Methods for Selecting Valid Instrumental Variables

By

XIAORAN LIANG

School of Economics
UNIVERSITY OF BRISTOL

Student Number: 1538079
Date: February 3, 2022

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of PHD ECONOMICS in the Faculty of Social Science and Law.

Abstract

In this thesis, we consider the problem of instrumental variable (IV) selection when we have a large number of available instruments. We allow that some of these candidate instruments may be invalid in the sense that they may violate the exclusion restriction and enter the model as explanatory variables. We propose three methods for selecting the valid IVs from the candidates. The first method is the Confidence Interval (CI) method. It selects as valid the largest group of instruments where all the confidence intervals of their instrument-specific causal estimates mutually overlap with each other. It can achieve consistent IV selection under the plurality rule, which assumes that all the valid instruments form the largest group, where instruments form a group if their instrument-specific estimators converge to the same value. We apply this method to estimate the effect of Body Mass Index (BMI) on diastolic blood pressure using 96 SNPs as candidate instruments. The second method is the adaptive Lasso IV selection method, which contributes to the literature by allowing for two endogenous regressors. Under the assumption that the number of invalid instruments is smaller than half of the total number of candidate instruments minus one, we develop a median-of-medians estimator, which is $\sqrt{n}$-consistent for the causal effects. Adaptive Lasso using the median-of-medians estimator as penalty weights can select valid instruments consistently. We apply this method to estimate the direct effects of educational attainment and cognitive ability on BMI. The third method combines the agglomerative hierarchical clustering (AHC) algorithm, a commonly used statistical learning method for clustering analysis, with the downward testing procedure based on the Sargan-Hansen test for overidentifying restrictions. Under the plurality assumption, the AHC method can select valid instruments consistently. The main advantage of this method is that it performs well in the presence of weak instruments, can be extended to allow for multiple endogenous regressors, and can be used to detect potential heterogeneous causal effects. We apply this method to estimate the short- and long-term effects of immigration on wages in the US labor market.

Dedication and Acknowledgements

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Taught Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, this work is my own work. Work done in collaboration with, or with the assistance of others, is indicated as such. I have identified all material in this dissertation which is not my own work through appropriate referencing and acknowledgement. Where I have quoted or otherwise incorporated material which is the work of others, I have included the source in the references. Any views expressed in the dissertation, other than referenced material, are those of the author.

Author's Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Taught Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, this work is my own work. Work done in collaboration with, or with the assistance of others, is indicated as such. I have identified all material in this dissertation which is not my own work through appropriate referencing and acknowledgement. Where I have quoted or otherwise incorporated material which is the work of others, I have included the source in the references. Any views expressed in the dissertation, other than referenced material, are those of the author.

SIGNED:

DATE

Contents

List of Figures

List of Tables

Introduction

Most empirical studies in economics and social sciences focus on causal relationships between different quantities. Econometrics plays an important role in developing the methodology to examine such causes and effects. A major concern in identifying causal effects is endogeneity, which arises in many practical settings. Relevant to this thesis, there might be unobserved factors that confounds the relationship between the exposure and the outcome. One of the most important and commonly used solutions to endogeneity is instrumental variable (IV) estimation, which was first proposed by Wright (1928) to identify the simultaneous relationships between price and quantity.

An instrumental variable is required to satisfy two conditions: (i) It must be correlated with the endogenous exposure, and the correlation should be strong enough that it provides sufficient variation in the exposure (the *relevance condition*). (ii) The only pathway from the instrument to the outcome variable is through the exposure. This implies that the instrument should not have a direct effect on the outcome, nor should it affect the outcome through unobserved confounders (the *exclusion restriction*). If we use instruments that fail to satisfy either of these conditions, it will lead to biased estimation. This thesis focuses on violations of the exclusion restriction, and, thus, when we refer to an instrument as "valid" we mean that it satisfies the exclusion restriction, if it does not, we call it "invalid". Throughout the thesis, we explore the same core research question: When there is a large number of candidate instruments and some of them might be invalid, how can we, without complete knowledge of their validity, select only the valid instruments for estimation?

When there are more instruments than endogenous variable(s), a common tool to diagnose violations of the *exclusion restriction* is the Sargan-Hansen test for

overidentifying restrictions (Sargan, 1958; Hansen, 1982). Intuitively, this test checks if all the just-identified models estimate the same parameter values (Windmeijer, 2019), which should be the case if all the instruments are valid. Therefore, if a model is rejected by the test, it indicates that the instrument-specific causal estimands are different. If we assume a constant causal effect, this heterogeneity among the instruments indicates violation of the exclusion restriction. While we can test for violations using the Sargan-Hansen test, it is not clear how to proceed with estimation if the test detects invalid instruments. Two main strands of the literature have studied the problem of IV estimation in the presence of invalid instruments. The majority of these studies focus on the setup where invalid instruments enter the model as explanatory variables, and, thus, have non-zero direct effects on the outcome.

The first strand proposes IV estimators that are robust to including invalid instruments. For example, under the assumption that the direct effects of the invalid instruments are orthogonal to their effects on the endogenous regressor, Kolesár et al. (2015) propose a bias-corrected 2SLS estimator. In epidemiology, they also research such robust IV estimators and commonly apply them in Mendelian randomization (MR) studies, where genetic variants are used as instruments to estimate the effect of an exposure on a health-related outcome. In MR studies, a main concern is horizontal pleiotropy, which means that some of the genetic variants may directly affect the outcome. To mitigate this, the literature on pleiotropy-robust IV estimators has expanded rapidly. For example, Bowden et al. (2015b) propose the MR-Egger regression. They treat IV estimation with multiple instruments as a meta-analysis, and suggest a de-biased procedure based on Egger regression. Another example is the robust adjusted profile score estimator, as proposed by Zhao et al. (2020). They use maximum profile likelihood regression to obtain a consistent and asymptotically normal estimator. Similar to Kolesár et al. (2015), both of these methods also impose assumptions on the direct effects of the invalid instruments.

The second strand of research focuses on the idea that we can first select the valid instruments among a set of candidates. Then, after selection, we perform standard IV estimation using only the selected valid instruments. Compared with the bias-corrected solutions, these methods, generally, do not impose restric-

tions on the direct effects of the invalid instruments. Instead, their identification assumptions are usually about the number of valid instruments. Andrews (1999) propose an IV selection procedure in a generalized method of moment (GMM) setup based on the Sargan-Hansen test of over-identifying restrictions. This method becomes computationally infeasible when there are many available instruments because it requires to evaluate a large number of models. In recent years, several studies have proposed computationally feasible solutions to the IV selection problem. Kang et al. (2016) set up a linear model framework for the IV selection problem, which is adopted by most later studies, including this thesis. They propose a method based on the least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996). Windmeijer et al. (2019) develop an IV selection method based on adaptive Lasso (Zou, 2006), which provides a theoretical guarantee for consistent IV selection under the assumption that more than half of the available instruments are valid (the *majority rule*). The IV selection method developed in Guo et al. (2018) is based on hard thresholding and a voting mechanism. They achieve consistent selection under the assumption that all the valid instruments form the largest group, where instruments are said to form a group if their instrument-specific causal estimators converge to the same value. This is called the *plurality rule*, and is a relaxation of the *majority rule*.

Following the aforementioned second strand of the literature, this thesis addresses the invalid instrument problem by developing IV selection methods. The major contributions are:

1. This thesis contributes to the literature by proposing three new methods for IV selection. The first is the Confidence Interval (CI) method, which can achieve consistent selection under the plurality assumption. The second is the adaptive Lasso IV selection method, which can select the instruments consistently under the adjusted majority rule, and allows for two endogenous variables. The third is the Agglomerative Hierarchical Clustering (AHC) method. This method can also achieve consistent selection under the plurality rule, and it allows for multiple endogenous variables. It also performs well in the presence of weak instruments, and can detect potential heterogenous treatment effects.

2. This thesis shows that the developed IV selection methods can be applied to different contexts, and that they can enhance the credibility of causal inference. For example, we use the CI method in an MR study to estimate the causal effect of BMI on diastolic blood pressure using 96 SNPs as instruments. We also apply the adaptive Lasso method to a multivariate MR (MVMR) study, where there are two endogenous exposures. Apart from MR, we also apply the methods to empirical economics, where we use the AHC method to estimate the effect of immigration on wages in the US labor market. Previously, all of these applications have been studied but without using IV selection. By revisiting them and applying IV selection, we find that the post-selection causal estimates can be quite different, which shows the importance of developing data-driven methods that explicitly account for the presence of invalid instruments.

3. This thesis combines machine learning techniques with traditional econometric methods to tackle the IV selection problem, and, thus, it contributes to the methodological literature on causal machine learning. While recent years have seen a significant increase in the usage of machine learning in economics and econometrics, it can be complicated to adapt such methods to causal inference, as noted in Athey and Imbens (2019). In particular, it can in many cases be problematic to apply off-the-shelf machine learning algorithms to causal inference because their emphasis tends to be prediction (Athey and Imbens, 2019). In this thesis, we exploit the structure of the IV selection problem from different perspectives and adapting different machine learning methods accordingly. First, the IV selection problem can be viewed as the task of selecting covariates with non-zero coefficients in a sparse linear model. In this scenario, a commonly used tool is the adaptive Lasso method. We develop a median-of-medians estimators, which can serve as the penalty weights of the adaptive Lasso, such that it can be applied to IV selection with two endogenous regressors. The second way of rephrasing the IV selection problem is that we want to find clusters of instruments such that the corresponding just-identified models estimate the same parameter value. For this purpose, we explore clustering algorithms, and adopt the agglomerative

hierarchical clustering (AHC) algorithm. In terms of adapting the AHC to IV selection, we do not tune the algorithm using common practices in machine learning, as they have no theoretical guarantee for consistent selection. Instead, we exploit that the AHC algorithm selects models in a downward step-wise manner, which makes it well suited as a dimension reduction device for the downward testing procedure proposed by Andrews (1999). Furthermore, although not explicitly discussed in this thesis, the CI method can be viewed as a novel clustering algorithm. It classifies instruments as being in the same cluster if the confidence intervals of their just-identified IV estimates mutually overlap with each other. Instead of clustering based on the point estimates, as the machine learning clustering algorithms usually do, the CI method instead takes the variance of the estimators into account.

In Chapter 1 of the thesis, we introduce the CI method for IV selection when there is a single endogenous variable. This method selects as valid the largest group of instruments where all the confidence intervals of their instrument-specific causal estimates mutually overlap with each other. At the time of its introduction, only the CI method and the Hard Thresholding (HT) with Voting method achieve consistent IV selection under the plurality rule. The advantage of the CI method over the HT method is that the number of instruments selected as valid decreases monotonically for decreasing values of the tuning parameter, which is not the case for the HT method. Therefore, we can combine the CI method with the downward testing procedure proposed by Andrews (1999), which is based on the Sargan-Hansen test. In this way, the CI method selects the model with the largest number of instruments selected as valid that passes the Sargan-Hansen test. We also show that the CI method has better finite sample performance. We apply the CI method to estimate the effect of BMI on diastolic blood pressure with 96 SNPs as instruments.

By the time of the introduction of the CI method, none of the existing IV selection methods could deal with multiple endogenous variables. To fill this gap, in Chapter 2, we propose the adaptive Lasso IV selection method which allows for two endogenous exposures. For consistent selection, the adaptive Lasso requires a $\sqrt{n}$-consistent estimator as penalty weights (Zou, 2006). For this purpose, we propose a

median-of-medians estimator that satisfies the requirement for the penalty weights under the adjusted plurality rule. We apply the method to a MVMR study where we estimate the direct effects of educational attainment and cognitive ability on BMI.

There are two main disadvantage of the existing IV selection methods. First, they may impose the assumption that all the available instruments are relevant for the endogenous regressor (e.g. Kang et al., 2016; Windmeijer et al., 2019; Windmeijer et al., 2021). This can be problematic in the presence of weak instruments. For example, the CI method tends to select all the weak invalid instruments as valid, causing severe bias in the post-selection IV estimator. Second, some of the methods can only detect a single group of instruments, and the methods select this group as the set of valid instruments, while they treat all the other instruments as invalid (e.g. Kang et al., 2016; Guo et al., 2018; Windmeijer et al., 2019). However, when there are heterogeneous causal effects, there might be multiple groups of valid instruments that represent different causal mechanisms. To address these two disadvantages, in Chapter 3, we propose an IV selection method, which combines the agglomerative hierarchical clustering algorithm with the downward testing procedure based on the Sargan-Hansen test (Andrews, 1999). This method performs well in the presence of weak instruments in the sense that it always selects invalid instruments as invalid regardless of their strength. It can identify all the IV groups, which makes it possible to analyse the possible multiple causal mechanisms. Also, the method can be extended to allow for more than two endogenous regressors. We apply this method to estimate the short- and long-term effects of immigration on wages in the US labor market.

Chapter 1

The Confidence Interval Method for Selecting Valid Instrumental Variables

Abstract

We propose a new method, the confidence interval (CI) method, to select valid instruments from a larger set of potential instruments for instrumental variables (IV) estimation of the causal effect of an exposure on an outcome. Invalid instruments are such that they fail the exclusion conditions and enter the model as explanatory variables. The CI method is based on the confidence intervals of the per instrument causal effects estimates and 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 standard IV, or two-stage least squares (2SLS) estimator has oracle properties. This result is the same as for the hard thresholding with voting (HT) method of Guo et al. (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. For the CI method, we can therefore use a downward testing procedure based on the Sargan (1958) test for overidentifying restrictions and a main advantage of the CI downward testing method is that it selects the model with the largest number of instruments selected as valid that passes the Sargan test.

This chapter is co-authored with Frank Windmeijer, Fernando P. Hartwig and Jack Bowden. Frank developed the core research idea and made major contributions to the paper. Xiaoran contributed substantially to the theory and the computational work in the simulations and empirical application, including developing an R package for the method. Fernando and Jack contributed to the development of the initial research idea, and advised on the empirical application. This paper is published in the Journal of the Royal Statistical Society, Series B.

1.1 Introduction

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

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

(a) are associated with the exposure (Condition 1),

(b) have no direct pathway to the outcome (Condition 2) and

(c) are not related to unmeasured variables that affect the exposure and the outcome (Condition 3).

Condition 1 is often referred to as the *relevance* condition and Conditions 2 and 3 as the *exclusion* conditions, see Section 1.2 for details.

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

As an example, Mendelian randomisation is a technique employed in epidemiology to learn about the causal effects of modifiable health exposures on disease. It posits that genetic variants, which are known to be associated with the exposure and hence satisfy Condition 1, additionally satisfy the exclusion conditions and are only associated with the outcome through the exposure. In our Mendelian randomisation application in Section 1.8, we utilise genetic variants as potential instruments for BMI in order to determine its causal effect on diastolic blood pressure. However, a genetic variant could be an invalid instrument for various reasons,

such as linkage disequilibrium and horizontal pleiotropy, see, for example, Lawlor et al. (2008) and Hinke et al. (2016).

