The observed outcome for an individual i is 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, or an indirect association with the outcome.

For two possible values of the exposure d^*, d and instruments $\mathbf{z}^*$, $\mathbf{z}$, assume the following potential outcomes model

$$Y_i^{(d^*,\mathbf{z}^*)} - Y_i^{(d,\mathbf{z})} = (\mathbf{z}^* - \mathbf{z})' \boldsymbol{\phi} + (d^* - d)\,\beta \quad (1)$$

$$E\left[Y_i^{(0,0)} | \mathbf{Z}_{i.}\right] = \mathbf{Z}_{i.}' \boldsymbol{\mu}, \quad (2)$$

where $\boldsymbol{\phi}$ measures the violation of the no-direct-effect assumption of the instruments on the outcome, and $\boldsymbol{\mu}$ represents the presence of unmeasured confounders that affect both the instruments and outcome. An instrument Z_j therefore does not satisfy the standard exclusion and independence assumptions if $\phi_j \neq 0$ and/or $\mu_j \neq 0$.

The two possible violations of the exclusion restriction considered in (1) are represented by the instruments Z_1 and Z_2 in the directed acyclic graph (DAG) as shown in Figure 1.

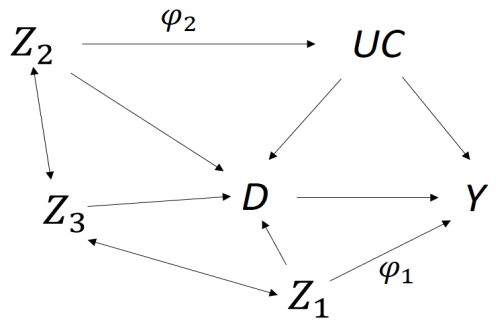

Figure 1: Causal DAG. UC represents unmeasured confounders. Z_1 and Z_2 are invalid instruments. Z_3 is a valid instrument after conditioning on Z_1 and Z_2, independent of any directional correlations between the instruments.

Z_1 has a direct effect on the outcome, whereas Z_2 has an effect on the outcome via UC, representing unmeasured confounders that affect both the outcome Y and the

exposure D. After conditioning on Z_1 and Z_2, as specified in the potential outcomes model (1), Z_3 is a valid instrument, independent of any possible directional correlation of the instruments.

Although Guo et al. (2018) assert that instruments can be correlated, the violation of the exclusion restriction (2) through unobserved confounding affecting the instrument and outcome needs some qualification when allowing for general correlations between the instruments. In Figure 2, Z_4 is an invalid instrument because of unobserved confounding UC affecting both Z_4 and the outcome Y.

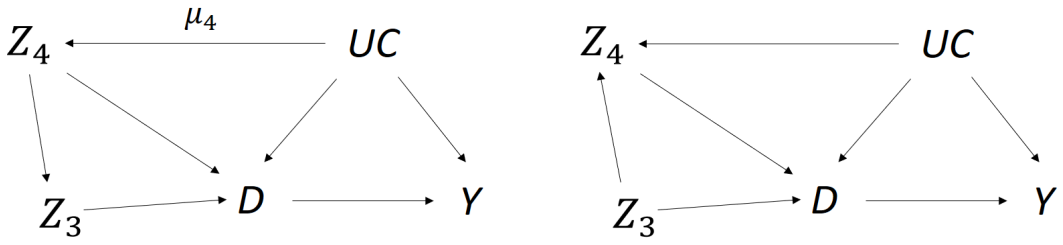

Figure 2: Instrument Z_4 is invalid. In the left panel, Z_3 is a valid instrument after conditioning on Z_4 as in (2). In the right panel, Z_3 becomes an invalid instrument after conditioning on Z_4.

In the left panel of the figure, there is directed correlation from Z_4 to Z_3, and after conditioning on Z_4 as in (2), Z_3 is a valid instrument. However, in the right panel of Figure 2 there is directed correlation from Z_3 to Z_4, resulting in Z_4 being a collider, and hence rendering Z_3 invalid after conditioning on Z_4. In other words, both μ_3 and μ_4 in (2) are then different from zero and the Lasso, CI and HT methods will not be able to select the oracle model. Therefore, in order to allow for this type of violation of the exclusion restriction one needs to make additional assumptions about the correlation structure of the instruments. A sufficient condition is that all invalid instruments of the type Z_4 are independent of the valid instruments Z_3, after conditioning on the Z_1 and Z_2 type invalid instruments from Figure 1. For ease of exposition, we assume in the following that this condition holds.

We have a random sample $\{Y_i, D_i, \mathbf{Z}_{i\cdot}'\}_{i=1}^n$. Combining (1) and (2), the observed data model for the random sample is given by

$$Y_i = D_i\beta + \mathbf{Z}_{i\cdot}'\boldsymbol{\alpha} + u_i, \tag{3}$$

where $\boldsymbol{\alpha} = \boldsymbol{\phi} + \boldsymbol{\mu}$;

$$u_i = Y_i^{(0,0)} - E\left[Y_i^{(0,0)} | \mathbf{Z}_{i.}\right]$$

and hence $E\left[u_i | \mathbf{Z}_{i.}\right] = 0$. We initially further assume conditional homoskedasticity, $E\left[u_i^2 | \mathbf{Z}_{i.}\right] = \sigma_u^2$.

The Kang et al. (2016) definition of a valid instrument is then linked to the exclusion restriction and given as follows: Instrument Z_j, $j \in \{1, ..., k_z\}$, is valid if $\alpha_j = 0$ and it is invalid if $\alpha_j \neq 0$. As in their setting, we are interested in the identification and estimation of the scalar treatment effect β in large samples with a fixed number k_z of potential instruments. We consider here only the fixed $k_z << n$ case, or the low dimensional setting in the terminology of Guo et al. (2018). This is a setting of interest in many applications and is the setting under which the oracle IV estimator has the standard limiting distribution as described below.

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 taken in deviation from their sample means. Other covariates can be partialled out in the same way. Let $\mathbf{Z}_{\mathcal{V}_0}$ and $\mathbf{Z}_{\mathcal{A}}$ be the sets of valid and invalid instruments, $\mathcal{V}_0 = \{j : \alpha_j = 0\}$, $\mathcal{A} = \{j : \alpha_j \neq 0\}$, with dimensions $k_{\mathcal{V}_0}$ and $k_{\mathcal{A}}$ respectively and $k_z = k_{\mathcal{V}_0} + k_{\mathcal{A}}$. The oracle model is given by

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

For a full column rank matrix $\mathbf{C}$ with n rows define $\mathbf{M}_C = \mathbf{I}_n - \mathbf{P}_C$, where $\mathbf{P}_C = \mathbf{C}\left(\mathbf{C}'\mathbf{C}\right)^{-1}\mathbf{C}'$ is the projection onto the column space of $\mathbf{C}$, and $\mathbf{I}_n$ is the n-dimensional identity matrix. Let $\widehat{\mathbf{d}} = \mathbf{P}_Z\mathbf{d}$, then the oracle 2SLS estimator for β is the OLS estimator in the specification

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

where ξ is defined implicitly, and is given by

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

Under standard assumptions, as defined below,

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

where

$$\sigma^2_{\beta_{or}} = \sigma^2_u \left(E\left[\mathbf{Z}_{i.} D_i\right]' E\left[\mathbf{Z}_{i.}\mathbf{Z}'_{i.}\right]^{-1} E\left[\mathbf{Z}_{i.} D_i\right] - E\left[\mathbf{Z}_{\mathcal{A},i.} D_i\right]' E\left[\mathbf{Z}_{\mathcal{A},i.}\mathbf{Z}'_{\mathcal{A},i.}\right]^{-1} E\left[\mathbf{Z}_{\mathcal{A},i.} D_i\right] \right)^{-1}. \tag{5}$$

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 projection

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

and hence $E\left[\mathbf{Z}_{i.}\varepsilon_{di}\right] = 0$. We initially assume that the k_z elements γ_j in $\boldsymbol{\gamma}$, are all different from 0:

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

This is the same assumption as in Kang et al. (2016) and Windmeijer et al. (2018). 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 $\boldsymbol{\Gamma} = E\left[\mathbf{Z}_{i.}\mathbf{Z}'_{i.}\right]^{-1} E\left[\mathbf{Z}_{i.} Y_i\right] = \boldsymbol{\gamma}\beta + \boldsymbol{\alpha}$. Then define β_j as

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

for $j = 1, ..., k_z$. Following Theorem 1 in Kang et al. (2018), a necessary and sufficient condition to identify β and the α_j, given $\boldsymbol{\Gamma}$ and $\boldsymbol{\gamma}$, is that the valid instruments form the largest group, where instruments form a group if they have the same value for β_j. This is the plurality rule. As in Guo et al. (2018), we maintain the assumption that this condition is satisfied:

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

For the sample $\{Y_i, D_i, \mathbf{Z}'_{i.}\}_{i=1}^n$, and models (3) and (6), we 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* $\mathbf{w}_i = (u_i \ \varepsilon_{di})'$. *Then* $E\left[\mathbf{w}_i\right] = 0$; $E\left[\mathbf{w}_i\mathbf{w}'_i\right] = \begin{bmatrix} \sigma^2_u & \sigma_{u\varepsilon_d} \\ \sigma_{u\varepsilon_d} & \sigma^2_{\varepsilon_d} \end{bmatrix} = \boldsymbol{\Sigma}$. *The elements of* $\boldsymbol{\Sigma}$ *are finite.*

Assumption 5 $\text{plim}\,(n^{-1}\mathbf{Z}'\mathbf{Z}) = E\,[\mathbf{Z}_{i.}\mathbf{Z}_{i.}'] = \mathbf{Q}$; $\text{plim}\,(n^{-1}\mathbf{Z}'\mathbf{d}) = E\,[\mathbf{Z}_{i.}D_i]$;
$\text{plim}\,(n^{-1}\mathbf{Z}'\mathbf{u}) = E\,[\mathbf{Z}_{i.}u_i] = 0$; $\text{plim}\,(n^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}_d) = E\,[\mathbf{Z}_{i.}\varepsilon_{di}] = 0$;
$\text{plim}\,(n^{-1}\sum_{i=1}^{n}\mathbf{w}_i) = 0$; $\text{plim}\,(n^{-1}\sum_{i=1}^{n}\mathbf{w}_i\mathbf{w}_i') = \boldsymbol{\Sigma}$.