The so-called plurality rule holds if the set of valid instruments forms the largest group, as specified in Section 1.2. An approach for selecting the valid instruments could then be to follow Andrews (1999) and estimate the causal effect for all $2^{k_z} - (k_z + 1)$ possible subsets of at least two instruments, where k_z denotes the total number of instruments, and to select the model that minimises an information criterion based on the Sargan (1958) test of overidentifying restrictions. A large value of the Sargan test statistic is an indication that invalid instruments are present. This approach is only feasible with a relatively small number of instruments, unlike in our application where we have 96 putative genetic instruments. We therefore need dimension reduction techniques, even though we are in a setting of a fixed number of instruments k_z with a large sample size n, the setting referred to as low dimensional by Guo et al. (2018).

Following the Lasso proposal by Kang et al. (2016), Windmeijer et al. (2019) proposed an adaptive Lasso estimator in combination with a downward testing procedure based on the Sargan test as in Andrews (1999). When the majority rule holds, meaning that more than 50% of the potential instruments are valid, then this approach results in consistent selection of the invalid instruments and oracle properties of the resulting standard IV, or two-stage least squares (2SLS) estimator. This means that the limiting distribution of the estimator is the same as the oracle estimator, which is the 2SLS estimator when the set of invalid instruments is known. Guo et al. (2018) proposed a two-stage hard thresholding with voting (HT) method that results in consistent selection of the valid instruments and oracle properties of the 2SLS estimator when the weaker plurality rule holds.

In this paper we develop an alternative method, which we call the confidence interval (CI) method as presented in Section 1.3. This method simply selects as valid instruments the largest group of instruments where all confidence intervals of the instrument specific causal effect estimates overlap, with a tuning parameter varying the width of the confidence intervals. Like the Guo et al. (2018) method, we show that the CI method results in consistent selection and oracle properties of the resulting 2SLS estimator when the plurality rule holds. An advantage of the CI method is that the number of instruments selected as valid decreases monotonically

for decreasing values of the tuning parameter, which is not the case for the HT method as we discuss in Section 1.4. For the CI method, we can therefore use a downward testing procedure based on the Sargan test and a main advantage of this CI method is that it selects the model with the largest number of instruments selected as valid that passes the Sargan test.

Whilst initially making the assumptions of conditional homoskedasticity and strong instruments in Section 1.2 for ease of exposition, we discuss in Section 1.5 how to adapt the methods to deal with general forms of heteroskedasticity. We further discuss the first-stage thresholding method of Guo et al. (2018) to dealing with weak instruments in Section 3.4.2.

We evaluate the two methods in the Monte Carlo exercise in Section 1.7, for a design very similar to that in Guo et al. (2018). We find that, overall, the CI method has a better finite sample performance than the HT method in this design. In the application in Section 1.8 we find that the HT method selects too few instruments as invalid, resulting in models that are rejected by the Sargan test. By design, the CI method selects models that pass the Sargan test. It produces results very similar to the adaptive Lasso method which suggests that the majority rule is not violated in this application.

We adopt the following notation. $\boldsymbol{x}$ denotes the vector with elements x_j. For a general matrix $\boldsymbol{X}$, $\boldsymbol{X}'$ denotes its transpose. All vectors are taken as column vectors, including $\mathbf{X}_{i.}$, where the row vector $\mathbf{X}_{i.}'$ is the i-th row of the matrix $\boldsymbol{X}$. For a full column-rank matrix $\mathbf{X}$ with n rows define $\mathbf{P}_X = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$, the projection onto the column space of $\mathbf{X}$, and $\mathbf{M}_X = \mathbf{I}_n - \mathbf{P}_X$, where $\mathbf{I}_n$ is the n-dimensional identity matrix. Proofs of Lemma 1.1 and Theorems 1.3 and 1.4 in Section 1.3 are presented in the Supplementary Appendix 1.A.1.

1.2 Model and Assumptions

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

starting from the additive linear, constant effects model of Holland (1988), arrived at the observed data model for the sample given by

$$Y_i = D_i\beta + \mathbf{Z}_{i.}'\boldsymbol{\alpha} + u_i, \tag{1.1}$$

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

Definition 1.1. *If $\alpha_j = 0$, then instrument j, $j = 1, \ldots, k_z$, is valid, it satisfies both Conditions 2 and 3. If $\alpha_j \neq 0$, then instrument j is invalid.*

We present some graphical representations of the causal model and possible violations of the exclusion conditions in Appendix 1.A.3.

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

The oracle model is then given by

$$\mathbf{y} = \mathbf{d}\beta + \mathbf{Z}_{\mathcal{A}_0}\boldsymbol{\alpha}_{\mathcal{A}_0} + \mathbf{u}. \tag{1.2}$$

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

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

where ξ is defined implicitly, and is given by

$$\widehat{\beta}_{or} = \left(\widehat{\mathbf{d}}'\mathbf{M}_{Z_{\mathcal{A}_0}}\widehat{\mathbf{d}}\right)^{-1}\widehat{\mathbf{d}}'\mathbf{M}_{Z_{\mathcal{A}_0}}\mathbf{y}. \tag{1.3}$$

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

$$\sqrt{n}\left(\widehat{\beta}_{or} - \beta\right) \xrightarrow{d} N\left(0, \sigma^2_{\beta_{or}}\right), \tag{1.4}$$

where

$$\begin{aligned}\sigma^2_{\beta_{or}} &= \sigma^2_u \left(\text{plim}\left(\frac{1}{n}\widehat{\mathbf{d}}'\mathbf{M}_{Z_{\mathcal{A}_0}}\widehat{\mathbf{d}}\right)^{-1}\right), \qquad (1.5)\\ &= \sigma^2_u \left(E\left[\mathbf{Z}_{i.}D_i\right]' E\left[\mathbf{Z}_{i.}\mathbf{Z}'_{i.}\right]^{-1} E\left[\mathbf{Z}_{i.}D_i\right] - E\left[\mathbf{Z}_{\mathcal{A}_0,i.}D_i\right]' E\left[\mathbf{Z}_{\mathcal{A}_0,i.}\mathbf{Z}'_{\mathcal{A}_0,i.}\right]^{-1} E\left[\mathbf{Z}_{\mathcal{A}_0,i.}D_i\right]\right)^{-1},\end{aligned}$$

see Appendix 1.A.2 for a derivation.

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

$$D_i = \mathbf{Z}'_{i.}\boldsymbol{\gamma} + \varepsilon_{di}, \tag{1.6}$$

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

Assumption 1.1. $\boldsymbol{\gamma} = \left(E\left[\mathbf{Z}_{i.}\mathbf{Z}'_{i.}\right]\right)^{-1} E\left[\mathbf{Z}_{i.}D_i\right]$, $\gamma_j \neq 0$, $j = 1, ..., k_z$.

This is the same assumption as in Kang et al. (2016) and Windmeijer et al. (2019). Guo et al. (2018) relaxed this assumption and proposed a first-stage hard thresholding procedure to consistently select only instruments with $\gamma_j \neq 0$. We will discuss this further in Section 3.4.2 and apply this first-stage thresholding in our application.

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

$$\beta_j \equiv \frac{\Gamma_j}{\gamma_j} = \beta + \frac{\alpha_j}{\gamma_j}, \tag{1.7}$$

for $j = 1, ..., k_z$. It follows from Definition 1.1 and Assumption 1.1 that for valid instruments, $j \in \mathcal{V}_0$, $\beta_j = \beta$. Following Theorem 1 in Kang et al. (2016) and Guo et al. (2018), a necessary and sufficient condition to identify 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 1.2. $|\mathcal{V}_0| > \max_{g\neq 0} |\mathcal{V}_g|$, *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 (1.1) and (1.6), we further assume that the following standard conditions hold:

Assumption 1.3. $E[\mathbf{Z}_{i.}\mathbf{Z}'_{i.}] = \mathbf{Q}$, *with* $\mathbf{Q}$ *a finite and full rank matrix.*

Assumption 1.4. *Let* $\mathbf{w}_i = (u_i \ \varepsilon_{di})'$. *Then* $E[\mathbf{w}_i] = 0$; $E[\mathbf{w}_i\mathbf{w}'_i] = \begin{bmatrix} \sigma_u^2 & \sigma_{u\varepsilon_d} \\ \sigma_{u\varepsilon_d} & \sigma_{\varepsilon_d}^2 \end{bmatrix} = \mathbf{\Sigma}$. *The elements of* $\mathbf{\Sigma}$ *are finite.*

Assumption 1.5. $plim\,(n^{-1}\mathbf{Z}'\mathbf{Z}) = E[\mathbf{Z}_{i.}\mathbf{Z}'_{i.}] = \mathbf{Q}$; $plim\,(n^{-1}\mathbf{Z}'\mathbf{d}) = E[\mathbf{Z}_{i.}D_i]$;
$plim\,(n^{-1}\mathbf{Z}'\mathbf{u}) = E[\mathbf{Z}_{i.}u_i] = 0$; $plim\,(n^{-1}\mathbf{Z}'\varepsilon_d) = E[\mathbf{Z}_{i.}\varepsilon_{di}] = 0$;
$plim\,(n^{-1}\sum_{i=1}^n \mathbf{w}_i) = 0$; $plim\,(n^{-1}\sum_{i=1}^n \mathbf{w}_i\mathbf{w}'_i) = \mathbf{\Sigma}$.

Assumption 1.6. $\frac{1}{\sqrt{n}}\sum_{i=1}^n vec\,(\mathbf{Z}_{i.}\mathbf{w}'_i) \xrightarrow{d} N(0, \mathbf{\Sigma}\otimes \boldsymbol{Q})$ *as* $n\to\infty$.

Whilst Assumption 1.5 holds if the observations are i.i.d., as the moments are assumed to exist, these conditions further hold under various weak dependence assumptions, see Staiger and Stock (1997, p 560).

Note that conditional homoskedasticity $E[\mathbf{w}_i\mathbf{w}'_i|\mathbf{Z}_{i.}] = \mathbf{\Sigma}$ is implicit in Assumption 1.6. We make this assumption primarily for ease of exposition and will relax this in Section 1.5.

The plurality rule, Assumption 1.2, 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 and γ_j can be arbitrary and arbitrarily correlated. The CI and HT methods are robust to any such correlation. Alternatively, the methods of Kolesár et al. (2015) and Bowden et al. (2015a) do not make the plurality assumption and can have all instruments invalid. A bias corrected 2SLS estimator is then consistent under the INstrument Strength Independent of Direct Effect (INSIDE) assumption that $Cov\,(\alpha_j, \gamma_j) = 0$, 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).

1.3 The Confidence Interval Method

From the plurality rule Assumption 1.2, it follows that consistent instrument selection procedures can be based on consistent and asymptotic normal estimators of the parameters β_j as defined in (1.7). Then groups of instruments are formed by similar estimates $\widehat{\beta}_j$, and, in large samples, the largest group will constitute the group of valid instruments under Assumption 1.2. Whilst in principle all combination of instruments could be tested separately, see Andrews (1999), in practice this may not be feasible when there are a large number of instruments. The Guo et al. (2018) method as described further in Section 1.4 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 model specifications

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

Under Assumptions 1.3-1.6 it follows that

$$\sqrt{n}\left(\begin{pmatrix}\widehat{\mathbf{\Gamma}}\\ \widehat{\boldsymbol{\gamma}}\end{pmatrix}-\begin{pmatrix}\mathbf{\Gamma}\\ \boldsymbol{\gamma}\end{pmatrix}\right)\xrightarrow{d}N\left(0,\mathbf{\Lambda}\right), \tag{1.8}$$

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

$$\widehat{\beta}_j=\frac{\widehat{\Gamma}_j}{\widehat{\gamma}_j}, \tag{1.9}$$

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 $\mathbf{Q}^{-1}$,

$$\sigma_j^2 = \frac{\tau_j^2 \mathbf{Q}_{jj}^{-1}}{\gamma_j^2}; \quad \tau_j^2 = \begin{pmatrix} 1 & -\beta_j \end{pmatrix} \mathbf{\Omega} \begin{pmatrix} 1 \\ -\beta_j \end{pmatrix}. \tag{1.10}$$

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

$$\widehat{Var}\left(\widehat{\beta}_j\right) = \frac{\widehat{\tau}_j^2 \left(\mathbf{Z}'\mathbf{Z}\right)_{jj}^{-1}}{\widehat{\gamma}_j^2}; \quad \widehat{\tau}_j^2 = \begin{pmatrix} 1 & -\widehat{\beta}_j \end{pmatrix} \widehat{\mathbf{\Omega}} \begin{pmatrix} 1 \\ -\widehat{\beta}_j \end{pmatrix}, \tag{1.11}$$

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

We show in Appendix 1.A.5 that $\widehat{\beta}_j$ is identical to the 2SLS estimator of β_j in the just-identified model

$$\mathbf{y} = \mathbf{d}\beta_j + \mathbf{Z}_{\{-j\}}\boldsymbol{\pi}^{[j]} + \mathbf{u}_j, \tag{1.12}$$

where $\mathbf{Z}_{\{-j\}} = \mathbf{Z} \backslash \{\mathbf{Z}_{\cdot j}\}$, using $\mathbf{Z}_{\cdot j}$ as the instrument for $\mathbf{d}$. This therefore implies that $\widehat{\beta}_j$ is the IV estimator for β based on instrument $\mathbf{Z}_{\cdot j}$ whilst treating all other instruments as invalid. The variance estimator $\widehat{Var}\left(\widehat{\beta}_j\right)$ as defined in (1.11) is also the same as the standard 2SLS variance estimator in the just-identified model (1.12).

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

$$ci_j\left(\psi_n\right) = \left[\widehat{\beta}_j - \widehat{v}_j\psi_n, \widehat{\beta}_j + \widehat{v}_j\psi_n\right], \tag{1.13}$$

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

Lemma 1.1. *Let the groups $\mathcal{V}_g$ be as defined in Assumption 1.2 and the confidence intervals $ci_j\left(\psi_n\right)$, $j = 1, \ldots, k_z$, as defined in (1.13). Then, under Assumptions 1.1 and 1.3-1.6, for $n \to \infty$, $\psi_n \to \infty$, $\psi_n = o\left(n^{1/2}\right)$, and $\forall g$, all confidence intervals*

$ci_j(\psi_n)$ within a group, $j \in \mathcal{V}_g$, will overlap with each other, whereas none of the confidence intervals in different groups, $ci_j(\psi_n)$, $ci_{j'}(\psi_n)$, $j \in \mathcal{V}_g$, $j' \in \mathcal{V}_{g'}$, will overlap with each other.

We can use the results of Lemma 1.1 to obtain a selection rule that consistently selects the valid instruments as valid, with the resulting 2SLS estimator having oracle properties. For any value ψ_n, classify the instruments in groups $\widehat{\mathcal{V}}_t^{over}(\psi_n)$, for $t = 1, \ldots, T(\psi_n)$, with $1 \leq T(\psi_n) \leq k_z$. For members $j \in \widehat{\mathcal{V}}_t^{over}(\psi_n)$, all $ci_j(\psi_n)$ overlap with each other. Only the largest of such groups are considered, and not their subdivisions. If for example all k_z confidence intervals overlap with each other, then $T(\psi_n) = 1$. It is clear from this definition that instruments can be members of multiple groups, and a group can be a singleton. For any value ψ_n, we then select as the group of valid instruments the largest group, denoted $\widehat{\mathcal{V}}_n$, defined as

$$\widehat{\mathcal{V}}_n := \left\{ \widehat{\mathcal{V}}_m(\psi_n) : \left| \widehat{\mathcal{V}}_m(\psi_n) \right| = \max_{t=1,\ldots,T(\psi_n)} \left| \widehat{\mathcal{V}}_t^{over}(\psi_n) \right| \right\}. \tag{1.14}$$

Note that for any value of ψ_n, there may be multiple groups with the largest number of overlapping confidence intervals. If that is the case, at this point we simply randomly select one of these in order to have a single set of instruments for each ψ_n. We will discuss selection using the Sargan test in Section 1.3.1.

The next theorem states the conditions under which the selection $\widehat{\mathcal{V}}_n$ is consistent, which follows directly from the results of Lemma 1.1, as $\mathcal{V}_0$ is the largest group by Assumption 1.2.

Theorem 1.1. *Let the $\widehat{\beta}_j$ be defined as in (1.9) and their confidence intervals as in (1.13). Let $\widehat{\mathcal{V}}_n$ be one of the largest groups of instruments for which all confidence intervals overlap with each other as defined in (1.14). For $\psi_n \to \infty$, $\psi_n = o\left(n^{1/2}\right)$, and under Assumptions 1.1-1.6 it follows that*

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

The next theorem states the oracle properties of the 2SLS estimator based on selecting $\mathbf{Z}_{\widehat{\mathcal{V}}_n}$ as the valid instruments and thus $\mathbf{Z}_{\widehat{\mathcal{A}}_n} = \mathbf{Z} \backslash \left\{ \mathbf{Z}_{\widehat{\mathcal{V}}_n} \right\}$ as the set of

invalid instruments. This result follows directly from Theorem 2 in Guo et al. (2018).

Theorem 1.2. *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 1.1, it follows that

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

For any value ψ_n the sets of overlapping confidence intervals can easily and rapidly be obtained as follows.

Algorithm 1.1. *Denote the lower and upper endpoints of* $ci_j(\psi_n)$ *as defined in (1.13) by* $cil_j(\psi_n)$ *and* $ciu_j(\psi_n)$*. Order the confidence intervals in ascending order of the lower endpoints, and use the notation* $cil_{[j]}(\psi_n)$ *and* $ciu_{[j]}(\psi_n)$ *for the ordered intervals. For* $j = 2, ..., k_z$*, let* $no_{[j]}(\psi_n) = \sum_{k=1}^{j-1} 1 \left(ciu_{[k]}(\psi_n) > cil_{[j]}(\psi_n) \right)$*. Then the largest set(s) of overlapping intervals are those associated with the maximum value of* $no_{[j]}(\psi_n)$*.*

For the sequences $\psi_n \to \infty$, $\psi_n = o\left(n^{1/2}\right)$, it follows from the results of Lemma 1.1 and Theorem 1.1 that $\widehat{\mathcal{V}}_n$ as defined in (1.14) converges to the unique set $\mathcal{V}_0$. It is therefore immaterial for consistent selection and oracle properties how we choose the set $\widehat{\mathcal{V}}_n$ for those values of ψ_n where there are multiple groups with the largest number of overlapping confidence intervals. We can extend the range of sequences ψ_n if we choose in that case the group with the minimum value of the Sargan test as we show next.

1.3.1 Sargan Test

For the oracle model (1.2),

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

with $\mathbf{X}_{\mathcal{A}_0} = \begin{bmatrix} \mathbf{d} & \mathbf{Z}_{\mathcal{A}_0} \end{bmatrix}$ and $\boldsymbol{\theta}_{\mathcal{A}_0} = \begin{pmatrix} \beta & \boldsymbol{\alpha}'_{\mathcal{A}_0} \end{pmatrix}'$, the Sargan (1958) test statistic is given by

$$S\left(\widehat{\boldsymbol{\theta}}_{\mathcal{A}_0}\right) = \frac{\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{A}_0})'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{A}_0})}{\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{A}_0})'\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{A}_0})/n}, \tag{1.15}$$

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

As $E\left[\mathbf{Z}_{i.}u_i\right] = 0$, and for $k_{\mathcal{A}_0} < k_z$, it follows under Assumptions 1.1 and 1.3-1.6 that $\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_{\mathcal{A}_0} - \boldsymbol{\theta}_{\mathcal{A}_0}\right) \xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{\Sigma}_0\right)$, with $\boldsymbol{\Sigma}_0 = \sigma_u^2 \text{plim}\left(\mathbf{X}'_{\mathcal{A}_0}\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}_{\mathcal{A}_0}/n\right)^{-1}$, and $S\left(\widehat{\boldsymbol{\theta}}_{\mathcal{A}_0}\right) \xrightarrow{d} \chi^2_{k_z - k_{\mathcal{A}_0} - 1}$. For any other selection $\mathbf{Z}_{\mathcal{A}} \neq \mathbf{Z}_{\mathcal{A}_0}$ with $k_{\mathcal{A}} \leq k_{\mathcal{A}_0}$, we have that $S\left(\widehat{\boldsymbol{\theta}}_{\mathcal{A}}\right) = O_p\left(n\right)$.

The results of the confidence interval selection method can be linked to the behaviour of the Sargan test statistic as it follows from the results of Theorems 1.1 and 1.2 that, under the conditions of Theorem 1.1, $S\left(\widehat{\boldsymbol{\theta}}_{\widehat{\mathcal{A}}_n}\right) \xrightarrow{d} \chi^2_{k_z - k_{\mathcal{A}_0} - 1}$.

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

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

$$\begin{aligned} \widehat{\mathcal{V}}^{sar}_n := \Bigg\{ & \widehat{\mathcal{V}}_m\left(\psi_n\right) : \left|\widehat{\mathcal{V}}_m\left(\psi_n\right)\right| = \max_{t=1,\ldots,T(\psi_n)} \left|\widehat{\mathcal{V}}^{over}_t\left(\psi_n\right)\right|, \\ & S\left(\widehat{\theta}_{\widehat{\mathcal{A}}_m(\psi_n)}\right) = \min_{m'=1,\ldots,M(\psi_n)} S\left(\widehat{\theta}_{\widehat{\mathcal{A}}^{max}_{m'}(\psi_n)}\right) \Bigg\}, \end{aligned} \tag{1.16}$$

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

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

Theorem 1.3. *Let the $\widehat{\beta}_j$ be defined as in (1.9) and their confidence intervals as in (1.13). Let $\widehat{\mathcal{V}}^{sar}_n$ be as defined in (1.16) and $\widehat{\mathcal{A}}^{sar}_n = \mathcal{V}\backslash\widehat{\mathcal{V}}^{sar}_n$. For $k_{\mathcal{V}_0} < k_z$, let $c_n = O\left(1\right) > 0$ be such that when $n \to \infty$, $\psi_n \to \infty$, for $\frac{\psi_n}{\sqrt{n}} \leq c_n$, $\max_{t=1,\ldots,T(\psi_n)} \left|\widehat{\mathcal{V}}^{over}_t\left(\psi_n\right)\right| \to k_{\mathcal{V}_0}$ and for $\frac{\psi_n}{\sqrt{n}} > c_n$, $\max_{t=1,\ldots,T(\psi_n)} \left|\widehat{\mathcal{V}}^{over}_t\left(\psi_n\right)\right| \to K$, with $K \geq k_{\mathcal{V}_0} + 1$. Then for $n \to \infty$, $\psi_n \to \infty$, $k_{\mathcal{V}_0} = k_z$, or $k_{\mathcal{V}_0} < k_z$ and*

$\frac{\psi_n}{\sqrt{n}} \leq c_n$, *and under Assumptions 1.1-1.6 it follows that*

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

and

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

1.3.2 Downward Testing Procedure

From the results of Theorem 1.3, we can devise a downward testing procedure as in Andrews (1999), reducing the dimension of the problem by evaluating only the models selecting the sets with the largest number of overlapping confidence intervals as valid instruments. The Andrews (1999) downward testing procedure uses the Sargan test statistic as a selection device for the consistent selection of the valid instruments. It starts with the model that selects all k_z instruments as valid. If the Sargan test rejects this model, then the procedure next evaluates the k_z models with $k_z - 1$ instruments selected as valid, treating each instrument in turn as invalid. If the minimum of the k_z Sargan test statistics does not reject the null, then the associated model is selected as the valid model. If the minimum rejects the null, then all $\binom{k_z}{2}$ models with $k_z - 2$ instruments selected as valid are evaluated. This gets repeated until a model with $k_z - k_{\mathcal{A}} - 1$ degrees of freedom has a Sargan test result that does not reject the null hypothesis. Denote the minimum of the $\binom{k_z}{k_{\mathcal{A}}}$ Sargan statistics of all possible models with $k_{\mathcal{A}}$ instruments selected as invalid by $S_{\min}(k_{\mathcal{A}})$. Let

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

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

$$\zeta_{n,k_z-k_{\mathcal{A}}-1} \to \infty \text{ for } n \to \infty, \text{ and } \zeta_{n,k_z-k_{\mathcal{A}}-1} = o(n), \tag{1.17}$$

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

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

$$\psi_{j,r}^{*}=\frac{\left|\widehat{\beta}_{j}-\widehat{\beta}_{r}\right|}{\widehat{v}_{j}+\widehat{v}_{r}}, \tag{1.18}$$

for $j=1,\ldots,k_z-1$, $r=j+1,\ldots,k_z$. From Algorithm 1.1 it follows that for $\psi_n \leq \psi_{j,r}^{*}$, $ci_j(\psi_n)$ and $ci_r(\psi_n)$ do not overlap, whereas they do when $\psi_n > \psi_{j,r}^{*}$. Let $\psi_{[k_z-1]}^{*}=\max_{j,r}\left(\psi_{j,r}^{*}\right)$. For $\psi_n > \psi_{[k_z-1]}^{*}$ all k_z confidence intervals overlap. At $\psi_n=\psi_{[k_z-1]}^{*}$, the number of overlapping confidence intervals in the largest groups drops by one to k_z-1, and there will be two groups, denoted as before as $\left\{\widehat{\mathcal{V}}_{m'}^{max}\left(\psi_{[k_z-1]}^{*}\right)\right\}$, $m'=1,\ldots,M\left(\psi_{[k_z-1]}^{*}\right)$, with $M\left(\psi_{[k_z-1]}^{*}\right)=2$. The next breakpoint where the size of the largest groups of overlapping confidence intervals is equal to k_z-2 is the minimum of the maximum of the breakpoints (1.18) in the two largest groups of size k_z-1. Denote these maximum group specific breakpoints by $\psi_{m',[k_z-2]}^{*}=\max_{\{j,r\}\in\widehat{\mathcal{V}}_{m'}^{max}\left(\psi_{[k_z-1]}^{*}\right)}\left(\psi_{j,r}^{*}\right)$, for $m'=1,2$, and the minimum by $\psi_{[k_z-2]}^{*}=\min_{m'}\left(\psi_{m',[k_z-2]}^{*}\right)$. Note that for $\psi_{[k_z-2]}^{*}<\psi_n\leq\psi_{[k_z-1]}^{*}$, the maximum group size remains k_z-1. Then at $\psi_n=\psi_{[k_z-2]}^{*}$, there will be $2\leq M\left(\psi_{[k_z-2]}^{*}\right)\leq 3$ groups with the maximum k_z-2 overlapping confidence intervals, and the next breakpoint where the size of the largest groups of overlapping confidence intervals is equal to k_z-3 is again determined by the minimum of the maxima of the breakpoints (1.18) in these groups. Repeating this, we get the k_z-2 breakpoints

$$\psi_{[2]}^{*}<\psi_{[3]}^{*}<\ldots<\psi_{[k_z-1]}^{*}, \tag{1.19}$$

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

Combining the results of Theorem 1.3 with the downward testing procedure of Andrews (1999) we get the following consistent selection and oracle properties.

Theorem 1.4. *Let the breakpoints $\left\{\psi_{[s]}^{*}\right\}_{s=2}^{k_z-1}$ be as defined in (1.19) and let $\psi_{[k_z]}^{*}=\psi_{[k_z-1]}^{*}+\delta$, for a constant $\delta>0$, so that the model with all k_z instruments selected*

as valid is included. Let

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

where $\widehat{\mathcal{A}}_n^{sar}\left(\psi_{[s]}^*\right) = \mathcal{V}\backslash\widehat{\mathcal{V}}_n^{sar}\left(\psi_{[s]}^*\right)$*, with* $\widehat{\mathcal{V}}_n^{sar}\left(\psi_{[s]}^*\right)$ *defined in (1.16), and where* $\zeta_{n,s-1}$ *satisfy the conditions stated in (1.17). Let* $\widehat{\mathcal{A}}_n^{dts} = \mathcal{V}\backslash\widehat{\mathcal{V}}_n^{dts}$*. Then under the conditions of Theorem 1.3 it follows that*

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

and

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

It follows from Theorem 1.4 that $\psi_n^* = O_p\left(n^{1/2}\right)$, as ψ_n^* is asymptotically equivalent to $\psi_{[k_{\mathcal{V}_0}]}^*$ and $\psi_{[k_{\mathcal{V}_0}]}^*/\sqrt{n}$ is asymptotically equivalent to c_n as specified in Theorem 1.3, see for details the proof in Appendix 1.A.1.

Following a result in Pötscher (1983), Andrews (1999) shows that (1.17) holds if the p-value of the Sargan test satisfies $p_n \to 0$ and $\log(p_n) = o(n)$. Therefore, instead of choosing values $\zeta_{n,s-1}$ for each s, we can choose a single sequence p_n for consistent selection. Windmeijer et al. (2019) choose as threshold p-value for the Sargan test $0.1/\log(n)$, following the suggestion of Belloni et al. (2012) and which satisfies the conditions for consistent model selection and oracle properties of the resulting 2SLS estimator.