Assumption 6 $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\text{vec}\,(\mathbf{Z}_{i.}\mathbf{w}_i') \xrightarrow{d} N\,(0, \boldsymbol{\Sigma}\otimes\mathbf{Q})$.

Note that conditional homoskedasticity $E\,[\mathbf{w}_i\mathbf{w}_i'|\mathbf{Z}_{i.}] = \boldsymbol{\Sigma}$ 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 α_j, the effect of the instrument on the outcome, and γ_j, the effect of the instrument on the exposure, also referred to as the strength of the instrument, can be arbitrary and arbitrarily correlated. For example, note from the DAG of Figure 1, that for instruments of type Z_2, the Z_2-D relationship itself is subject to unobserved confounding, which could lead to a correlation between the observed strength of the instruments and their effects on the outcome Y. The CI and HT methods are robust to any such correlation. Alternatively, the methods of Kolesár et al. (2015) and Bowden et al. (2015) do not make the plurality rule 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 et al. (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 β_j. 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. Whilst in principle all combination of instruments could be tested separately, see

Andrews (1999), in practice this may not be feasible, as is the case in our application, where we have 96 potential 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 $H_0:\beta_j=\beta_k$, combined with a voting scheme to group the instruments.

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

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

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

Under Assumptions 3-6 it follows that

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

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

Following Guo et al. (2018), let an estimator for β_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 $\mathbf{Q}^{-1}$,

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

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

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

where $\left(\mathbf{Z}'\mathbf{Z}\right)_{jj}^{-1}$ is the j-th diagonal element of $\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}$ and $\widehat{\mathbf{\Omega}}=\frac{1}{n}\sum_{i=1}^n\widehat{\boldsymbol{\varepsilon}}_i\widehat{\boldsymbol{\varepsilon}}_i'$, 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.1 that $\widehat{\beta}_j$ is identical to the 2SLS estimator of β_j in the just identified model

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

where $\mathbf{Z}_{\{-j\}} = \mathbf{Z} \backslash \{\mathbf{Z}_{.j}\}$, using $\mathbf{Z}_{.j}$ as the instrument for $\mathbf{d}$. This therefore implies that $\widehat{\beta}_j$ is the IV estimator for β based on instrument $\mathbf{Z}_{.j}$ whilst 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.

The CI method is a fast method that consistently selects the valid instruments. Given a value ψ_n, define the confidence interval $ci_j(\psi_n)$ for $\widehat{\beta}_j$ as

$$ci_j(\psi_n) = \left(\widehat{\beta}_j - \widehat{v}_j\psi_n, \widehat{\beta}_j + \widehat{v}_j\psi_n\right), \tag{12}$$

for $j = 1, ..., k_z$, where $\widehat{v}_j = \sqrt{\widehat{Var}\left(\widehat{\beta}_j\right)}$ with $\sqrt{n}\widehat{v}_j \xrightarrow{p} \sigma_j$. Let $\widehat{\sigma}_j = \sqrt{n}\widehat{v}_j$. Instruments are then classified as groups $\widehat{\mathcal{V}}_g(\psi_n)$. For members $j \in \widehat{\mathcal{V}}_g(\psi_n)$, all $ci_j(\psi_n)$ overlap with each other. It is clear from this definition that instruments can be members of multiple groups, and a group can be a singleton. We then select as the group of valid instruments the largest group, denoted $\widehat{\mathcal{V}}_n$ defined as

$$\widehat{\mathcal{V}}_n = \left\{\widehat{\mathcal{V}}_m(\psi_n) : \left|\widehat{\mathcal{V}}_m(\psi_n)\right| = \max_g \left|\widehat{\mathcal{V}}_g(\psi_n)\right|\right\}. \tag{13}$$

The next Theorem states the conditions under which this selection method is consistent.

Theorem 1 *Let the $\widehat{\beta}_j$ be defined as in (9) and their confidence intervals as in (12). Let $\widehat{\mathcal{V}}_n$ be the largest group of instruments for which all confidence intervals overlap with each other as defined in (13). For $\psi_n \to \infty$, $\psi_n = o\left(n^{1/2}\right)$, and under Assumptions 1-6 it follows that*

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

Proof. First consider a valid instrument Z_q and invalid instrument Z_s. Consider wlog the case with $\beta_s > \beta$. The joint limiting distribution of the estimators $\widehat{\beta}_q$ and $\widehat{\beta}_s$ is given by

$$\sqrt{n}\left(\begin{pmatrix}\widehat{\beta}_q\\ \widehat{\beta}_s\end{pmatrix} - \begin{pmatrix}\beta\\ \beta_s\end{pmatrix}\right) \xrightarrow{d} N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{bmatrix}\sigma_q^2 & \sigma_{qs}\\ \sigma_{qs} & \sigma_s^2\end{bmatrix}\right).$$

Then the confidence intervals will not overlap when $n \to \infty$, as

$$\lim_{n\to\infty} P\left(\widehat{\beta}_q + \widehat{v}_q\psi_n < \widehat{\beta}_s - \widehat{v}_q\psi_n\right) = \lim_{n\to\infty} P\left(\widehat{\beta}_q - \widehat{\beta}_s < -\psi_n\left(\widehat{v}_q + \widehat{v}_s\right)\right)$$

$$= \lim_{n\to\infty} P\left(\frac{\sqrt{n}\left(\left(\widehat{\beta}_q - \widehat{\beta}_s\right) - \left(\beta - \beta_s\right)\right)}{\sqrt{\sigma_q^2 + \sigma_s^2 - 2\sigma_{qs}}} < -\psi_n\frac{\widehat{\sigma}_q + \widehat{\sigma}_s}{\sqrt{\sigma_q^2 + \sigma_s^2 - 2\sigma_{qs}}} + \frac{\sqrt{n}\left(\beta_s - \beta\right)}{\sqrt{\sigma_q^2 + \sigma_s^2 - 2\sigma_{qs}}}\right)$$

$$= 1,$$

as

$$\frac{\sqrt{n}\left(\left(\widehat{\beta}_q - \widehat{\beta}_s\right) - \left(\beta - \beta_s\right)\right)}{\sqrt{\sigma_q^2 + \sigma_s^2 - 2\sigma_{qs}}} \xrightarrow{d} N\left(0,1\right)$$

and $\psi_n = o\left(n^{1/2}\right)$.

For any pair of valid instruments Z_q and Z_k, we have that the confidence intervals will overlap with probability 1 when $n \to \infty$, as

$$\begin{aligned}\lim_{n\to\infty} P\left(\widehat{\beta}_q + \widehat{v}_q\psi_n > \widehat{\beta}_k - \widehat{v}_k\psi_n\right) &= \lim_{n\to\infty} P\left(\widehat{\beta}_q - \widehat{\beta}_k > -\psi_n\left(\widehat{v}_q + \widehat{v}_k\right)\right) \\ &= \lim_{n\to\infty} P\left(\frac{\sqrt{n}\left(\left(\widehat{\beta}_q - \widehat{\beta}_k\right)\right)}{\sqrt{\sigma_q^2 + \sigma_k^2 - 2\sigma_{qk}}} > -\psi_n\frac{\widehat{\sigma}_q + \widehat{\sigma}_k}{\sqrt{\sigma_q^2 + \sigma_k^2 - 2\sigma_{qk}}}\right) \\ &= 1.\end{aligned}$$

The above results hold for all groups $\mathcal{V}_g$. For $n \to \infty$, all confidence intervals for the instruments within a group will overlap, whereas none of the confidence intervals of instruments in different groups $\mathcal{V}_g$ and $\mathcal{V}_{g'}$ will overlap. ■

Following the results in Guo et al. (2018), 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 $\mathbf{Z}_{\widehat{\mathcal{A}}_n} = \mathbf{Z}\backslash\left\{\mathbf{Z}_{\widehat{\mathcal{V}}_n}\right\}$ as the set of invalid instruments.