With this strategy, there is a maximum of $k_z(k_z-1)/2$ models to be evaluated. Together with the use of Algorithm 1.1, which has a computational cost of $O(k_z \log(k_z))$, at at most $k_z - 2$ breakpoints, the computational cost of this downward testing algorithm is of the order $O(k_z^2 \log(k_z))$. We give a stepwise description of the full downward testing algorithm in Appendix 1.A.4, together with an illustration using a single generated data set.[1]

Under the plurality Assumption 1.2, the CI downward testing procedure will

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

consistently select the set of valid instruments. In any application it may well be the case that multiple sets of maximum size are found for which the Sargan test statistics do not reject the null. The method of Andrews (1999) is then to select the model with the minimum value of the Sargan test statistics for these models with the same degrees of freedom, which is replicated by $\widehat{\mathcal{V}}_n^{dts}$. In practice, however, a researcher should acknowledge the fact that there are multiple such models, which could be an indication of a violation of Assumption 1.2, and investigate their results, which could lead to additional insights on the possible pathways from instruments to exposure and from exposure to outcomes.

Whilst the CI method achieves dimension reduction by ignoring the covariances between the estimators $\widehat{\beta}_j$ when constructing the sets with overlapping confidence intervals, by using the downward Sargan based testing procedure 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.

1.4 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, see e.g. Davidson and MacKinnon (2004, pp 422-424), 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 1.4.3, 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

$$\widehat{\pi}_k^{[j]} = \widehat{\Gamma}_k - \frac{\widehat{\Gamma}_j}{\widehat{\gamma}_j}\widehat{\gamma}_k. \tag{1.20}$$

It follows from the delta method that $\sqrt{n}\left(\widehat{\pi}_k^{[j]} - \pi_k^{[j]}\right) \xrightarrow{d} N\left(0, \sigma^2_{\pi_k^{[j]}}\right)$, with $\sigma^2_{\pi_k^{[j]}} =$

$\tau_j^2 \left(\mathbf{Q}_{kk}^{-1} - 2 \left(\frac{\gamma_k}{\gamma_j} \right) \mathbf{Q}_{kj}^{-1} + \left(\frac{\gamma_k}{\gamma_j} \right)^2 \mathbf{Q}_{jj}^{-1} \right)$, where τ_j^2 is as defined in (1.10). An estimator for the variance of $\widehat{\pi}_k^{[j]}$ is therefore given by

$$\widehat{Var}\left(\widehat{\pi}_k^{[j]}\right) = \widehat{\tau}_j^2 \left((\mathbf{Z}'\mathbf{Z})_{kk}^{-1} - 2 \left(\frac{\widehat{\gamma}_k}{\widehat{\gamma}_j} \right) (\mathbf{Z}'\mathbf{Z})_{kj}^{-1} + \left(\frac{\widehat{\gamma}_k}{\widehat{\gamma}_j} \right)^2 (\mathbf{Z}'\mathbf{Z})_{jj}^{-1} \right), \tag{1.21}$$

where $\widehat{\tau}_j^2$ is as defined in (1.11), with $n\widehat{Var}\left(\widehat{\pi}_k^{[j]}\right) \xrightarrow{p} \sigma^2_{\pi_k^{[j]}}$.

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

$$t_k^{[j]} = \frac{\widehat{\pi}_k^{[j]}}{\widehat{v}_{\pi_k^{[j]}}} \tag{1.22}$$

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

$$\lim_{n\to\infty} P\left(\left|t_k^{[j]}\right| \leq \psi_n\right) = 1, \tag{1.23}$$

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

$$\lim_{n\to\infty} P\left(\left|t_k^{[j]}\right| \leq \psi_n\right) = \lim_{n\to\infty} P\left(\left| \frac{\sqrt{n}\left(\widehat{\pi}_k^{[j]} - \pi_k^{[j]}\right)}{\widehat{\sigma}_{\pi l}^{[j]}} + \frac{\sqrt{n}\pi_k^{[j]}}{\widehat{\sigma}_{\pi k}^{[j]}} \right| \leq \psi_n\right) = 0. \tag{1.24}$$

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

$$\widehat{\mathcal{V}}_n^{[j]} = \left\{ k : \left|t_k^{[j]}\right| \leq \psi_n \right\}. \tag{1.25}$$

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, for $\psi_n \to \infty$, $\psi_n = o\left(n^{1/2}\right)$, if $\beta_k = \beta_j$, $\lim_{n\to\infty} P\left(k \in \widehat{\mathcal{V}}_n^{[j]}\right) = 1$ and if $\beta_k \neq \beta_j$, $\lim_{n\to\infty} P\left(k \in \widehat{\mathcal{V}}_n^{[j]}\right) = 0$.

[2]We provide detail of the correspondence between the specification in Guo et al. (2018) and our notation in Appendix 1.A.6.

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

$$\widehat{\mathcal{V}}_M = \left\{k : VM_k > \frac{k_z}{2}\right\}. \tag{1.26}$$

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

$$\widehat{\mathcal{V}}_P = \left\{k : VM_k = \max_l VM_l\right\}. \tag{1.27}$$

Let $\widehat{\mathcal{V}}_n^{HT} = \widehat{\mathcal{V}}_M \cup \widehat{\mathcal{V}}_P$, then Guo et al. (2018, pp 13-14) show that under Assumptions 1.1-1.6 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 $\widehat{\beta}_n^{HT} = \left(\widehat{\mathbf{d}}'\mathbf{M}_{Z_{\widehat{\mathcal{A}}_n}^{HT}}\widehat{\mathbf{d}}\right)^{-1}\widehat{\mathbf{d}}'\mathbf{M}_{Z_{\widehat{\mathcal{A}}_n}^{HT}}\mathbf{y}$, $\mathbf{Z}_{\widehat{\mathcal{A}}_n}^{HT} = \mathbf{Z}\backslash\left\{\mathbf{Z}_{\widehat{\mathcal{V}}_n}^{HT}\right\}$.

1.4.1 Choice of Tuning Parameter

Guo et al. (2018) do not treat ψ_n as a classical tuning parameter and they do not specify the rate, $\psi_n \to \infty$, $\psi_n = o\left(n^{1/2}\right)$, as obtained for results (1.23) and (1.24) above. They set $\psi_n = \sqrt{2.01^2 \log\left(\max\left(k_z, n\right)\right)}$ which in the setting here with fixed k_z and $n > k_z$ would lead to $\psi_n = \sqrt{2.01^2 \log\left(n\right)}$. The motivation seems to be from the fact that there are $k_z\left(k_z - 1\right)$ statistics $t_k^{[j]}$. If they were all independent

$N(0,1)$ distributed random variables, then it follows that for an increasing number of instruments k_z,

$$\lim_{k_z \to \infty} P\left(\max_{k,j}\left(\left|t_k^{[j]}\right|\right) > \sqrt{2\log\left(k_z\left(k_z-1\right)\right)}\right) = 0, \tag{1.28}$$

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

$$E\left[\max_{k,j}\left(t_k^{[j]}\right)\right] < \sqrt{2\log\left(k_z\left(k_z-1\right)\right)}. \tag{1.29}$$

It is unclear how the result in (1.29) 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 $\psi_n = \sqrt{2.01^2 \log(n)}$ 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 $\max(k_z, n)$ 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 $\psi = \sqrt{2.01^2 \log(k_z)}$, in line with the results (1.28) and (1.29) above. 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.

1.4.2 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 1.1. 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 (1.27) then leads to selecting these four instruments all as valid, $\widehat{\mathcal{V}}_{n,2}^{HT} = \{1, 2, 3, 4\}$, and so the number of valid instruments selected here increases for a decreasing value of ψ. Because of this, the Andrews (1999) Sargan test based downward testing procedure can not be applied in general to the HT method.

Table 1.1: Examples 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

As is clear from Table 1.1, the voting mechanism can select the instruments in non-overlapping groups all as valid. One way to overcome the problem of ties in the voting matrix is to find the maximal cliques, but as this problem is np complete, Karp (1972), this negates the dimension reduction properties of the voting scheme. This problem is circumvented in the CI method, which keeps track of the groupings and selects the group of instruments with the smallest value of the Sargan test in case of ties.

Further note that for the HT method the number of instruments selected as valid can be both larger and smaller than the number of votes, as the examples in Table 1.1 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.

1.4.3 Relationship with Sargan Test

Proposition 1.A.5 in Appendix 1.A.5 shows that $t_k^{[j]}$ as defined in (1.22) can equivalently be specified as

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

after 2SLS estimation of the parameters in the just-identified model (1.12)

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

with $\mathbf{Z}_{\{-j\}} = \mathbf{Z}\backslash\{\mathbf{Z}_{.j}\}$, using $\mathbf{Z}_{.j}$ as the instrument for $\mathbf{d}$, and using the notation $\widehat{\pi}_{2sls}^{[j]} = \left(\widehat{\pi}_{k,2sls}^{[j]}\right)_{k\neq j}$. 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

$$\mathbf{y} = \mathbf{d}\beta_{jk} + \mathbf{Z}_{\{-jk\}}\pi^{[jk]} + \mathbf{u}_{jk}, \tag{1.30}$$

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 1.A.5. Denoting this Sargan statistic by S_{jk}, then under the null $H_0 : E[\mathbf{Z}_{i.}u_{jk,i}] = 0$, and under Assumptions 1.1 and 1.3-1.6, $S_{jk} \xrightarrow{d} \chi_1^2$.

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 $k_z(k_z-1)/2$ 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 $\widehat{\mathcal{V}}_n^{[j]} = \left\{k : \sqrt{S_{jk}} \leq \psi_n\right\}$.

1.5 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 $E[\mathbf{w}_i\mathbf{w}_i'|\mathbf{Z}_{i.}] = \mathbf{\Sigma}(\mathbf{Z}_{i.})$

and $E\left[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'|\mathbf{Z}_{i.}\right] = \boldsymbol{\Lambda}\left(\mathbf{Z}_{i.}\right)$, with the functions $\boldsymbol{\Sigma}\left(\mathbf{Z}_{i.}\right)$ and $\boldsymbol{\Lambda}\left(\mathbf{Z}_{i.}\right)$ unknown. Let $\widehat{\boldsymbol{\eta}} = \begin{pmatrix} \widehat{\boldsymbol{\Gamma}}' & \widehat{\boldsymbol{\gamma}}' \end{pmatrix}'$, then a robust estimator of $Var\left(\widehat{\boldsymbol{\eta}}\right)$ is given by

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

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

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 (1.2),

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

the two-step GMM estimator is given by

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

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

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

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

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

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

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

1.6 Weak Instruments

The relevance Assumption 1.1 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 variation of the exposure.

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

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

$$\left|t_{\gamma_j}\right| = \left|\frac{\widehat{\gamma}_j}{\sqrt{\widehat{Var}\left(\widehat{\gamma}_j\right)}}\right| < \omega_n, \tag{1.31}$$

with $\omega_n = \sqrt{2.01 \log\left\{\max\left(k_z, n\right)\right\}}$, 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 1.4.1, the threshold parameter is set to $\omega_n = \sqrt{2.01 \log\left(k_z\right)}$ in the R routine TSHT.R (Kang, 2018), also for the low dimensional, fixed k_z case, and we will apply this first-stage thresholding in our application.

A potential problem with this first-stage thresholding is that, as the instru-

ments are not *a priori* considered to be valid, there is a chance that invalid instruments are more likely to cross the threshold. This may occur for instruments of the type Z_2 as displayed in Figure 1.A.1 in Appendix 1.A.3. As Z_2 affects the unmeasured confounders that in turn affect the exposure, the Z_2-exposure relationship itself is confounded and could result in a stronger observed effect of the instrument on the exposure than it otherwise would have been, and a larger chance of crossing the first-stage threshold.

1.7 Some Monte Carlo Results

In order to illustrate how the CI and HT methods utilise the available information, following the discussion in Sections 1.3 and 1.4, we consider a design similar to that in Guo et al. (2018, Table 2) who considered a setting with a small number of potential instruments, $k_z = 7$, in their design where the majority rule is violated, but the plurality rule holds. We consider here such setting but with a larger number of potential instruments, $k_z = 21$. We present a replication of their $k_z = 7$ design in Appendix 1.A.7.

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 $\beta = 1$; $k_z = 21$; $\rho = 0.25$; $k_{\mathcal{A}_0} = 12$, $\boldsymbol{\alpha} = c_a\left(\boldsymbol{\iota}_6', 0.5\boldsymbol{\iota}_6', \mathbf{0}_9'\right)'$ and $\boldsymbol{\gamma} = c_\gamma \times \boldsymbol{\iota}_{k_z}$, where $\boldsymbol{\iota}_r$ is an r-vector of ones, and $\mathbf{0}_r$ is an r-vector of zeros. There are therefore 3 groups of instruments, $\mathcal{V}_{c_\alpha/c_\gamma} = \{1, 2, \ldots, 6\}$, $\mathcal{V}_{0.5c_\alpha/c_\gamma} = \{7, 8, \ldots, 12\}$ and $\mathcal{V}_0 = \{13, 14, \ldots, 21\}$. $\mathcal{V}_0$ is the largest group and so the plurality rule holds, but not the majority rule. The elements of $\boldsymbol{\Sigma}_z$ are given by $\boldsymbol{\Sigma}_{z,jk} = \rho_z^{|j-k|}$. We set $\rho_z = 0.5$ and $c_\alpha = c_\gamma = 0.4$. As in Guo et al. (2018), in this setting all instruments are

strong, and the first-stage thresholding is omitted. Note that this simple design represents invalid instruments with a direct effect on the outcome of the type Z_1 as displayed in Figure 1.A.1 in Appendix 1.A.3.

Before evaluating estimation results using the downward testing CI method and the HT method as described above, Figure 1.1 shows the frequency of selection of the oracle model for the HT and CI methods, the latter on the basis of $\widehat{\mathcal{V}}_n^{sar}(\psi)$ as defined in (1.16), for $10,000$ Monte Carlo replications, as a function of values $\psi = (0.15, 0.20, ..., 6.95, 7)$ and for a sample size of $n = 2000$. It is clear that the CI method utilises the available information better in this case and obtains a maximum frequency of selecting the oracle model of 0.98 at $\psi = 2.60$, whereas the maximum frequency for the HT method is only 0.60 at $\psi = 2.40$.

Figure 1.2 shows the average total number of instruments selected as invalid, $\left|\widehat{\mathcal{A}}_n\right|$, and the average number of invalid instruments selected as invalid as a function of 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 1.2, the number of selected instruments as invalid is not monotonically increasing for decreasing values of the threshold ψ for the HT method, as discussed in Section 1.4.2, whereas it is for the CI method.