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

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

Proof. As $\lim_{n\to\infty} P\left(\widehat{\mathcal{V}}_n = \mathcal{V}_0\right) = 1$, the result follows directly from Theorem 2 in Guo et al. (2018). ■

3.1 Algorithm

For any ψ_n the sets of overlapping confidence intervals can easily and rapidly be obtained as follows. Denote the lower and upper endpoints of ci_j by cil_j and ciu_j. Order the confidence interval in ascending order of the lower endpoints, and use the notation $cil_{[j]}$ and $ciu_{[j]}$ for the ordered intervals. For $j = 2, ..., k_z$, let $no_{[j]} = \sum_{k=1}^{j-1} 1\left(ciu_{[k]} > cil_{[j]}\right)$. Then the largest set(s) of overlapping intervals are those associated with the maximum value of $no_{[j]}$.

3.2 Choice of Tuning Parameter, Sargan Test

Whilst any sequence ψ_n such that $\psi_n \to \infty$ and $\psi_n = o\left(n^{1/2}\right)$ will lead to consistent selection, for any given sample there is the usual trade-off in the sense that larger values of ψ_n result in larger probabilities of selecting the valid instruments as valid, but smaller probabilities of selecting the invalid instruments as invalid, and vice versa. Whilst standard cross validation techniques can be used for a data driven selection of the tuning parameter, these are well-known to not result in consistent selection and to select too many instruments as invalid in general. This was illustrated by Windmeijer et al. (2018) for the Lasso selection of invalid instruments.

Instead of choosing a value for ψ_n, one can instead focus on choosing the critical value of the Sargan test for overidentifying restrictions, following Andrews (1999), see also Windmeijer et al. (2018). For the oracle model

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

with $\mathbf{X}_{\mathcal{A}} = \left[\begin{array}{cc} \mathbf{d} & \mathbf{Z}_{\mathcal{A}} \end{array}\right]$ and $\boldsymbol{\theta}_A = \left(\begin{array}{cc} \beta & \boldsymbol{\alpha}'_{\mathcal{A}} \end{array}\right)'$, the Sargan test is given by

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

where $\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}_{\mathcal{A}}\widehat{\boldsymbol{\theta}}_{\mathcal{A}}$, with $\widehat{\boldsymbol{\theta}}_{\mathcal{A}}$ the 2SLS estimator of $\boldsymbol{\theta}_{\mathcal{A}}$. Then, under the null that the moment conditions are correct, $H_0 : E\left[\mathbf{Z}_{i\cdot}u_i\right] = 0$, $S\left(\widehat{\boldsymbol{\theta}}_{\mathcal{A}}\right) \xrightarrow{d} \chi^2_{k_z - k_{\mathcal{A}} - 1}$. For any set $\mathbf{Z}_{A^+}$, such that $\mathbf{Z}_{\mathcal{A}} \subset \mathbf{Z}_{A^+}$ it follows that $S\left(\widehat{\boldsymbol{\theta}}_{A^+}\right) \xrightarrow{d} \chi^2_{k_z - k_{A^+} - 1}$, whereas for any set $\mathbf{Z}_{A^-}$, such that $\mathbf{Z}_{\mathcal{A}} \not\subset \mathbf{Z}_{A^-}$, $S\left(\widehat{\boldsymbol{\theta}}_{A^-}\right)$ is $O_p\left(n\right)$.

The Sargan test $S\left(\widehat{\boldsymbol{\theta}}_{\mathcal{A}}\right)$ in the oracle model is equal to the score test for testing $H_0 : \boldsymbol{\alpha}_B = 0$ after 2SLS estimation of the just identified specification

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

where $\mathbf{Z}_B$ is any $k_{\mathcal{V}_0} - 1$ subset of instruments from $\mathbf{Z}_{\mathcal{V}_0}$, see Newey and West (1987), making clear the link between the Sargan test and the specification of invalid instruments as in (3). In the oracle model, it is also a test for the joint null hypothesis $H_0 : \beta_1^{\mathcal{V}_0} = \beta_2^{\mathcal{V}_0} = ... = \beta_{k_{\mathcal{V}_0}}^{\mathcal{V}_0}$, where the $\beta_r^{\mathcal{V}_0}$ are the β_j coefficients (7) for $j \in \mathcal{V}_0$, see Windmeijer (2019). Therefore, whilst the CI method achieves dimension reduction by ignoring the covariances between the estimators $\widehat{\beta}_j$, by using the Sargan test as a stopping rule as described below, the selected model is the one with the largest number of instruments with overlapping confidence intervals for which the joint null hypothesis is not rejected, incorporating the full covariance structure.

Starting from a large value ψ_1 such that all confidence intervals overlap and so $\widehat{\mathcal{V}}_1 = \mathbf{Z}$ and $\widehat{\mathcal{A}}_1 = \varnothing$, the full model selection path is obtained from the sequence of values ψ_s, which is the maximum value of ψ for which the maximum number of overlapping confidence intervals drops at each step. For any value ψ_s, there may be multiple groups with the largest number of overlapping confidence intervals. If that is the case, we follow Andrews (1999) and select the group of instruments for which the Sargan test for overidentifying restrictions is smallest.

Let $\psi_{j,r}^* = \left|\widehat{\beta}_j - \widehat{\beta}_r\right| / (\widehat{v}_j + \widehat{v}_r)$, then all possible breakpoints are given by $\psi_{j,r}^* - \varepsilon$ for some small value $\varepsilon > 0$, $j, r = 1, ..., k_z$, $j \neq r$. In practice, one only needs to consider the next breakpoint for the currently selected largest group of instruments, which is given by $\psi_s = \max_{j,r \in \widehat{\mathcal{V}}_{s-1}} \left(\psi_{j,r}^*\right) - \varepsilon$. When there are multiple groups with the maximum number of overlapping confidence intervals, then only the smallest value of the collection of next group specific breakpoints need to be considered.

For the monotonic sequence of models thus obtained we can use the downward testing procedure of Andrews (1999) and select the model with the largest degrees of freedom for which the Sargan test is smaller than the critical value $\zeta_{n,k_z-k_{\widehat{\mathcal{A}}_s}-1}$ of the $\chi^2_{k_z-k_{\widehat{\mathcal{A}}_s}-1}$ distribution. For consistent model selection, the critical values $\zeta_{n,k_z-k_{\widehat{\mathcal{A}}_s}-1}$ need to satisfy

$$\zeta_{n,k_z-k_{\widehat{\mathcal{A}}_s}-1} \to \infty \text{ for } n \to \infty, \text{ and } \zeta_{n,k_z-k_{\widehat{\mathcal{A}}_s}-1} = o\left(n\right), \tag{14}$$

see Andrews (1999). Windmeijer et al. (2018) 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 are a maximum of $k_z - 1$ models to be evaluated. Clearly, if the last model with 2 possible valid instruments gets rejected, there is no evidence that any of the instruments are valid.

4 Hard Thresholding Method

Consider next pairwise testing of the null hypotheses $H_0 : \beta_j = \beta_k$, $j = 1, ..., k_z - 1; k = j + 1, ..., k_z$. 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 nonlinear restriction, and whilst 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

$$\widehat{\pi}_k^{[j]} = \widehat{\Gamma}_k - \frac{\widehat{\Gamma}_j}{\widehat{\gamma}_j}\widehat{\gamma}_k. \tag{15}$$

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 τ_j^2 is as defined in (10). An estimator for the variance of $\widehat{\pi}_k^{[j]}$ is therefore given by

$$\widehat{Var}\left(\widehat{\pi}_k^{[j]}\right) = \widehat{\tau}_j^2\left((\mathbf{Z}'\mathbf{Z})_{kk}^{-1} - 2\left(\frac{\widehat{\gamma}_k}{\widehat{\gamma}_j}\right)(\mathbf{Z}'\mathbf{Z})_{kj}^{-1} + \left(\frac{\widehat{\gamma}_k}{\widehat{\gamma}_j}\right)^2(\mathbf{Z}'\mathbf{Z})_{jj}^{-1}\right), \tag{16}$$

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[1]

$$t_k^{[j]} = \frac{\widehat{\pi}_k^{[j]}}{\widehat{v}_{\pi_k^{[j]}}} \tag{17}$$

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

for $k, j = 1, .., k_z$, $k \neq j$, where $\widehat{v}_{\pi_k^{[j]}} = \sqrt{\widehat{Var}\left(\widehat{\pi}_k^{[j]}\right)}$ with $\sqrt{n}\widehat{v}_{\pi_k^{[j]}} \xrightarrow{p} \sigma_{\pi_k^{[j]}}$. 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 $\pi_k^{[j]} = 0$,

$$\lim_{n\to\infty} P\left(\left|t_k^{[j]}\right| \leq \psi_n\right) = 1, \tag{18}$$

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

$$\lim_{n\to\infty} P\left(\left|t_k^{[j]}\right| \leq \psi_n\right) = \lim_{n\to\infty} P\left(\left|\frac{\sqrt{n}\left(\widehat{\pi}_k^{[j]} - \pi_k^{[j]}\right)}{\widehat{\sigma}_{\pi l}^{[j]}} + \frac{\sqrt{n}\pi_k^{[j]}}{\widehat{\sigma}_{\pi k}^{[j]}}\right| \leq \psi_n\right) = 0. \tag{19}$$