The proposed threshold value for the HT method, $\psi_n = \sqrt{2.01^2 \log(n)} = 5.54$ is clearly too large a value in this design. The alternative choice is $\psi = \sqrt{2.01^2 \log(k_z)} = 3.51$. As shown in Figure 1.1, the probability of selecting the oracle model at this value is equal to only 0.018. Figure 1.2 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

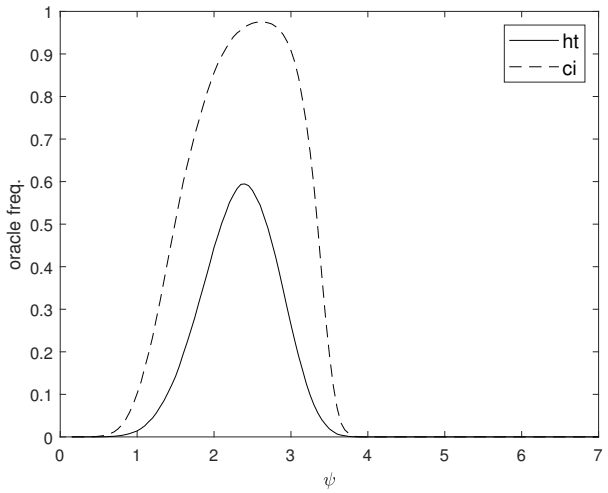


Figure 1.1: Frequency of selecting oracle model as a function of ψ. $n = 2000$, $k_z = 21$, $k_{\mathcal{A}_0} = 12$, $c_\alpha = c_\gamma = 0.4$.

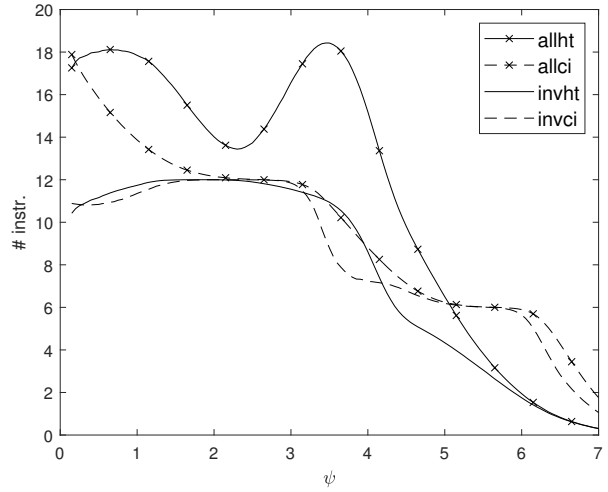


Figure 1.2: Average total number of instruments selected as invalid (all) and number of invalid instruments selected as invalid (inv) as a function of ψ. $n = 2000$, $k_z = 21$, $k_{\mathcal{A}_0} = 12$, $c_\alpha = c_\gamma = 0.4$.

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 1.2 shows estimation results for the downward testing CI method and the HT method for this design for different values of the sample size $n = 500, 1000, 2000, 5000$, for $10,000$ Monte 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 intervals. In addition, we present the average number of instruments selected as invalid, $\left|\widehat{\mathcal{A}}_n\right|$, the frequency of selecting the oracle model, p_{or}, and the frequency of selecting all invalid instruments as invalid, p_{allinv}. The 95% confidence interval is given by $\left[\widehat{\beta}_{\widehat{\mathcal{A}}_n} - 1.96\widehat{v}_{\widehat{\beta}_{\widehat{\mathcal{A}}_n}}, \widehat{\beta}_{\widehat{\mathcal{A}}_n} + 1.96\widehat{v}_{\widehat{\beta}_{\widehat{\mathcal{A}}_n}}\right]$, with $\widehat{v}_{\widehat{\beta}_{\widehat{\mathcal{A}}_n}} = \sqrt{\widehat{Var}\left(\widehat{\beta}_{\widehat{\mathcal{A}}_n}\right)}$, the 2SLS standard error.

Table 1.2: 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.

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 1.3.2 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 1.A.7 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.

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

1.8 Application: The Effect of BMI on Blood Pressure

We use data on $105,276$ individuals from the UK Biobank and investigate the effect of BMI on diastolic blood pressure, DBP. See for further details Windmeijer et al. (2019). We use 96 SNPs as potential instruments for BMI as identified in independent GWAS studies, see Locke et al. (2015). Because of skewness, we log-transformed both BMI and DBP. The linear model specification includes age, age^2 and sex, together with 15 principal components of the genetic relatedness matrix as additional explanatory variables. Because of the log-transformation, the interpretation of the causal parameter of interest β is that of an elasticity, i.e. an increase of BMI by 1% changes DPB by β%.

Table 1.3 presents the estimation results. R code for the estimation procedure is available at https://github.com/xlbristol/CIIV. We present here the results based on the assumption of conditional homoskedasticity. Robust methods as discussed in Section 1.5 produce virtually identical results. The first set of results is based on the full set of instruments, not performing a first-stage thresholding, or in other words setting $\omega_n = 0$ in (1.31). The OLS estimate of the causal parameter is equal to 0.206 (se 0.002), whereas the 2SLS estimate treating all 96 instruments as valid is much smaller at 0.087 (se 0.016). The Sargan test, however, rejects the null that all the instruments are valid with a p-value of 2.05e-19.

The 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.

Table 1.3: 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 $n = 105,276$.

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. (2019), also using a downward Sargan p-value based testing procedure. This method selects 11 instruments as invalid, resulting in an estimate of 0.163 (se 0.018) and a p-value of the Sargan test of 0.013. This method has oracle properties if more than 50% of the instruments are valid, an assumption that does not appear to be invalid given the estimation results of the 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 3.4.2. We can therefore apply the first-stage thresholding in order to exclude these instruments from consideration.

The second set of results presented in Table 1.3 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 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 2 instruments as invalid 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. The 9 instruments selected as invalid for $\omega_n = 3.03$ are a subset of the 13 instruments selected for $\omega_n = 0$ for CI_{sar}. For the Lasso procedure, 8 of the 9 instruments selected as invalid for $\omega_n = 3.03$ were also selected as invalid for $\omega_n = 0$.

Figure 1.A.4 in Appendix 1.A.8 displays the confidence intervals for the $\omega_n = 3.03$, $k_z = 62$ case at the selected final breakpoint $\psi_n^* = 2.35$. Only one of the instruments selected as invalid has a positive estimate for the causal effect, whereas the other 8 have negative estimates, resulting in a larger estimate of the causal effect when these instruments are treated as invalid.

In order to compare the results to those found by Zhao et al. (2019) we also performed the analysis on the untransformed BMI and DPB variables. The results for OLS in this case are 0.559 (0.0062), for 2SLS, 0.248 (0.0452), and for CI_{sar}, 0.568 (0.0565), with 13 instruments found to be invalid. For the pre-selected $k_z = 62$ case, the results for 2SLS are 0.244 (0.0469), and for CI_{sar}, 0.494 (0.0557), with 9 instruments found to be invalid. In the latter case these invalid instruments are identical to the ones found above, but this is not the case when $k_z = 96$. Again these results suggest that the original OLS results suffer much less from unobserved

confounding bias than the naive 2SLS estimator suggests. These results are similar to those found in the two-sample summary data analysis of Zhao et al. (2019), who found profile score, RAPS, IVW and weighted median estimates of 0.601 (0.054), 0.402 (0.106), 0.514 (0.102) and 0.472 (0.176) respectively in their analysis with 160 SNPs as potential instruments.

1.9 Conclusion and Discussion

We have shown that the confidence interval method for selecting the set of valid instruments from a putative set of instruments that may include invalid ones for an instrumental variables analysis is a viable alternative to the hard thresholding method and the adaptive Lasso method when the plurality rule holds. The methods developed for selecting invalid instruments thus far have only considered a single endogenous treatment variable. Recent analyses have considered models with multiple treatments, see e.g. Sanderson et al. (2019) for an examination of multivariable Mendelian randomisation. An extension of the instrument selection methods for multiple treatment models is not straightforward. When the majority rule applies, the adaptive Lasso method can be utilised by constructing an initial consistent median-of-medians estimator, see Chapter 2. For the HT and CI methods, such an extension is the subject of future research.

1.A Appendix

1.A.1 Proofs of Lemma 1.1 and Theorems 1.3 and 1.4

Lemma 1.1

Proof. It suffices to show the result for $\mathcal{V}_0$. Invoking Assumption 1.2, it follows that $|\mathcal{V}_0| > 1$. Consider a valid instrument Z_q, $q \in \mathcal{V}_0$, and invalid instrument Z_s, $s \in \mathcal{A}_0$. Consider wlog the case with $\beta_s > \beta$. Let $\widehat{\sigma}_j = \sqrt{n}\widehat{v}_j$. 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

$$\lim_{n\to\infty} P\left(\widehat{\beta}_q + \widehat{v}_q\psi_n < \widehat{\beta}_s - \widehat{v}_s\psi_n\right) = \lim_{n\to\infty} P\left(\widehat{\beta}_q - \widehat{\beta}_s < -\psi_n\left(\widehat{v}_q + \widehat{v}_s\right)\right) \qquad (1.A.1)$$

$$\begin{aligned} &= \lim_{n\to\infty} P\left(\frac{\sqrt{n}\left(\left(\widehat{\beta}_q - \widehat{\beta}_s\right) - (\beta - \beta_s)\right)}{\sqrt{\sigma_q^2 + \sigma_s^2 - 2\sigma_{qs}}} < -\psi_n\frac{\widehat{\sigma}_q + \widehat{\sigma}_s}{\sqrt{\sigma_q^2 + \sigma_s^2 - 2\sigma_{qs}}} + \frac{\sqrt{n}\left(\beta_s - \beta\right)}{\sqrt{\sigma_q^2 + \sigma_s^2 - 2\sigma_{qs}}}\right)\\ &= 1, \end{aligned}$$

as

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

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

For any pair of valid instruments Z_q and Z_k, $q, k \in \mathcal{V}_0$, we have that the confidence intervals will overlap with probability 1 when $n \to \infty$, as

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

as $\psi_n \to \infty$. The above results hold for all groups $\mathcal{V}_g$. Therefore, for $n \to \infty$, $\psi_n \to \infty$, $\psi_n = o\left(n^{1/2}\right)$, all confidence intervals of 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. □

Theorem 1.3

Proof. It follows directly from Theorem 1.1 that $\widehat{\mathcal{V}}_n^{sar} \to \mathcal{V}_0$ when $n \to \infty$, $\psi_n \to \infty$ and $\psi_n = o\left(n^{1/2}\right)$. For $\psi_n = O\left(n^{1/2+\delta}\right)$ for $\delta > 0$, it follows from Lemma 1.1 that the confidence intervals of all k_z instruments overlap with each other. As $\max_{t=1,\ldots,T(\psi_n)} \left|\widehat{\mathcal{V}}_t^{over}(\psi_n)\right|$ is nondecreasing in ψ_n, there exists c_n satisfying the conditions stated if $k_{\mathcal{V}_0} < k_z$. From Lemma 1.1 it follows that the sequence c_n depends on the values of β, β_s, $\widehat{\sigma}_q$ and $\widehat{\sigma}_s$, $s = 1, \ldots, k_{\mathcal{A}_0} \in \mathcal{A}_0$, $q = 1, \ldots, k_{\mathcal{V}_0} \in \mathcal{V}_0$. Then, for $\frac{\psi_n}{\sqrt{n}}$ increasing to c_n, for $n \to \infty$ we have that all confidence intervals of the valid instruments will overlap with each other, but there will be at least one and a maximum of $k_z - k_{\mathcal{V}_0}$ additional sets of $k_{\mathcal{V}_0}$ overlapping confidence intervals from a mixture of valid and invalid instruments. For these mixtures, the Sargan test statistic is $O_p(n)$ and only for the set of valid instruments it has a limiting $\chi^2_{k_z - k_{\mathcal{A}_0} - 1}$ distribution. Therefore, $\lim_{n\to\infty} P\left(\min_{m'=1,\ldots,M(\psi_n)} S\left(\widehat{\theta}_{\widehat{\mathcal{A}}_{m'}^{max}(\psi_n)}\right) = S\left(\widehat{\theta}_{\mathcal{A}_0}\right)\right) = 1$ and the results follow. □

Theorem 1.4

Proof. For $k_{\mathcal{V}_0} = k_z$, the Sargan test statistic for the model with the full set of instruments selected as valid has a limiting $\chi^2_{k_z-1}$ distribution and hence in the model with the largest value of $\psi^*_{[s]} = \psi^*_{[k_z]}$ it follows that

$$\lim_{n\to\infty} P\left(S\left(\widehat{\theta}_{\widehat{\mathcal{A}}_n^{sar}\left(\psi^*_{[k_z]}\right)}\right) < \zeta_{n,k_z-1}\right) = 1,$$

as $\zeta_{n,k_z-1} \to \infty$.

Next consider $k_{\mathcal{V}_0} < k_z$. For any $s > k_{\mathcal{V}_0}$ the set of instruments selected as valid is a mixture of valid and invalid instruments and hence

$$\lim_{n\to\infty, s>k_{\mathcal{V}_0}} P\left(S\left(\widehat{\theta}_{\widehat{\mathcal{A}}_n^{sar}\left(\psi^*_{[s]}\right)}\right) < \zeta_{n,s-1}\right) = 0,$$

as $S\left(\widehat{\theta}_{\widehat{\mathcal{A}}_n^{sar}\left(\psi^*_{[s]}\right)}\right) = O_p(n)$ for $s > k_{\mathcal{V}_0}$, and $\zeta_{n,s-1} = o(n)$.