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

$$\widehat{\mathcal{V}}_n^{[j]} = \left\{k : \left|t_k^{[j]}\right| \leq \psi_n\right\}. \tag{20}$$

These are the instruments $k = 1, ..., k_z$, for which $H_0 : \pi_k^{[j]} = 0$ is not rejected using critical value, or threshold, ψ_n. Note that instrument j is always contained in $\widehat{\mathcal{V}}_n^{[j]}$. It follows that 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

$$\widehat{\mathcal{V}}_M = \left\{k : VM_k > \frac{k_z}{2}\right\}. \tag{21}$$

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

$$\widehat{\mathcal{V}}_P = \left\{k : VM_k = \max_l VM_l\right\}. \tag{22}$$

Let $\widehat{\mathcal{V}}_n^{HT} = \widehat{\mathcal{V}}_M \cup \widehat{\mathcal{V}}_P$, then Guo et al. (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 $\widehat{\beta}_n^{HT} = \left(\widehat{\mathbf{d}}'\mathbf{M}_{Z_{\widehat{\mathcal{A}}_n^{HT}}}\widehat{\mathbf{d}}\right)^{-1}\widehat{\mathbf{d}}'\mathbf{M}_{Z_{\widehat{\mathcal{A}}_n^{HT}}}\mathbf{y}$, $\mathbf{Z}_{\widehat{\mathcal{A}}_n^{HT}} = \mathbf{Z}\backslash\left\{\mathbf{Z}_{\widehat{\mathcal{V}}_n^{HT}}\right\}$.

4.1 Choice of Tuning Parameter

From the results in (18) and (19) it follows that there is the usual trade-off in the value of ψ_n in classifying instruments as potentially valid or invalid. However, Guo et al. (2018) do not treat ψ_n as a classical tuning parameter, indeed stressing the fact that their method is tuning parameter free, and they do not specify the rate for ψ_n as obtained 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 $N\left(0,1\right)$ distributed random variables, then it follows that if the number of instruments k_z increases,

$$\lim_{k_z\to\infty} P\left(\max_{k,j}\left(\left|t_k^{[j]}\right|\right) > \sqrt{2\log\left(k_z\left(k_z-1\right)\right)}\right) = 0, \tag{23}$$

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

$$E\left[\max_{k,j}\left(t_k^{[j]}\right)\right] < \sqrt{2\log\left(k_z\left(k_z-1\right)\right)}. \tag{24}$$

It is unclear how the result in (24) translates into an optimal choice ψ_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 (23) and (24) above, but in

principle this choice of ψ 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 $\psi = \sqrt{2.01 \log(k_z)}$. 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 ψ_n. Consider the example as depicted in Table 1. There are 5 potential instruments. In the left panel of the table, for a value ψ_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 (22) 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 ψ. Because of this, the Andrews (1999) Sargan test based downward testing procedure can not 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

Table 1: Example of voting

ψ_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	x	x	-	-	-	2	1	x	x	-	-	-	2
2	x	x	x	-	-	3	2	x	x	-	-	-	2
3	-	x	x	x	-	3	3	-	-	x	x	-	2
4	-	-	x	x	-	2	4	-	-	x	x	-	2
5	-	-	-	-	x	1	5	-	-	-	-	x	1

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 Sargan Test

Proposition A1 in Appendix A.1 shows that $t_k^{[j]}$ as defined in (17) 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 model (A.1)

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

with $\mathbf{Z}_{\{-j\}} = \mathbf{Z}\backslash\{\mathbf{Z}_{.j}\}$, using $\mathbf{Z}_{.j}$ as the instrument for $\mathbf{d}$, and using the notation $\widehat{\boldsymbol{\pi}}_{2sls}^{[j]} = \left(\widehat{\pi}_{k,2sls}^{[j]}\right)_{k\neq j}$. Instead of the t, or Wald test, one could perform a score test for the null $H_0 : \pi_k^{[j]} = 0$, 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

$$\mathbf{y} = \mathbf{d}\beta_{jk} + \mathbf{Z}_{\{-jk\}}\boldsymbol{\pi}^{[jk]} + \mathbf{u}_{jk}, \tag{25}$$

where $\mathbf{Z}_{\{-jk\}} = \mathbf{Z}\backslash\{\mathbf{Z}_{.j}, \mathbf{Z}_{.k}\}$, 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.1. Denoting this Sargan statistic by S_{jk}, then under the null and maintained assumptions, $S_{jk} \xrightarrow{d} \chi_1^2$.

Unlike the $t_k^{[j]}$, for which $t_k^{[j]} \neq t_j^{[k]}$, 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 $\widehat{\mathcal{V}}_n^{[j]} = \left\{k : \left|t_k^{[j]}\right| \leq \psi_n\right\}$, we can use the asymptotically equivalent rule $\widehat{\mathcal{V}}_n^{[j]} = \left\{k : \sqrt{S_{jk}} \leq \psi_n\right\}$.

5 Robustness to Heteroskedasticity

Both the confidence interval 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 $\widehat{\boldsymbol{\eta}}_j = \begin{pmatrix} \widehat{\Gamma}_j & \widehat{\gamma}_j \end{pmatrix}'$, then a robust estimator of $Var\left(\widehat{\boldsymbol{\eta}}_j\right)$ is given by

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

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

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

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

the two-step GMM estimator is given by

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

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

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

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

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

Under the null $H_0 : E\left[\mathbf{Z}_{i.}u_i\right] = 0$, $J\left(\widehat{\boldsymbol{\theta}}_{\mathcal{A},2}, \widehat{\boldsymbol{\theta}}_{\mathcal{A},1}\right) \xrightarrow{d} \chi^2_{k_z - k_{\mathcal{A}} - 1}$, 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 $j = 1, ..., k_z$. 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). Whilst 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 variance 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. Analogously 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 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

$$\left|t_{\gamma_j}\right| = \left|\frac{\widehat{\gamma}_j}{\sqrt{\widehat{Var}\left(\widehat{\gamma}_j\right)}}\right| < \omega_n, \tag{26}$$

with $\omega_n = \sqrt{2.01 \log\{\max(k_z, n)\}}$, and where $\widehat{Var}\left(\widehat{\gamma}_j\right)$ can be a robust variance estimator in case of heteroskedasticity. As with the setting of ψ_n discussed in Section 4.1, the threshold parameter is set to $\omega_n = \sqrt{2.01 \log(k_z)}$ 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 priory* 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 in the DAG of Figure 1. As Z_2 affects the unmeasured confounders UC that in turn affects the exposure D, the Z_2-D 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

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.3.

The data are generated from

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

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 $\beta = 1$; $k_z = 21$; $\rho = 0.25$; $k_{\mathcal{A}} = 12$, $\boldsymbol{\alpha} = c_a\left(\boldsymbol{\iota}_6', 0.5\boldsymbol{\iota}_6', \mathbf{0}_9'\right)'$ 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. The largest group is formed by the valid instruments and so the plurality rule holds, but not the majority rule. The elements of $\boldsymbol{\Sigma}_z$ are given by $\boldsymbol{\Sigma}_{z,jk} = \rho_z^{|j-k|}$. We set $\rho_z = 0.5$ and first show selection information as a function of the tuning parameter ψ graphically for the HT and CI methods for a sample size of $n = 2000$ 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 in the DAG of Figure 1.

Figure 3 shows the frequency of selection of the oracle model for the HT and CI methods, for $10,000$ Monte Carlo replications and a grid of values $\psi = (0.15, 0.20, ..., 6.95, 7)$.

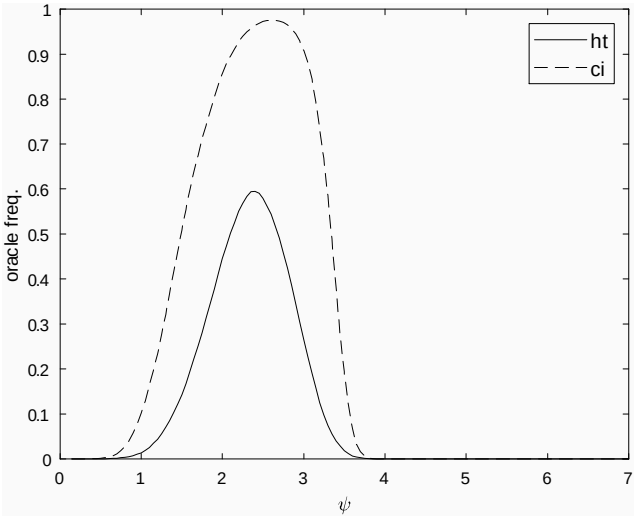

Figure 3: Frequency of selecting oracle model as a function of ψ. $n = 2000$, $k_z = 21$, $k_{\mathcal{A}} = 12$, $c_\alpha = c_\gamma = 0.4$.

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 4 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 ψ. Whilst both methods can correctly select the 12 invalid instruments as invalid for a range of values of ψ, 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 4, the number of selected instruments as invalid is not monotonic in decreasing values of the threshold ψ 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 as used in the TSHT.R routine,

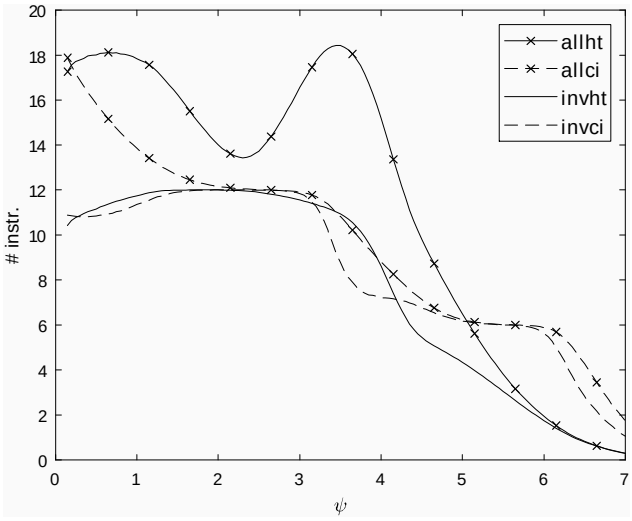

Figure 4: Average total number of instruments selected as invalid (all) and number of invalid instruments selected as invalid (inv) as a function of ψ. $n = 2000$, $k_z = 21$, $k_{\mathcal{A}} = 12$, $c_\alpha = c_\gamma = 0.4$.

Kang (2018), is $\psi = \sqrt{2.01^2 \log(k_z)} = 3.51$. As shown in Figure 3, the probability of selecting the oracle model at this value is equal to only 0.018. Figure 4 shows that the average number of correctly selected invalid instruments at this value of ψ 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 this design for different values of the sample size $n = 500, 1000, 2000, 5000$, for $10,000$ Monte Carlo replications. As in Guo et al. (2018), we present the median absolute error (mae), the coverage probability of the 95% confidence interval for β and the average length of the confidence interval. 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% confidence interval is given by $\left(\widehat{\beta}_{\widehat{\mathcal{A}}_n} - 1.96\widehat{v}_{\widehat{\beta}_{\widehat{\mathcal{A}}_n}}, \widehat{\beta}_{\widehat{\mathcal{A}}_n} + 1.96\widehat{v}_{\widehat{\beta}_{\widehat{\mathcal{A}}_n}}\right)$,

Table 2: Estimation Results, $k_z = 21$

	mae	coverage	CI length	$\widehat{\mathcal{A}}_n$	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.

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 $\psi = \sqrt{2.01^2 \log(k_z)} = 3.51$ and $\psi = \sqrt{2.01 \log(k_z)} = 2.47$ as threshold values, denoted HT$_{4k_z}$ and 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 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 CI$_{sar}$ estimator is better behaved than the HT estimators, especially at the

smaller sample sizes $n = 500$ and $n = 1000$, with the 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 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 confidence interval being the same as that of the oracle estimator and equal to 0.066. In contrast, the mae for HT_{2k_z} at $n = 1000$ is equal to 0.31. Its coverage probability is only 0.088, and the average length of the confidence interval 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 CI_{sar}. In terms of mae and coverage probability HT_{2k_z} is better behaved than 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 HT_{4k_z} is now better behaved overall than HT_{2k_z} as HT_{2k_z} still selects on average too many instruments as invalid, 12.69, versus 12.03 and 12.01 for HT_{4k_z} and 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.3 show again a better performance of the 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.

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. (2018). 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.

Table 3 presents the estimation results. 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 $\omega_n = 0$

Table 3: Estimation results, the effect of $ln(BMI)$ on $ln(DBP)$

	estimate	st err	$\lvert\widehat{A}_n\rvert$	p-value Sargan test
$\omega_n = 0$, $k_z = 96$				
OLS	0.206	0.002		
2SLS	0.087	0.016	0	2.05e-19
HT_{4k_z}	0.087	0.016	0	2.05e-19
HT_{2k_z}	0.104	0.016	3	3.11e-11
CI_{sar}	0.140	0.019	13	0.011
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
HT_{4k_z}	0.098	0.016	1	5.29e-14
HT_{2k_z}	0.104	0.017	2	1.90e-11
CI_{sar}	0.174	0.020	9	0.014
Post-ALasso_{sar}	0.174	0.020	9	0.014

Notes: sample size $n = 105,276$.

in (26). 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 HT_{4k_z} method does not select any instruments as invalid, whereas HT_{2k_z} selects 3 instruments as invalid. The 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 CI_{sar} procedure results in a selection of 13 instruments as invalid. The 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. (2018), 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 CI_{sar} method. It is more efficient in this case than the CI_{sar} method as it finds a model with a larger group of valid instruments that passes the Sargan test.

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 for consideration in the CI and HT methods.

The second set of results presented in Table 3 performs a first-stage thresholding using the Guo et al. (2018) recommended value of $\omega_n = \sqrt{2.01 \log(k_z)} = 3.03$. 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 HT_{4k_z} estimator is here the default estimator using the TSHT.R routine. It only 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 HT_{2k_z} procedure selects only 2 instruments and the model is also rejected by the Sargan test. Interestingly, the CI_{sar} and post adaptive Lasso procedures result in the same model selection with the same 9 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.

9 Conclusions

We have proposed here a new method, the confidence interval method, for selecting valid instruments from a set of potential instruments that may contain invalid ones. We showed that this method has oracle properties when the plurality rule applies, i.e. that the group of valid instruments is the largest group. This result is the same as for the hard thresholding with voting method of Guo et al. (2018), but a difference between the two methods is that for the CI method the number of instruments selected as valid is monotonically decreasing in decreasing values of the tuning parameter, whereas this

is not the case in general for the HT method. Therefore the Sargan based downward testing procedure can be applied to the CI method. It was found in simulations that this method performed better overall than the HT method in a design very similar to that of Guo et al. (2018), but with a larger number of potential instruments. In the application of the effect of BMI on blood pressure it was found that the HT method selected too few instruments as invalid, whereas the selection of the CI method was similar to the adaptive Lasso one, indicating also that the majority rule was not violated.

The fast and simple to compute CI method is therefore a viable alternative to the HT method and also to the adaptive Lasso method for when the majority rule does not hold but the plurality rule does.

References

Andrews, D.W.K., (1999), Consistent Moment Selection Procedures for Generalized Method of Moments Estimation, *Econometrica* 67, 543-564.

Belloni, A., D. Chen, V. Chernozhukov and C. Hansen, (2012), Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain, *Econometrica* 80, 2369-2429.

Bowden, J., G.D. Smith, S. Burgess, (2015), Mendelian Randomization with Invalid Instruments: Effect Estimation and Bias Detection through Egger Regression, *International Journal of Epidemiology* 44, 512-525.

Bowden, J., G. Davey Smith, P.C. Haycock and S. Burgess, (2016), Consistent Estimation in Mendelian Randomization with Some Invalid Instruments using a Weighted Median Estimator, *Genetic Epidemiology* 40, 304-314.

Burgess, S., J. Bowden, F. Dudbridge and S.G. Thompson, (2018), Robust Instrumental Variable Methods using Multiple Candidate Instruments with Application to Mendelian Randomization, arXiv:1606.03729.

Burgess, S., D.S. Small and S.G. Thompson, (2017), A Review of Instrumental Variable Estimators for Mendelian Randomization, *Statistical Methods in Medical Research 26*, 2333-2355.

Clarke, P.S. and F. Windmeijer, (2012), Instrumental Variable Estimators for Binary Outcomes, *Journal of the American Statistical Association* 107, 1638-1652.

Davies, N.M., S. von Hinke Kessler Scholder, H. Farbmacher, S. Burgess, F. Windmeijer and G. Davey Smith, (2015), The Many Weak Instruments Problem and Mendelian Randomization, *Statistics in Medicine* 34, 454-468.

Donoho, D.L. and I.M. Johnstone, (1994), Ideal Spatial Adaptation by Wavelet Shrinkage, *Biometrika* 81, 425-455.

Guo, Z., H. Kang, T. Cai and D. Small, (2018), Confidence Intervals for Causal Effects with Invalid Instruments 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., G. Davey Smith and J. Bowden, (2017), Robust Inference in Summary Data Mendelian Randomization via the Zero Modal Pleiotropy Assumption, *International Journal of Epidemiology* 46, 1985–1998.

von Hinke, S., G. Davey Smith, D.A. Lawlor, C. Propper and F. Windmeijer, (2016), Genetic Markers as Instrumental Variables, *Journal of Health Economics* 45, 131-148.

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., A. Zhang, T.T. Cai and D.S. Small, (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 R. E. Miller, J. W. Thatcher, J.D. Bohlinger (eds.). *Complexity of Computer Computations*. New York: Plenum, 85-103.

Kolesár, M., R. Chetty, J. Friedman, E. Glaeser and G.W. Imbens, (2015), Identification and Inference with Many Invalid Instruments, *Journal of Business and Economic Statistics* 33, 474-484.

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

Locke, A.E., et al. (2015), Genetic Studies of Body Mass Index Yield New Insights for Obesity Biology, *Nature* 518, 197–206.

Newey, W.K., and K.D. West, (1987), Hypothesis Testing with Efficient Method of Moments Estimation, *International Economic Review* 28, 777-787.