For $s = k_{\mathcal{V}_0}$, by construction we have that $\max_{t=1,\ldots,T(\psi^*_{[k_{\mathcal{V}_0}]})} \left|\widehat{\mathcal{V}}_t^{over}\left(\psi^*_{[k_{\mathcal{V}_0}]}\right)\right| = k_{\mathcal{V}_0}$, and for $\psi_n > \psi^*_{[k_{\mathcal{V}_0}]}$, $\max_{t=1,\ldots,T(\psi_n)} \left|\widehat{\mathcal{V}}_t^{over}(\psi_n)\right| > k_{\mathcal{V}_0}$. Further, it follows from Lemma 1.1 that $\psi^*_{[k_{\mathcal{V}_0}]} = O_p\left(n^{1/2}\right)$ and therefore $\psi^*_{[k_{\mathcal{V}_0}]}/\sqrt{n}$ satisfies the conditions for c_n as described in Theorem 1.3. For $n \to \infty$, for $\psi_n = \psi^*_{[k_{\mathcal{V}_0}]}$, the set of groups with the maximum $k_{\mathcal{V}_0}$ overlapping confidence intervals, $\left\{\widehat{\mathcal{V}}_{m'}^{max}\left(\psi^*_{[k_{\mathcal{V}_0}]}\right)\right\}$, $m' = 1, \ldots, M\left(\psi^*_{[k_{\mathcal{V}_0}]}\right)$, therefore includes $\mathcal{V}_0$ and contains other sets that are mixtures of valid and invalid instruments. It follows that

$$\lim_{n\to\infty} P\left(S\left(\widehat{\theta}_{\widehat{\mathcal{A}}_n^{sar}(\psi^*_{[k_{\mathcal{V}_0}]})}\right) < \zeta_{n,k_{\mathcal{V}_0}-1}\right) = 1,$$

and

$$\lim_{n\to\infty} P\left(S\left(\widehat{\theta}_{\widehat{\mathcal{A}}_n^{sar}(\psi^*_{[k_{\mathcal{V}_0}]})}\right) = S\left(\widehat{\theta}_{\mathcal{A}_0}\right)\right) = 1.$$

The breakpoint $\psi^*_{[k_{\mathcal{V}_0}]}$ is the maximum of the $\left\{\psi^*_{[s]}\right\}_{s=2}^{k_z}$ for which this occurs and hence,

$$\lim_{n\to\infty} P\left(S\left(\widehat{\theta}_{\widehat{\mathcal{A}}_n^{sar}(\psi_n^*)}\right) = S\left(\widehat{\theta}_{\mathcal{A}_0}\right)\right) = 1,$$

and the results follow. □

1.A.2 Limiting Distribution of Oracle 2SLS Estimator $\widehat{\beta}_{or}$.

The oracle model is given by

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

with $\mathbf{X}_{\mathcal{A}_0} = [\mathbf{d}\ \mathbf{Z}_{\mathcal{A}_0}]$ and $\boldsymbol{\theta}_{\mathcal{A}_0} = \left(\beta\ \boldsymbol{\alpha}'_{\mathcal{A}_0}\right)'$. The instrument matrix is given by $\mathbf{Z} = [\mathbf{Z}_{\mathcal{V}_0}\ \mathbf{Z}_{\mathcal{A}_0}]$ and the 2SLS estimator for $\boldsymbol{\theta}_{\mathcal{A}_0}$ is

$$\widehat{\boldsymbol{\theta}}_{\mathcal{A}_0} = \left(\mathbf{X}'_{\mathcal{A}_0}\mathbf{P}_Z\mathbf{X}_{\mathcal{A}_0}\right)^{-1}\mathbf{X}'_{\mathcal{A}_0}\mathbf{P}_Z\mathbf{y}.$$

For $k_{\mathcal{A}_0} < k_z$ and under Assumptions 1.1 and 1.3-1.6, as $n \to \infty$,

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_{\mathcal{A}_0} - \boldsymbol{\theta}_{\mathcal{A}_0}\right) \xrightarrow{d} N\left(0, \sigma_u^2 \text{plim}\left(\left(\frac{1}{n}\mathbf{X}'_{\mathcal{A}_0}\mathbf{P}_Z\mathbf{X}_{\mathcal{A}_0}\right)^{-1}\right)\right),$$

see e.g. Davidson and MacKinnon (2004, p 322).

As $\mathbf{P}_Z\mathbf{X}_{\mathcal{A}_0} = [\mathbf{P}_Z\mathbf{d}\ \mathbf{Z}_{\mathcal{A}_0}]$,it follows that

$$\mathbf{X}'_{\mathcal{A}_0}\mathbf{P}_Z\mathbf{X}_{\mathcal{A}_0} = \begin{bmatrix} \mathbf{d}'\mathbf{P}_Z \\ \mathbf{Z}_{\mathcal{A}_0} \end{bmatrix}[\mathbf{P}_Z\mathbf{d}\ \mathbf{Z}_{\mathcal{A}_0}] = \begin{bmatrix} \mathbf{d}'\mathbf{P}_Z\mathbf{d} & \mathbf{d}'\mathbf{Z}_{\mathcal{A}_0} \\ \mathbf{Z}'_{\mathcal{A}_0}\mathbf{d} & \mathbf{Z}'_{\mathcal{A}_0}\mathbf{Z}_{\mathcal{A}_0} \end{bmatrix}.$$

For the inverse of a partitioned matrix we have

$$\left(\mathbf{X}'_{\mathcal{A}_0}\mathbf{P}_Z\mathbf{X}_{\mathcal{A}_0}\right)^{-1} = \begin{bmatrix} e & \mathbf{f}' \\ \mathbf{f} & \mathbf{G} \end{bmatrix};\ e = \left(\mathbf{d}'\mathbf{P}_Z\mathbf{d} - \mathbf{d}'\mathbf{P}_{Z_{\mathcal{A}_0}}\mathbf{d}\right)^{-1},$$

and, as $\mathbf{P}_{Z_{\mathcal{A}_0}}\mathbf{P}_Z = \mathbf{P}_{Z_{\mathcal{A}_0}}$, it follows that $\mathbf{d}'\mathbf{P}_Z\mathbf{d} - \mathbf{d}'\mathbf{P}_{Z_{\mathcal{A}_0}}\mathbf{d} = \widehat{\mathbf{d}}'\widehat{\mathbf{d}} - \widehat{\mathbf{d}}'\mathbf{P}_{Z_{\mathcal{A}_0}}\widehat{\mathbf{d}} = \widehat{\mathbf{d}}'\mathbf{M}_{Z_{\mathcal{A}_0}}\widehat{\mathbf{d}}$.

It therefore follows that, as $n \to \infty$,

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{or} - \beta\right) \xrightarrow{d} N\left(0, \sigma_u^2 \text{plim}\left(\left(\frac{1}{n}\widehat{\mathbf{d}}'\mathbf{M}_{Z_{\mathcal{A}_0}}\widehat{\mathbf{d}}\right)^{-1}\right)\right),$$

with, under Assumption 5,

$$\begin{aligned} \text{plim}\left(\frac{1}{n}\widehat{\mathbf{d}}'\mathbf{M}_{Z_{\mathcal{A}_0}}\widehat{\mathbf{d}}\right) &= \text{plim}\left(\frac{1}{n}\left(\mathbf{d}'\mathbf{P}_Z\mathbf{d} - \mathbf{d}'\mathbf{P}_{Z_{\mathcal{A}_0}}\mathbf{d}\right)\right) \\ &= E\left[\mathbf{Z}_{i\cdot}D_i\right]' E\left[\mathbf{Z}_{i\cdot}\mathbf{Z}'_{i\cdot}\right]^{-1} E\left[\mathbf{Z}_{i\cdot}D_i\right] \\ &\quad - E\left[\mathbf{Z}_{\mathcal{A}_0,i\cdot}D_i\right]' E\left[\mathbf{Z}_{\mathcal{A}_0,i\cdot}\mathbf{Z}'_{\mathcal{A}_0,i\cdot}\right]^{-1} E\left[\mathbf{Z}_{\mathcal{A}_0,i\cdot}D_i\right]. \end{aligned}$$

1.A.3 Correlated Instruments and Violations of the Exclusion Conditions

We next give some graphical representations of the causal model and possible violations of the exclusion conditions. First consider the directed acyclic graph (DAG), Pearl (2009), shown in Figure 1.A.1. The nodes are the treatment D, outcome Y, unobserved confounders UC and three putative instruments Z_1, Z_2 and Z_3. The directed edges represent the direction of effect. There is a causal effect from D to Y, which is confounded by UC. The three instruments all satisfy Condition 1, as they all have directed edges to D, but there is a direct pathway form Z_1 to Y and an indirect pathway from Z_2 to Y via UC. Hence Z_1 and Z_2 are invalid instruments, with $\alpha_1 \neq 0$ and $\alpha_2 \neq 0$ in model (1.1). Z_3 is a valid instrument after conditioning on Z_1 and Z_2, i.e. after including Z_1 and Z_2 as explanatory variables in model (1.1). Note that this is independent of the correlation structure between the instruments, as indicated by the bidirectional edges, as conditioning on Z_1 and Z_2 blocks any pathway from Z_3 to Y other than via D. Hence Z_3 then satisfies both Conditions 2 and 3, $\alpha_3 = 0$, and is a valid instrument.

Next consider violation of exclusion Condition 3 with unobserved confounders affecting the instrument and outcome. In Figure 1.A.2, Z_4 is an invalid instrument because unobserved confounding UC affects both Z_4 and the outcome Y.

In the left panel of Figure 1.A.2, there is a directed effect from Z_4 to Z_3, and after conditioning on Z_4, Z_3 is a valid instrument as then the pathway from UC to Z_3 is blocked. In the right panel of Figure 1.A.2 there is a directed effect from Z_3 to Z_4. This results in Z_4 being a collider and hence rendering Z_3 invalid after conditioning on Z_4. Both α_3 and α_4 are then different from zero in model (1.1).

1.A.4 Downward Testing Algorithm and Illustration

The algorithm for the downward testing CI method is as follows:

1. Select all instruments as valid and compute the Sargan test statistic. Choose $p_n \to 0$, $\log(p_n) = o(n)$, for example $p_n = 0.1/\log(n)$. If the p-value of the Sargan test statistic, using the $\chi^2_{k_z-1}$ distribution, is larger than p_n, then stop, there is no evidence that any of the instruments are invalid. If the

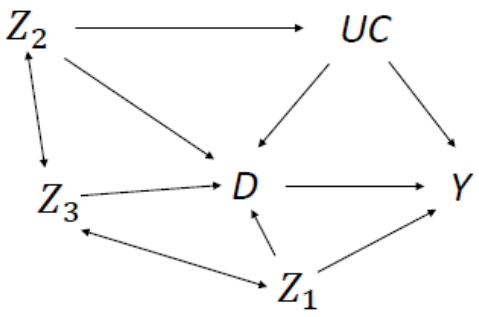


Figure 1.A.1: Directed Acyclic Graph. UC represents unmeasured confounders. Z_1 and Z_2 are invalid instruments. Z_3 is a valid instrument after conditioning on Z_1 and Z_2, independent of any directional correlations between the instruments.

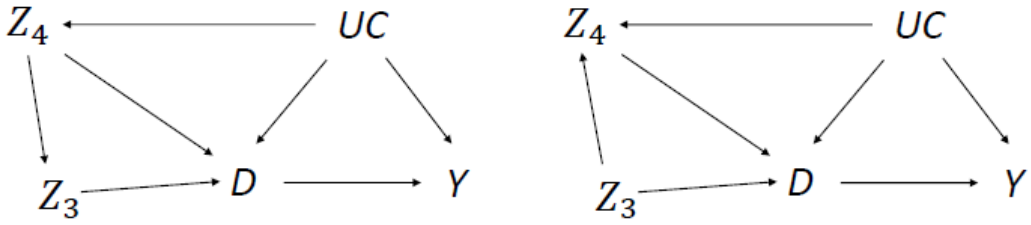


Figure 1.A.2: Instrument Z_4 is invalid. In the left panel, Z_3 is a valid instrument after conditioning on Z_4. In the right panel, Z_3 becomes an invalid instrument after conditioning on Z_4.

p-value is smaller than p_n then go to step 2.

2. Calculate the set of $k_z\left(k_z-1\right)/2$ possible breakpoints $\psi_{j,r}^{*}=\left|\widehat{\beta}_j-\widehat{\beta}_r\right|/\left(\widehat{v}_j+\widehat{v}_r\right)$, $j=1,\ldots,k_z-1$, $r=j+1,\ldots,k_z$. Let $s=k_z-1$ and set $\psi_n=\psi_{[s]}^{*}=\max_{j,r}\left(\psi_{j,r}^{*}\right)$.

3. Find the groups with the largest number of overlapping confidence intervals using Algorithm 1.1.

4. Compute the Sargan test statistics for the groups found in step 3. If the p-value of the minimum of these Sargan test statistics, using the χ_{s-1}^2 distribution, is larger than p_n, then stop and select the associated group as valid

instruments. If the p-value is smaller then p_n, then go to step 5.

5. Set $s = s - 1$ and set $\psi_n = \psi^*_{[s]}$, which is the smallest of the maximum breakpoints for the groups found in step 3, see (1.19).

6. Repeat steps 3-5 until the p-value in step 4 is larger than p_n and $s > 1$. If no such model can be found, then no group of instruments can be classified as valid.

The computational cost of the ordering Algorithm 1.1 is $O(k_z \log(k_z))$, and therefore the computational cost of this algorithm is of the order $O(k_z^2 \log(k_z))$.

We illustrate the procedure using one set of data generated from a design with $k_z = 7$, $n = 1000$, $k_{\mathcal{A}_0} = 4$, $\alpha = (0.8, 0.4, 0.4, 0.2, \mathbf{0}_3')$ and $\gamma = (0.8, 0.4, 0.6, 0.8, 0.8, 0.4, 0.4)$. All other parameter settings are as in the Monte Carlo design in Section 1.7. In this setting there are four groups of instruments, $\mathcal{V}_0 = \{5, 6, 7\}$, $\mathcal{V}_{0.25} = \{4\}$, $\mathcal{V}_{0.67} = \{3\}$ and $\mathcal{V}_1 = \{1, 2\}$. For this sample size, we set $p_n = 0.1/\log(n) = 0.015$.

For the generated data set, the Sargan test treating all instruments as valid rejects the null. The estimates $\widehat{\beta}_j$, standard errors $\widehat{v}_j$ and breakpoints $\psi^*_{j,r} = \left|\widehat{\beta}_j - \widehat{\beta}_r\right| / (\widehat{v}_j + \widehat{v}_r)$ are given by

$\widehat{\beta}_j$	$\widehat{v}_j$				$\psi^*_{j,r}$			
2.08	0.058	$j \backslash r$	2	3	4	5	6	7
1.84	0.111	1	1.46	3.29	7.30	10.25	7.06	7.49
1.67	0.069	2		0.95	3.41	5.36	4.41	4.13
1.28	0.052	3			3.18	5.80	4.48	4.14
0.98	0.050	4				3.02	2.71	1.76
0.81	0.122	5					0.96	0.58
1.05	0.080	6						1.19

Therefore $\psi^*_{[k_z-1]} = \psi^*_{[6]} = 10.25$, and for $\psi_n = \psi^*_{[6]}$ we have two groups with six overlapping confidence intervals using Algorithm 1.1, as displayed in Figure 1.A.3. Table 1.A.1, together with Figure 1.A.3, shows the step-by-step results of the downward testing CI method, here resulting in correctly selecting instruments $\{5, 6, 7\}$ as the set of valid instruments.