Sargan, J. D., (1958), The Estimation of Economic Relationships Using Instrumental Variables, *Econometrica* 26, 393–415.

Windmeijer, F., H. Farbmacher, N. Davies and G. Davey Smith, (2018), On the Use of the Lasso for Instrumental Variables Estimation with Some Invalid Instruments, *Journal of the American Statistical Association*, in press, DOI:10.1080/01621459.2018.1498346.

Windmeijer, F., (2019), Two-Stage Least Squares as Minimum Distance, *The Econometrics Journal* 22, 1-9.

Appendix

A.1 Alternative Representation of Estimators $\widehat{\beta}_j$ and $\widehat{\pi}_k^{[j]}$

Consider the model specifications

$$\mathbf{y} = \mathbf{d}\beta_j + \mathbf{Z}_{\{-j\}}\boldsymbol{\pi}^{[j]} + \mathbf{u}_j, \tag{A.1}$$

for $j = 1, ..., k_z$, where $\mathbf{Z}_{\{-j\}} = \mathbf{Z} \backslash \{\mathbf{Z}_{.j}\}$, the instrument matrix with the j-th instrument omitted. From models (3) and (6) it follows that

$$\begin{aligned}
\mathbf{u}_j &= \mathbf{u} + \frac{\alpha_j}{\gamma_j}\boldsymbol{\varepsilon}_d \\
\beta_j &= \beta + \frac{\alpha_j}{\gamma_j} \\
\pi_k^{[j]} &= \alpha_k - \frac{\alpha_j}{\gamma_j}\gamma_k \\
&= \beta\gamma_k + \alpha_k - \left(\beta + \frac{\alpha_j}{\gamma_j}\right)\gamma_k = \Gamma_k - \beta_j\gamma_k
\end{aligned}$$

where here the index $k = 1, 2, ..., j-1, j+1, ...k_z$ is the index for the included instruments. For example for $k_z = 3$, $\boldsymbol{\pi}^{[1]} = \begin{pmatrix} \pi_2^{[1]} & \pi_3^{[1]} \end{pmatrix}'$, $\boldsymbol{\pi}^{[2]} = \begin{pmatrix} \pi_1^{[2]} & \pi_3^{[2]} \end{pmatrix}'$ and $\boldsymbol{\pi}^{[3]} = \begin{pmatrix} \pi_1^{[3]} & \pi_2^{[3]} \end{pmatrix}'$.

For estimating the parameters in (A.1) by 2SLS using instruments $\mathbf{Z}$, this is a just identified model as $\mathbf{Z}_{.j}$ is the only excluded instrument. Let $\mathbf{X}_j = \begin{bmatrix} \mathbf{d} & \mathbf{Z}_{\{-j\}} \end{bmatrix}$, then the 2SLS estimator for $\boldsymbol{\theta}_j = \begin{pmatrix} \beta_j & \boldsymbol{\pi}^{[j]\prime} \end{pmatrix}'$ is given by

$$\widehat{\boldsymbol{\theta}}_{j,2sls} = \left(\mathbf{X}_j'\mathbf{P}_Z\mathbf{X}_j\right)^{-1}\mathbf{X}_j'\mathbf{P}_Z\mathbf{y} = \left(\mathbf{Z}'\mathbf{X}_j\right)^{-1}\mathbf{Z}'\mathbf{y}, \tag{A.2}$$

and so

$$\widehat{\beta}_{j,2sls} = \widehat{\boldsymbol{\theta}}_{j,2sls,1}; \tag{A.3}$$

$$\widehat{\pi}_{k,2sls}^{[j]} = \widehat{\boldsymbol{\theta}}_{j,2sls,k^*}, \tag{A.4}$$

where $k^* = k + 1\,(k < j)$. The estimator for the variance of $\widehat{\boldsymbol{\theta}}_{j,2sls}$ is given by

$$\widehat{Var}\left(\widehat{\boldsymbol{\theta}}_{j,2sls}\right) = \widehat{\sigma}_{u_j}^2\left(\mathbf{X}_j'\mathbf{P}_Z\mathbf{X}_j\right)^{-1}, \tag{A.5}$$

where $\widehat{\sigma}_{u_j}^2 = \widehat{\mathbf{u}}_{j,2sls}'\widehat{\mathbf{u}}_{j,2sls}/n$, $\widehat{\mathbf{u}}_{j,2sls} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\theta}}_{j,2sls}$, and hence

$$\widehat{Var}\left(\widehat{\beta}_{j,2sls}\right) = \widehat{\sigma}_{u_j}^2\left(\mathbf{X}_j'\mathbf{P}_Z\mathbf{X}_j\right)_{11}^{-1} \tag{A.6}$$

$$\widehat{Var}\left(\widehat{\pi}_{k,2sls}^{[j]}\right) = \widehat{\sigma}_{u_j}^2\left(\mathbf{X}_j'\mathbf{P}_Z\mathbf{X}_j\right)_{k^*,k^*}^{-1}. \tag{A.7}$$

The following proposition establishes the equivalences of $\widehat{\beta}_j$ and $\widehat{\beta}_{j,2sls}$; $\widehat{Var}\left(\widehat{\beta}_j\right)$ and $\widehat{Var}\left(\widehat{\beta}_{j,2sls}\right)$; $\widehat{\pi}_k^{[j]}$ and $\widehat{\pi}_{k,2sls}^{[j]}$; and $\widehat{Var}\left(\widehat{\pi}_k^{[j]}\right)$ and $\widehat{Var}\left(\widehat{\pi}_{k,2sls}^{[j]}\right)$.

Proposition A1 *Consider the estimators $\widehat{\beta}_j$, $\widehat{\beta}_{j,2sls}$, $\widehat{\pi}_k^{[j]}$ and $\widehat{\pi}_{k,2sls}^{[j]}$ as given in (9), (A.3), (15) and (A.4) respectively, and the variance estimators $\widehat{Var}\left(\widehat{\beta}_j\right)$, $\widehat{Var}\left(\widehat{\beta}_{j,2sls}\right)$, $\widehat{Var}\left(\widehat{\pi}_k^{[j]}\right)$ and $\widehat{Var}\left(\widehat{\pi}_{k,2sls}^{[j]}\right)$ as defined in (11), (A.6), (16) and (A.7) respectively. Then $\widehat{\beta}_j = \widehat{\beta}_{j,2sls}$; $\widehat{\pi}_k^{[j]} = \widehat{\pi}_{k,2sls}^{[j]}$; $\widehat{Var}\left(\widehat{\beta}_j\right) = \widehat{Var}\left(\widehat{\beta}_{j,2sls}\right)$; and $\widehat{Var}\left(\widehat{\pi}_k^{[j]}\right) = \widehat{Var}\left(\widehat{\pi}_{k,2sls}^{[j]}\right)$.*

Proof. Recall that we have the reduced-form and first-stage specifications

$$\begin{aligned} \mathbf{y} &= \mathbf{Z}\boldsymbol{\Gamma} + \boldsymbol{\varepsilon}_y \\ \mathbf{d} &= \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_d \end{aligned}$$

with the OLS estimators denoted $\widehat{\boldsymbol{\Gamma}}$ and $\widehat{\boldsymbol{\gamma}}$. The estimators for β_j are given $\widehat{\beta}_j = \frac{\widehat{\Gamma}_j}{\widehat{\gamma}_j}$ and the Guo et al. (2018) hard thresholding method is based on comparing the estimators $\widehat{\pi}_k^{[j]} = \widehat{\Gamma}_k - \widehat{\beta}_j\widehat{\gamma}_k = \widehat{\Gamma}_k - \frac{\widehat{\Gamma}_j}{\widehat{\gamma}_j}\widehat{\gamma}_k$to 0. Define $\widehat{\pi}^{[j]} = \left(\widehat{\pi}_1^{[j]}, ..., \widehat{\pi}_{j-1}^{[j]}, \widehat{\pi}_{j+1}^{[j]}, ..., \widehat{\pi}_{k_z}^{[j]}\right)'$. Let the OLS residuals be $\widehat{\boldsymbol{\varepsilon}}_y = \mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\Gamma}}$ and $\widehat{\boldsymbol{\varepsilon}}_d = \mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\gamma}}$, and define $\widehat{\boldsymbol{\Omega}} = \frac{1}{n}\begin{pmatrix} \widehat{\boldsymbol{\varepsilon}}_y & \widehat{\boldsymbol{\varepsilon}}_d \end{pmatrix}'\begin{pmatrix} \widehat{\boldsymbol{\varepsilon}}_y & \widehat{\boldsymbol{\varepsilon}}_d \end{pmatrix}$. Then the estimator for the variance of $\widehat{\beta}_j$, using the delta method, is given by

$$\widehat{Var}\left(\widehat{\beta}_j\right) = \frac{\widehat{\tau}_j^2}{\widehat{\gamma}_j^2}\left(\mathbf{Z}'\mathbf{Z}\right)_{jj}^{-1},$$

where

$$\begin{aligned} \widehat{\tau}_j^2 &= \begin{pmatrix} 1 & -\widehat{\beta}_j \end{pmatrix}\widehat{\boldsymbol{\Omega}}\begin{pmatrix} 1 \\ -\widehat{\beta}_j \end{pmatrix} = \frac{1}{n}\left(\widehat{\boldsymbol{\varepsilon}}_y - \widehat{\beta}_j\widehat{\boldsymbol{\varepsilon}}_d\right)'\left(\widehat{\boldsymbol{\varepsilon}}_y - \widehat{\beta}_j\widehat{\boldsymbol{\varepsilon}}_d\right) \\ &= \frac{1}{n}\left(\mathbf{y} - \widehat{\beta}_j\mathbf{d}\right)'\mathbf{M}_Z\left(\mathbf{y} - \widehat{\beta}_j\mathbf{d}\right). \end{aligned}$$