Table 1.A.1: Step-by-step results of algorithm

s	ψ_n	$\widehat{\mathcal{V}}_{m'}^{max}$	$\widehat{\mathcal{A}}_{m'}^{max}$	p-value Sargan	$\psi^*_{m',[s-1]}$	$\psi^*_{[s-1]}$, (j,r)
7	>10.25	$\{1,2,3,4,5,6,7\}$	$\emptyset$	2.20e-92	10.25	10.25, $(1,5)$
6	10.25	$\{2,3,4,5,6,7\}$	$\{1\}$	5.61e-41	5.80	5.80, $(3,5)$
		$\{1,2,3,4,6,7\}$	$\{5\}$	4.88e-55	7.49	
5	5.80	$\{2,4,5,6,7\}$	$\{1,3\}$	1.98e-17	5.36	
		$\{2,3,4,6,7\}$	$\{1,5\}$	1.19e-25	4.48	4.48, $(3,6)$
4	4.48	$\{4,5,6,7\}$	$\{1,2,3\}$	1.41e-05	3.02	3.02, $(4,5)$
		$\{2,4,6,7\}$	$\{1,3,5\}$	3.03e-15	4.41	
		$\{2,3,4,7\}$	$\{1,5,6\}$	2.82e-13	4.14	
3	3.02	$\{5,6,7\}$	$\{1,2,3,4\}$	$0.36 > 0.015$		
		$\{4,5,7\}$	$\{1,2,3,6\}$	1.56e-5		

1.A.5 Alternative Representation of Estimators $\widehat{\beta}_j$ and $\widehat{\pi}_k^{[j]}$

Consider the model specifications

$$\mathbf{y} = \mathbf{d}\beta_j + \mathbf{Z}_{\{-j\}}\pi^{[j]} + \mathbf{u}_j, \tag{1.A.2}$$

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 (1.1) and (1.6) it follows that

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

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

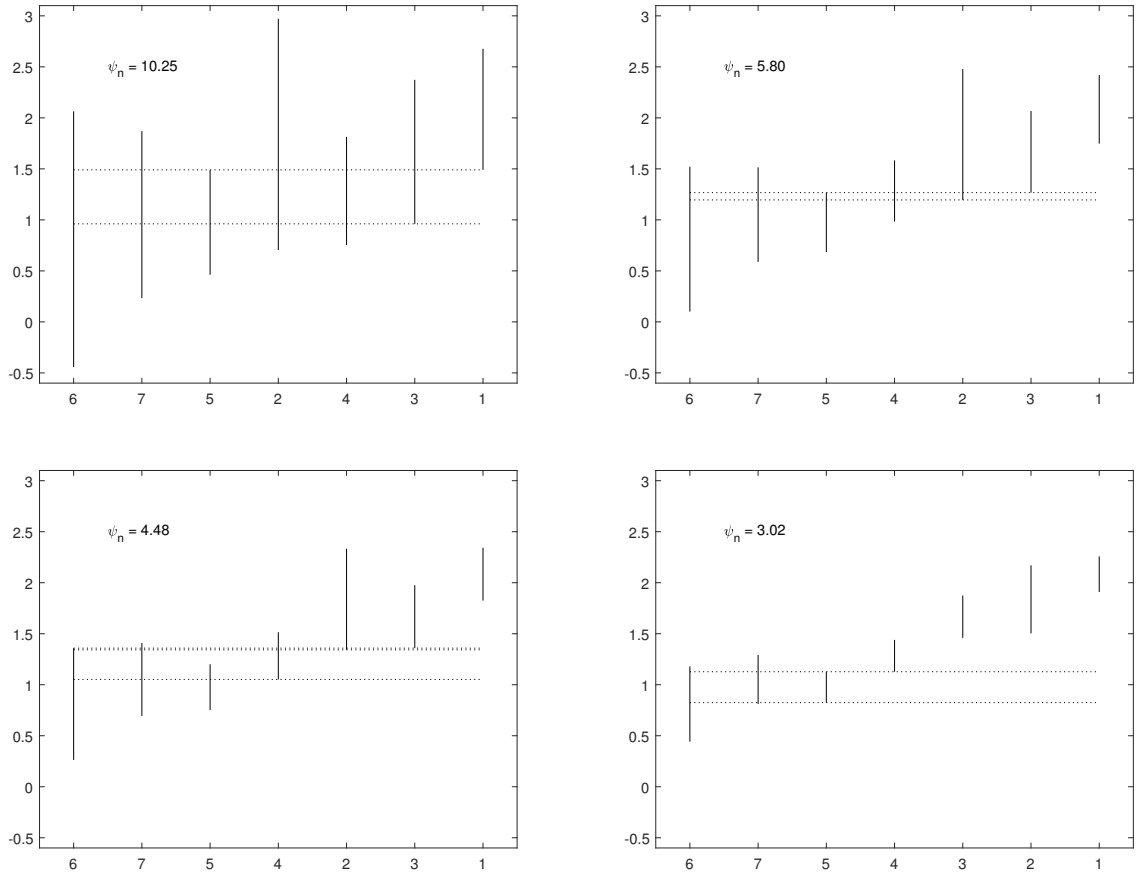


Figure 1.A.3: Confidence intervals for values of $\psi_n = \psi^*_{[s]}$, for $s = 6, ..., 3$, with largest groups of overlapping confidence intervals indicated by intersections with dotted horizontal lines. Instrument number j on x-axis.

and $\pi^{[3]} = \begin{pmatrix} \pi_1^{[3]} & \pi_2^{[3]} \end{pmatrix}'$.

For estimating the parameters in (1.A.2) 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 $\theta_j = \begin{pmatrix} \beta_j & \pi^{[j]\prime} \end{pmatrix}'$ is given by

$$\widehat{\theta}_{j,2sls} = \left(\mathbf{X}_j'\mathbf{P}_Z\mathbf{X}_j\right)^{-1}\mathbf{X}_j'\mathbf{P}_Z\mathbf{y} = \left(\mathbf{Z}'\mathbf{X}_j\right)^{-1}\mathbf{Z}'\mathbf{y}, \tag{1.A.3}$$

and so

$$\widehat{\beta}_{j,2sls} = \widehat{\theta}_{j,2sls,1}; \tag{1.A.4}$$

$$\widehat{\pi}^{[j]}_{k,2sls} = \widehat{\theta}_{j,2sls,k^*}, \tag{1.A.5}$$

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

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

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

$$\widehat{Var}\left(\widehat{\beta}_{j,2sls}\right) = \widehat{\sigma}^2_{u_j}\left(\mathbf{X}_j'\mathbf{P}_Z\mathbf{X}_j\right)^{-1}_{11} \tag{1.A.7}$$

$$\widehat{Var}\left(\widehat{\pi}^{[j]}_{k,2sls}\right) = \widehat{\sigma}^2_{u_j}\left(\mathbf{X}_j'\mathbf{P}_Z\mathbf{X}_j\right)^{-1}_{k^*,k^*}. \tag{1.A.8}$$

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

Consider the estimators $\widehat{\beta}_j$, $\widehat{\beta}_{j,2sls}$, $\widehat{\pi}^{[j]}_k$ and $\widehat{\pi}^{[j]}_{k,2sls}$ as given in (1.9), (1.A.4), (1.20) and (1.A.5) respectively, and the variance estimators $\widehat{Var}\left(\widehat{\beta}_j\right)$, $\widehat{Var}\left(\widehat{\beta}_{j,2sls}\right)$, $\widehat{Var}\left(\widehat{\pi}^{[j]}_k\right)$ and $\widehat{Var}\left(\widehat{\pi}^{[j]}_{k,2sls}\right)$ as defined in (1.11), (1.A.7), (1.21) and (1.A.8) respectively. Then $\widehat{\beta}_j = \widehat{\beta}_{j,2sls}$; $\widehat{\pi}^{[j]}_k = \widehat{\pi}^{[j]}_{k,2sls}$; $\widehat{Var}\left(\widehat{\beta}_j\right) = \widehat{Var}\left(\widehat{\beta}_{j,2sls}\right)$; and $\widehat{Var}\left(\widehat{\pi}^{[j]}_k\right) = \widehat{Var}\left(\widehat{\pi}^{[j]}_{k,2sls}\right)$.

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

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\Gamma} + \varepsilon_y$$

$$\mathbf{d} = \mathbf{Z}\gamma + \varepsilon_d,$$

with the OLS estimators denoted $\widehat{\boldsymbol{\Gamma}}$ and $\widehat{\gamma}$. 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 $\widehat{\pi}^{[j]}_k = \widehat{\Gamma}_k - \widehat{\beta}_j\widehat{\gamma}_k = \widehat{\Gamma}_k - \frac{\widehat{\Gamma}_j}{\widehat{\gamma}_j}\widehat{\gamma}_k$ to 0. Define $\widehat{\pi}^{[j]} = \left(\widehat{\pi}^{[j]}_1, ..., \widehat{\pi}^{[j]}_{j-1}, \widehat{\pi}^{[j]}_{j+1}, ..., \widehat{\pi}^{[j]}_{k_z}\right)'$. Let the OLS residuals be $\widehat{\varepsilon}_y = \mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\Gamma}}$ and $\widehat{\varepsilon}_d = \mathbf{y} - \mathbf{Z}\widehat{\gamma}$, and define $\widehat{\boldsymbol{\Omega}} = \frac{1}{n}\begin{pmatrix} \widehat{\varepsilon}_y & \widehat{\varepsilon}_d \end{pmatrix}'\begin{pmatrix} \widehat{\varepsilon}_y & \widehat{\varepsilon}_d \end{pmatrix}$. 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

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

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 $\gamma = \begin{pmatrix} \gamma_1 & \gamma_2' \end{pmatrix}'$ and $\mathbf{\Gamma} = \begin{pmatrix} \Gamma_1 & \mathbf{\Gamma}_2' \end{pmatrix}'$. Then consider the specification

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

Let $\mathbf{Z}^* = \begin{bmatrix} \widehat{\mathbf{d}} & \mathbf{Z}_2 \end{bmatrix}$, then $\mathbf{Z}^* = \mathbf{Z}\widehat{\mathbf{H}}$, with

$$\widehat{\mathbf{H}} = \begin{bmatrix} \widehat{\gamma}_1 & 0 \\ \widehat{\gamma}_2 & \mathbf{I}_{k_z-1} \end{bmatrix};\ \ \widehat{\mathbf{H}}^{-1} = \begin{bmatrix} \widehat{\gamma}_1^{-1} & 0 \\ -\widehat{\gamma}_2\widehat{\gamma}_1^{-1} & \mathbf{I}_{k_z-1} \end{bmatrix}.$$

The 2SLS estimator for $\theta_1 = \begin{pmatrix} \beta_1 & \pi^{[1]\prime} \end{pmatrix}'$ is given by

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

Hence

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

Let $\widehat{\mathbf{u}}_{1,2sls} = \mathbf{y} - \mathbf{d}\widehat{\beta}_{1,2sls} - \mathbf{Z}_2\widehat{\pi}^{[1]}_{2sls}$. 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 $\widehat{\theta}_{1,2sls}$ is given by

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

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

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

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 $H_0 : \pi_k^{[1]}$ in the just-identified model

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

This generalises to any j.

Next, partition $\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_{.1} & \mathbf{Z}_{.2} & \mathbf{Z}_3 \end{bmatrix}$, $\pi^{[1]} = \begin{pmatrix} \pi_2^{[1]} & \pi_3^{[1]\prime} \end{pmatrix}'$ and consider the test for $H_0 : \pi_2^{[1]} = 0$ in

$$\mathbf{y} = \mathbf{d}\beta_1 + \mathbf{Z}_{.2}\pi_2^{[1]} + \mathbf{Z}_3\pi_3^{[1]} + \mathbf{u}_1.$$

The model under the null is then given by

$$\mathbf{y} = \mathbf{d}\beta_1 + \mathbf{Z}_3\pi_3^{[1]} + \mathbf{u}_1 \tag{1.A.9}$$

and the score test for $H_0 : \pi_2^{[1]} = 0$ is then the same as the Sargan test for overidentifying restrictions in (1.A.9) 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 $t_2^{[1]} \neq t_1^{[2]}$, whereas the Sargan test is symmetric, i.e. the score test for testing $H_0 : \pi_2^{[1]} = 0$ is identical to the score test for testing $H_0 : \pi_1^{[2]} = 0$ in the specification

$$\mathbf{y} = \mathbf{d}\beta_2 + \mathbf{Z}_{.1}\pi_1^{[2]} + \mathbf{Z}_3\pi_3^{[2]} + \mathbf{u}_2.$$

1.A.6 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 (1.11). The formulation of the threshold set $\widehat{\mathcal{V}}^{[j]}$ is given in Guo et al.

(2018, equation (7), p 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 (1.21).

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

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

1.A.7 Some Further Monte Carlo Results

Table 1.A.2 presents results for the same design as in Guo et al. (2018, Table 2), with $k_z = 7$, $k_{\mathcal{A}_0} = 4$, $\alpha = c_a\left(\iota_2', 0.5\iota_2', \mathbf{0}_3'\right)'$, $\rho_z = 0$, $c_\alpha = 0.2$, 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 1.2 due to the smaller number of instruments.

Table 1.A.2: 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.

1.A.8 Confidence Intervals for Application

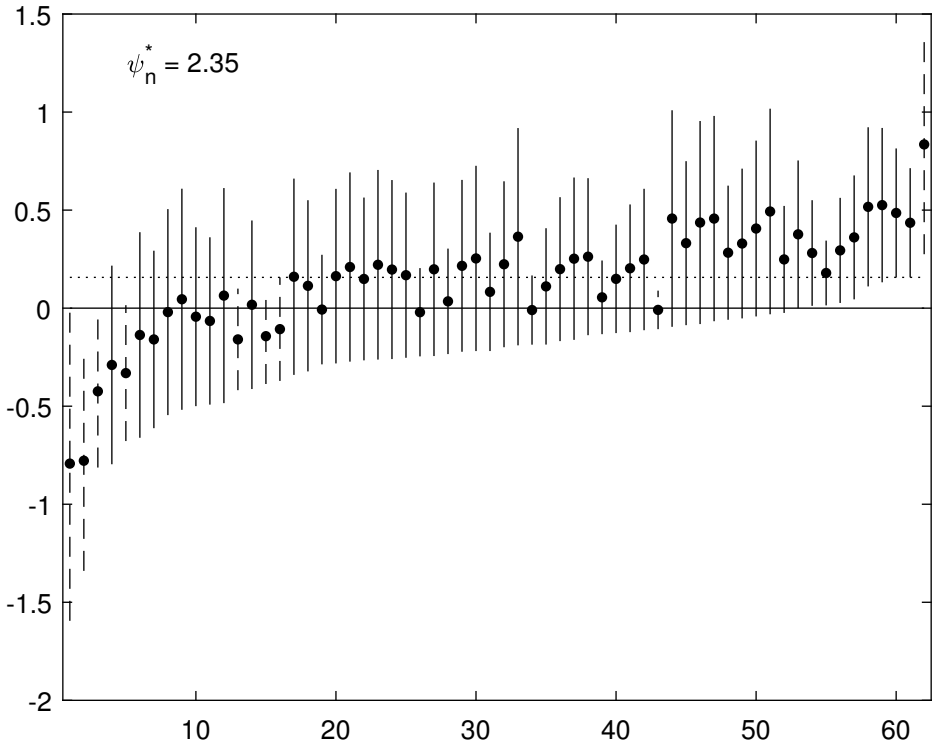


Figure 1.A.4: Confidence intervals for $\psi_n^* = 2.35$ for effect of $ln\,(BMI)$ on $ln\,(DPB)$, $\omega_n = 3.03$, $k_z = 62$. Largest group of instruments with overlapping confidence intervals with p-value for Sargan test statistic less than p_n indicated by intersections with dotted horizontal line. CIs for instruments selected as invalid represented by dashed lines, for those selected as valid by solid lines. Mid-points of CIs are point estimates $\widehat{\beta}_j$ represented by solid circles.