For $\widehat{\pi}_k^{[j]}$ we have the variance estimator

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

For ease of exposition and wlog, let $j = 1$, and partition $\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_{.1} & \mathbf{Z}_2 \end{bmatrix}$, where $\mathbf{Z}_2$ is an $n \times (k_z - 1)$ matrix. Equivalently, partition $\boldsymbol{\gamma} = \begin{pmatrix} \gamma_1 & \boldsymbol{\gamma}_2' \end{pmatrix}'$ and $\boldsymbol{\Gamma} = \begin{pmatrix} \Gamma_1 & \boldsymbol{\Gamma}_2' \end{pmatrix}'$. Then consider the specification

$$\mathbf{y} = \mathbf{d}\beta_1 + \mathbf{Z}_2\boldsymbol{\pi}^{[1]} + \mathbf{u}_1.$$

Let $\mathbf{Z}^* = \left[\begin{array}{cc} \widehat{\mathbf{d}} & \mathbf{Z}_2 \end{array} \right]$, then $\mathbf{Z}^* = \mathbf{Z}\widehat{\mathbf{H}}$, with

$$\widehat{\mathbf{H}} = \left[\begin{array}{cc} \widehat{\gamma}_1 & 0 \\ \widehat{\boldsymbol{\gamma}}_2 & \mathbf{I}_{k_z-1} \end{array} \right]; \quad \widehat{\mathbf{H}}^{-1} = \left[\begin{array}{cc} \widehat{\gamma}_1^{-1} & 0 \\ -\widehat{\boldsymbol{\gamma}}_2\widehat{\gamma}_1^{-1} & \mathbf{I}_{k_z-1} \end{array} \right].$$

The 2SLS estimator for $\boldsymbol{\theta}_1 = \left(\begin{array}{cc} \beta_1 & \boldsymbol{\pi}^{[1]\prime} \end{array} \right)'$ is given by

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{1,2sls} &= \left(\mathbf{Z}^{*\prime}\mathbf{Z}^*\right)^{-1}\mathbf{Z}^{*\prime}\mathbf{y} = \widehat{\mathbf{H}}^{-1}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{y} \\ &= \widehat{\mathbf{H}}^{-1}\widehat{\boldsymbol{\Gamma}}. \end{aligned}$$

Hence

$$\begin{aligned} \widehat{\beta}_{1,2sls} &= \frac{\widehat{\Gamma}_1}{\widehat{\gamma}_1} = \widehat{\beta}_1 \\ \widehat{\boldsymbol{\pi}}^{[1]}_{2sls} &= \widehat{\boldsymbol{\Gamma}}_2 - \widehat{\boldsymbol{\gamma}}_2\frac{\widehat{\Gamma}_1}{\widehat{\gamma}_1} = \widehat{\boldsymbol{\Gamma}}_2 - \widehat{\beta}_1\widehat{\boldsymbol{\gamma}}_2 \\ &= \widehat{\boldsymbol{\pi}}^{[1]}. \end{aligned}$$

Let $\widehat{\mathbf{u}}_{1,2sls} = \mathbf{y} - \mathbf{d}\widehat{\beta}_{1,2sls} - \mathbf{Z}_2\widehat{\boldsymbol{\pi}}^{[1]}_{2sls}$. As the model is just identified, it follows that $\mathbf{Z}'\widehat{\mathbf{u}}_{1,2sls} = 0$, hence $\widehat{\mathbf{u}}_{1,2sls} = \mathbf{M}_Z\widehat{\mathbf{u}}_{1,2sls} = \mathbf{M}_Z\left(\mathbf{y} - \widehat{\beta}_1\mathbf{d}\right)$. Therefore,

$$\begin{aligned} \widehat{\sigma}^2_{u_1} &= \frac{1}{n}\widehat{\mathbf{u}}'_{1,2sls}\widehat{\mathbf{u}}_{1,2sls} = \frac{1}{n}\widehat{\mathbf{u}}'_{1,2sls}M_Z\widehat{\mathbf{u}}_{1,2sls} \\ &= \left(\mathbf{y} - \widehat{\beta}_1\mathbf{d}\right)'\mathbf{M}_Z\left(\mathbf{y} - \widehat{\beta}_1\mathbf{d}\right) = \widehat{\tau}^2_1. \end{aligned}$$

The estimator of the variance of the 2SLS estimator $\widehat{\boldsymbol{\theta}}_{1,2sls}$ is given by

$$\widehat{Var}\left(\widehat{\boldsymbol{\theta}}_{1,2sls}\right) = \widehat{\sigma}^2_{u_1}\left(\mathbf{Z}^{*\prime}\mathbf{Z}^*\right)^{-1} = \widehat{\sigma}^2_{u_1}\widehat{\mathbf{H}}^{-1}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\widehat{\mathbf{H}}^{-1\prime}.$$

Let $\widehat{\mathbf{H}}^{-1}_{1.}$ be the first row of $\widehat{\mathbf{H}}^{-1}$. Then

$$\begin{aligned} \widehat{Var}\left(\widehat{\beta}_{1,2sls}\right) &= \widehat{\sigma}^2_{u_1}\widehat{\mathbf{H}}^{-1}_{1.}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\left(\widehat{\mathbf{H}}^{-1}_{1.}\right)' \\ &= \widehat{\sigma}^2_{u_1}\left(\begin{array}{cc} \widehat{\gamma}_1^{-1} & 0 \end{array} \right)\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\left(\begin{array}{c} \widehat{\gamma}_1^{-1} \\ 0 \end{array} \right) \\ &= \frac{\widehat{\tau}^2_1}{\widehat{\gamma}^2_1}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}_{11} = \widehat{Var}\left(\widehat{\beta}_1\right). \end{aligned}$$

For $k = 2, ..., k_z$, let $\mathbf{e}^{k-1}_{k_z-1}$ be a $k_z - 1$ dimensional unit vector with $(k-1)$-th element

equal to 1. Then,

$$\begin{aligned}
\widehat{Var}\left(\widehat{\pi}^{[1]}_{k,2sls}\right) &= \widehat{\sigma}^2_{u_1}\widehat{\mathbf{H}}^{-1}_{k.}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\left(\widehat{\mathbf{H}}^{-1}_{k.}\right)' \\
&= \widehat{\tau}^2_1\begin{pmatrix} -\frac{\widehat{\gamma}_k}{\widehat{\gamma}_1} & \left(\mathbf{e}^{k-1}_{k_z-1}\right)' \end{pmatrix}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\begin{pmatrix} -\frac{\widehat{\gamma}_k}{\widehat{\gamma}_1} \\ \mathbf{e}^{k-1}_{k_z-1} \end{pmatrix} \\
&= \widehat{\tau}^2_1\left(\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}_{kk} - 2\left(\frac{\widehat{\gamma}_k}{\widehat{\gamma}_1}\right)\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}_{k1} + \left(\frac{\widehat{\gamma}_k}{\widehat{\gamma}_1}\right)^2\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}_{11}\right) \\
&= \widehat{Var}\left(\widehat{\pi}^{[1]}_k\right).
\end{aligned}$$

■

It therefore follows that the t-test statistic for testing $H_0 : \pi^{[1]}_k = 0$, given by

$$t^{[1]}_k = \frac{\widehat{\pi}^{[1]}_k}{\sqrt{\widehat{Var}\left(\widehat{\pi}^{[1]}_k\right)}},$$

is identical to the 2SLS t-statistic for testing the null $H_0 : \pi^{[1]}_k$ in the just identified model

$$\mathbf{y} = \mathbf{d}\beta_1 + \mathbf{Z}_2\boldsymbol{\pi}^{[1]} + \mathbf{u}_1.$$

This generalises to any j.

Next, partition $\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_{.1} & \mathbf{Z}_{.2} & \mathbf{Z}_3 \end{bmatrix}$, $\boldsymbol{\pi}^{[1]} = \begin{pmatrix} \pi^{[1]}_2 & \boldsymbol{\pi}^{[1]\prime}_3 \end{pmatrix}'$ and consider the test for $H_0 : \pi^{[1]}_2 = 0$ in

$$\mathbf{y} = \mathbf{d}\beta_1 + \mathbf{Z}_{.2}\pi^{[1]}_2 + \mathbf{Z}_3\boldsymbol{\pi}^{[1]}_3 + \mathbf{u}_1.$$

The model under the null is then given by

$$\mathbf{y} = \mathbf{d}\beta_1 + \mathbf{Z}_3\boldsymbol{\pi}^{[1]}_3 + \mathbf{u}_1 \tag{A.8}$$

and the score test for $H_0 : \pi^{[1]}_2 = 0$ is then the same as the Sargan test for overidentifying restrictions in (A.8) after estimation by 2SLS using instruments $\mathbf{Z}$, see Newey and West (1987). The Guo et al. (2018) method is a Wald test approach, which is asymmetric, that is $t^{[1]}_2 \neq t^{[2]}_1$, whereas the Sargan test is symmetric, i.e. the score test for testing $H_0 : \pi^{[1]}_2 = 0$ is identical to the score test for testing $H_0 : \pi^{[2]}_1 = 0$ in the specification

$$\mathbf{y} = \mathbf{d}\beta_2 + \mathbf{Z}_{.1}\pi^{[2]}_1 + \mathbf{Z}_3\boldsymbol{\pi}^{[2]}_3 + \mathbf{u}_2.$$

A.2 Formulation of Threshold Set by Guo et al. (2018)

In their formulation of the model, Guo et al. (2018) explicitly include exogenous explanatory variables $\mathbf{X}$, and their matrix $\mathbf{W} = \begin{bmatrix} \mathbf{Z} & \mathbf{X} \end{bmatrix}$. In the low dimension setting we consider here, the $\mathbf{X}$ variables have been partialled out, and $\mathbf{W} = \mathbf{Z}$, where it is implicitly understood that $\mathbf{Z}$ are the residuals after linear regression on $\mathbf{X}$. Then, following their notation, $\widehat{\mathbf{U}} = (\mathbf{Z}'\mathbf{Z}/n)^{-1}$ and $\widehat{\sigma^2}^{[j]}$ is the same as $\widehat{\tau}_j^2$ as defined in (11). The formulation of the threshold set $\widehat{\mathcal{V}}^{[j]}$ is given in Guo et al. (2018, equation (7), page 9) as

$$\widehat{\mathcal{V}}^{[j]} = \left\{ k : \left|\widehat{\pi}_k^{[j]}\right| \leq \sqrt{\widehat{\sigma^2}^{[j]}} \frac{\left\| \mathbf{W} \left\{ \widehat{\mathbf{U}}_{.k} - \left(\frac{\widehat{\gamma}_k}{\widehat{\gamma}_j} \right) \widehat{\mathbf{U}}_{.j} \right\} \right\|_2}{\sqrt{n}} \sqrt{\frac{2.01^2 \log\left(\max\left(k_z, n\right)\right)}{n}} \right\}.$$

Denote $\sqrt{2.01^2 \log\left(\max\left(k_z, n\right)\right)} = \psi_n$. Then consider

$$\begin{aligned}
& \frac{\widehat{\sigma^2}^{[j]}}{n^2} \left\| \mathbf{W} \left\{ \widehat{\mathbf{U}}_{.k} - \left(\frac{\widehat{\gamma}_k}{\widehat{\gamma}_j} \right) \widehat{\mathbf{U}}_{.j} \right\} \right\|_2^2 \\
= & \frac{\widehat{\sigma^2}^{[j]}}{n^2} \left\| \mathbf{Z} \left\{ \widehat{\mathbf{U}}_{.k} - \left(\frac{\widehat{\gamma}_k}{\widehat{\gamma}_j} \right) \widehat{\mathbf{U}}_{.j} \right\} \right\|_2^2 \\
= & \widehat{\tau}_j^2 \left((\mathbf{Z}'\mathbf{Z})_{.k}^{-1} - \left(\frac{\widehat{\gamma}_k}{\widehat{\gamma}_j} \right) (\mathbf{Z}'\mathbf{Z})_{.j}^{-1} \right)' \mathbf{Z}'\mathbf{Z} \left((\mathbf{Z}'\mathbf{Z})_{.k}^{-1} - \left(\frac{\widehat{\gamma}_k}{\widehat{\gamma}_j} \right) (\mathbf{Z}'\mathbf{Z})_{.j}^{-1} \right) \\
= & \widehat{\tau}_j^2 \left((\mathbf{Z}'\mathbf{Z})_{kk}^{-1} - 2 \left(\frac{\widehat{\gamma}_k}{\widehat{\gamma}_j} \right) (\mathbf{Z}'\mathbf{Z})_{kj}^{-1} + \left(\frac{\widehat{\gamma}_k}{\widehat{\gamma}_j} \right)^2 (\mathbf{Z}'\mathbf{Z})_{jj}^{-1} \right) \\
= & \widehat{Var}\left(\widehat{\pi}_k^{[j]} \right),
\end{aligned}$$

as defined in (16).

Therefore,

$$\begin{aligned}
& \sqrt{\widehat{\sigma^2}^{[j]}} \frac{\left\| \mathbf{W} \left\{ \widehat{\mathbf{U}}_{.k} - \left(\frac{\widehat{\gamma}_k}{\widehat{\gamma}_j} \right) \widehat{\mathbf{U}}_{.j} \right\} \right\|_2}{\sqrt{n}} \sqrt{\frac{2.01^2 \log\left(\max\left(k_z, n\right)\right)}{n}} \\
= & \frac{\sqrt{\widehat{\sigma^2}^{[j]}}}{n} \left\| \mathbf{Z} \left\{ \widehat{\mathbf{U}}_{.k} - \left(\frac{\widehat{\gamma}_k}{\widehat{\gamma}_j} \right) \widehat{\mathbf{U}}_{.j} \right\} \right\|_2 \psi_n \\
= & \sqrt{\widehat{Var}\left(\widehat{\pi}_k^{[j]} \right)} \psi_n
\end{aligned}$$

and hence

$$\begin{aligned}\widehat{\mathcal{V}}^{[j]} &= \left\{k : \left|\widehat{\pi}_k^{[j]}\right| \leq \sqrt{\widehat{Var}\left(\widehat{\pi}_k^{[j]}\right)}\psi_n\right\} \\ &= \left\{k : \left|\frac{\widehat{\pi}_k^{[j]}}{\sqrt{\widehat{Var}\left(\widehat{\pi}_k^{[j]}\right)}}\right| \leq \psi_n\right\} = \left\{k : \left|t_k^{[j]}\right| \leq \psi_n\right\}\end{aligned}$$

A.3 Some Further Monte Carlo Results

Table A1 presents results for the same design as in Guo et al. (2018, Table 2), with $k_z = 7$, $k_{\mathcal{A}} = 4$, $\boldsymbol{\alpha} = c_a\left(\boldsymbol{\iota}_2', 0.5\boldsymbol{\iota}_2', \mathbf{0}_3'\right)'$, $\rho_z = 0$, $c_\alpha = 0.2$, and $c_\gamma = 0.6$. The results for mae and CI length for the HT_{2k_z} estimator are very similar to those reported in Guo et al. (2018). There are some differences in coverage probabilities, but this is due to the fact that they report results from only 500 Monte Carlo repetitions, whereas we do 10,000 replications. The results show again a better performance of the CI_{sar} estimator in terms of mae and coverage probability compared to the HT estimators, although the difference are overall smaller than those presented in Table 2 due to the smaller number of instruments.

Table A1: Estimation Results, $k_z = 7$

	mae	coverage	CI length	$\widehat{\mathcal{A}}_n$	p_{or}	p_{allinv}
$n = 500$						
2SLS or	0.029	0.949	0.169	4.000	1.000	1.000
2SLS	0.143	0.002	0.110	0.000	0.000	0.000
HT_{4k_z}	0.136	0.059	0.114	0.441	0.000	0.000
HT_{2k_z}	0.120	0.194	0.127	1.691	0.000	0.004
CI_{sar}	0.102	0.291	0.127	1.756	0.001	0.001
$n = 1000$						
2SLS or	0.020	0.946	0.119	4.000	1.000	1.000
2SLS	0.144	0.000	0.078	0.000	0.000	0.000
HT_{4k_z}	0.123	0.076	0.087	1.405	0.000	0.001
HT_{2k_z}	0.096	0.266	0.120	3.454	0.026	0.113
CI_{sar}	0.071	0.332	0.099	2.674	0.044	0.044
$n = 2000$						
2SLS or	0.015	0.946	0.084	4.000	1.000	1.000
2SLS	0.143	0.000	0.055	0.000	0.000	0.000
HT_{4k_z}	0.088	0.206	0.088	3.657	0.039	0.143
HT_{2k_z}	0.040	0.590	0.098	4.236	0.385	0.601
CI_{sar}	0.026	0.654	0.079	3.568	0.558	0.558
$n = 5000$						
2SLS or	0.009	0.950	0.053	4.000	1.000	1.000
2SLS	0.143	0.000	0.035	0.000	0.000	0.000
HT_{4k_z}	0.010	0.892	0.055	4.054	0.900	0.953
HT_{2k_z}	0.010	0.924	0.057	4.114	0.871	0.970
CI_{sar}	0.009	0.938	0.053	4.009	0.985	0.988
$n = 10000$						
2SLS or	0.007	0.952	0.038	4.000	0.000	0.000
2SLS	0.143	0.000	0.025	0.000	0.000	0.000
HT_{4k_z}	0.007	0.951	0.038	4.020	0.986	0.999
HT_{2k_z}	0.007	0.932	0.040	4.115	0.879	0.975
CI_{sar}	0.007	0.943	0.038	4.011	0.989	0.993

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.