1.A.9 Summary Data

The CI method can be applied with multi-sample (e.g. GWAS) summary data, where only $\left\{\widehat{\Gamma}_j, se\left(\widehat{\Gamma}_j\right)\right\}_{j=1}^{k_z}$ and $\{\widehat{\gamma}_j, se\,(\widehat{\gamma}_j)\}_{j=1}^{k_z}$ are available from different, independent sources, under the assumption that the instruments are independent. The individual estimates are often obtained from bivariate regressions, and hence these are only valid when the instruments are independent. From the sets of individual estimates, we obtain the estimates $\widehat{\beta}_j = \widehat{\Gamma}_j/\widehat{\gamma}_j$ and $\widehat{v}_j = \sqrt{\widehat{Var}\left(\widehat{\beta}_j\right)}$, and so we can construct the confidence intervals $ci_j\,(\psi_n)$ as in (1.13). Starting from the model

where all instruments are assumed to be valid, the Sargan test is replaced by the minimum distance Q-statistic,

$$Q\left(\widehat{\beta}^{k_z}\right) = \sum_{j=1}^{k_z} \left(\frac{\widehat{\beta}_j - \widehat{\beta}^{k_z}}{\widehat{v}_j} \right)^2,$$

where $\widehat{\beta}^{k_z}$ is the minimum-distance/inverse-variance weighted (IVW) estimator given by

$$\widehat{\beta}^{k_z} = \frac{\sum_{j=1}^{k_z} w_j \widehat{\beta}_j}{\sum_{j=1}^{k_z} w_j}$$

where $w_j = \widehat{v}_j^{-2}$.

If all instruments are valid, then $Q\left(\widehat{\beta}^{k_z}\right) \xrightarrow{d} \chi^2_{k_z-1}$, and we can then follow the steps of the algorithm described in Appendix 1.A.4, where now an estimator based on s selected instruments as valid, denoted $r = 1, ..., s$, is given by

$$\widehat{\beta}^{s} = \frac{\sum_{r=1}^{s} w_r \widehat{\beta}_r}{\sum_{r=1}^{s} w_r},$$

with associated Q-statistic

$$Q\left(\widehat{\beta}^{s}\right) = \sum_{r=1}^{s} \left(\frac{\widehat{\beta}_r - \widehat{\beta}^{s}}{\widehat{v}_r} \right)^2.$$

So, instead of including invalid instruments as explanatory variables in the model, here invalid instruments are excluded from the analysis altogether, which gives equivalent results when instruments are independent.

Chapter 2

Adaptive Lasso Method for Selecting Valid Instrumental Variables with Two Endogenous Variables

Abstract

In a linear instrumental variable (IV) setup with two endogenous exposure variables, we investigate the adaptive Lasso as a method for selecting valid instrumental variables from a set of available instruments that may contain invalid ones. An instrument is invalid if it has a direct effect on the outcome or affects the outcome through unobserved factors. Following Windmeijer et al. (2018), we propose a median-in-medians estimator. This estimator is consistent for the causal effects under the condition that the number of invalid instruments is smaller than half of the total number of the candidate instruments minus one. We show that the adaptive Lasso, which uses the median-in-medians estimator for the penalty weights, can achieve oracle selection and estimation. This is evidenced by our Monte Carlo simulation results. We apply the method to estimate the causal effects of educational attainment and cognitive ability on body mass index (BMI).

This chapter is co-authored with Eleanor Sanderson and Frank Windmeijer. Xiaoran made major contributions to the paper, including theory and the computational work, and developed the research idea. Elanor contributed to the empirical application. Frank contributed substantially to the theory and developed the research idea.

2.1 Introduction

Instrumental variables (IV) are widely used to determine the causal effect of a treatment/exposure on an outcome when their relationship is potentially confounded by unobserved factors. In IV estimation, it is crucial that all the instruments are valid. This requires that (a) the instruments must be associated with the endogenous exposure variable (the relevance condition), and (b) the only pathway from the instruments to the outcome is through the exposure; the instruments must not have direct effects on the outcome nor affect the outcome through unobservables (the exclusion restriction). In our research, we are concerned with the situation where we have a large number of available instruments that satisfy the relevance condition. However, some of the instruments may violate the exclusion restriction and hence be invalid. If we include these invalid instruments in IV estimation, it will result in inconsistent estimation. One mitigation strategy is to select just the valid instruments and use these for IV estimation. This is the strategy we use in this work.

Previous work has tackled the IV selection problem in the case of a single endogenous variable. Kang et al. (2016) establish the model setup for IV selection, which has been adopted by most later work. They develop the identification conditions and propose a selection method based on the least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996). Windmeijer et al. (2019) propose a method based on the adaptive Lasso (Zou, 2006) under the assumption that more than half of the candidate instruments are valid; the so-called majority rule. Compared to Lasso, the adaptive Lasso method suggested by Windmeijer et al. (2019) theoretically guarantees consistent selection without assuming the restrictive Irrepresentable Condition. Guo et al. (2018) refine the identification condition proposed by Kang et al. (2016) and they establish the sufficient and necessary identification condition which is the plurality rule. It states that all the valid instruments form the largest group where instruments form a group if their instrument-specific estimators for the causal effect of interest converge to the same value. The Hard Thresholding with Voting method proposed by Guo et al. (2018) can achieve consistent selection under the plurality rule, which is a relaxation of the majority rule. Also assuming the plurality rule, Windmeijer et al. (2021)

propose the Confidence Interval method which theoretically guarantees consistent selection, and also has better finite sample performance compared to the Hard Thresholding method.

Unlike the existing literature above, we consider the case of two endogenous variables. This setting can be motivated by Mendelian Randomization (MR) which is used in epidemiology. In MR studies, genetic variants are used as instruments for estimating the causal effect of an endogenous exposure on a health-related outcome. In many cases, there may be additional endogenous variables that we want to control for apart from the primary (endogenous) exposure. For example, Sanderson et al. (2019) estimate the effect of educational attainment on body mass index (BMI) conditional on cognitive ability. Both educational attainment and cognitive ability are endogenous, and, thus, to use IV selection, the method must handle multiple endogenous variables.

We contribute to the literature by extending the adaptive Lasso method in Windmeijer et al. (2019) to allow for two endogenous variables. To this end, we propose a median-of-medians estimator which is $\sqrt{n}$-consistent under the so-called modified majority rule. This rule requires that the number of invalid instruments is less than half of the total number of the candidate instruments minus one. Using this median-of-medians estimator for the penalty weights, the adaptive Lasso can select the valid instruments consistently, and the resulting post-selection IV estimator has the same limiting distribution as if we knew the true set of valid instruments. These properties are jointly as the oracle property (Fan and Li, 2001).

To obtain the median-of-medians estimator, the inputs we need are all the just-identified estimators for the causal effects. To guarantee the identification of the casual parameters in each of the just-identified models, we need to impose the full rank assumption, which states that the 2×2 matrices formed by the first-stage coefficients of all pairs of instruments must have full rank. In general, this assumption requires that all instruments should be relevant for both endogenous variables. In practice, however, this might not be the case, as the candidate instruments may be identified separately for each endogenous variable. For example, genetic variants that are candidate instruments for educational attainment and cognitive ability are identified in separate GWAS studies. Therefore, an individual instrument might only be relevant for one of the endogenous exposures,

which violates the full rank assumption. To relax the assumption, we also propose an adjustment to the method that uses a block structure.

The rest of the paper is structured as follows. Section 2.2 introduces the model setup for the IV selection and estimation problem. Section 2.3 develops the adaptive Lasso IV selection method using the median-of-medians estimator as penalty weights. We also combine the adaptive Lasso with the downward testing procedure for model selection proposed by Andrews (1999). In Section 2.4, we introduce the block structure variation of the method that accounts for violation of the full rank assumption. Section 2.5 presents the Monte Carlo simulation results. In Section 2.6, we apply our method to Mendelian randomisation and estimate the causal effects of educational attainment and cognitive ability on BMI. Section 2.7 concludes.

Notation. In the remainder of the paper, let $\|.\|_q$ denote the l_q-norm of a vector. For a matrix $\mathbf{X}_{n\times p}$ with full column rank, let $\mathbf{P}_X = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{M}_X = \mathbf{I}_n - \mathbf{P}_X$ where $\mathbf{I}_n$ is the n-dimensional identity matrix.

2.2 Model Setup

As in Kang et al. (2016), we consider the following potential outcomes model which allows for potentially invalid instruments. For $i = 1, ..., n$, the observed outcome for an individual i is denoted by the scalar Y_i, the two endogenous exposure/treatment variables by $\mathbf{X}_i \in \mathbb{R}^2$, and the k_z candidate instrumental variables by $\mathbf{Z}_i \in \mathbb{R}^{k_z}$ ($k_z > 2$). Some of the candidate instruments may be invalid in the sense that they may have direct effects on the outcome or affect the outcome through unobserved factors. Let $Y_i^{(\mathbf{x},\mathbf{z})}$ be the outcome if individual i were to have exposure value $\mathbf{x}$ and instrument value $\mathbf{z}$. For a set of values $(\mathbf{x}^*, \mathbf{z}^*)$, the potential outcomes model is

$$Y_i^{(\mathbf{x}^*,\mathbf{z}^*)} = Y_i^{(0,0)} + \mathbf{x}^{*\prime}\boldsymbol{\theta} + \mathbf{z}^{*\prime}\boldsymbol{\phi}.$$

According to Holland (1988), it follows that the observed data model is

$$Y_i = Y_i^{(0,0)} + \mathbf{X}_i'\boldsymbol{\theta} + \mathbf{Z}_i'\boldsymbol{\phi}, \tag{2.1}$$

where $\{Y_i, \mathbf{X}_i', \mathbf{Z}_i'\}_{i=1}^n$ is a random sample, $\boldsymbol{\theta}$ is the parameter of interest, and $\boldsymbol{\phi}$ measures the direct effect of possibly invalid instruments on the outcome.

Apart from a direct effect, another possible violation of the exclusion restriction is if the invalid instruments affect the outcome through unobservables. To capture this indirect effect, we model the unobserved factor $Y_i^{(0,0)}$ in the following way:

$$E[Y_i^{(0,0)}|\mathbf{Z}_i] = \mathbf{Z}_i'\boldsymbol{\mu}, \tag{2.2}$$

where the indirect effect is measured by the parameter $\boldsymbol{\mu}$. If we combine (2.1) and (2.2), the observed date model is

$$Y_i = \mathbf{X}_i'\boldsymbol{\theta} + \mathbf{Z}_i'\boldsymbol{\alpha} + u_i, \tag{2.3}$$

where $\boldsymbol{\alpha} = \boldsymbol{\phi} + \boldsymbol{\mu}$. By definition,

$$u_i = Y_i^{(0,0)} - E[Y_i^{(0,0)}|\mathbf{Z}_i]$$

and hence $E[u_i|\mathbf{Z}_i] = 0$. The parameter $\boldsymbol{\alpha}$ captures the violation of the exclusion restriction; a valid instrument, which has zero direct effect as well as zero indirect effect, has an entry in $\boldsymbol{\alpha}$ that equals 0. Formally, following the definition of an invalid instrument proposed by Kang et al. (2016), for $j \in 1, ..., k_z$, an instrument Z_j is invalid if $\alpha_j \neq 0$ and valid if $\alpha_j = 0$. Let $\mathcal{V}$ and $\mathcal{A}$ be the sets of indices of the valid and invalid instruments respectively: $\mathcal{V} = \{j : \alpha_j = 0\}$, $\mathcal{A} = \{j : \alpha_j \neq 0\}$, with dimensions $k_\mathcal{V}$ and $k_\mathcal{A}$, then $k_z = k_\mathcal{V} + k_\mathcal{A}$. Let $\mathbf{Z}_\mathcal{V}$ and $\mathbf{Z}_\mathcal{A}$ be the sets of valid and invalid instruments. In this setup, we are interested in the identification and estimation of both $\mathcal{V}$ and $\boldsymbol{\theta}$ in large samples with fixed k_z.

Let $\mathbf{y}$ be the n-vector of n observations on Y_i, and let $\mathbf{X}$ and $\mathbf{Z}$ be the $n \times 2$ and $n \times k_z$ matrices of the endogenous treatment/exposure variables and candidate instrumental variables, respectively. Rewrite the outcome equation (2.3) in matrix form:

$$\mathbf{y} \;=\; \mathbf{X}\boldsymbol{\theta} + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{u}, \tag{2.4}$$

where $\mathbf{u}$ is the $n-$ vector of u_i. The first-stage linear projection of $\mathbf{X}$ on $\mathbf{Z}$ is

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\Pi} + \boldsymbol{v}_x, \tag{2.5}$$

where $\boldsymbol{\Pi} = E[\mathbf{Z}'\mathbf{Z}]^{-1}E[\mathbf{Z}'\mathbf{X}]$ and $E[\mathbf{Z}_i v'_{xi}] = \mathbf{0}$.

The ideal properties of the estimators for the set of valid instruments and the causal effects, denoted by $\widehat{\mathcal{V}}$ and $\widehat{\boldsymbol{\theta}}$, are summarized by the concept of the oracle property (Fan and Li, 2001). This property states that the probability of selecting the true set of valid instruments converges to 1 as sample size, n, goes to infinity, i.e. $P(\widehat{\mathcal{V}} = \mathcal{V}) \to 1$, and that the nonzero estimators have the same asymptotic distributions as they would have had if the set of valid instruments was known in advance. The oracle model is given by

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

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

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

where $\boldsymbol{\xi}$ is defined implicitly. The oracle estimator for $\boldsymbol{\theta}$ is given by

$$\widehat{\boldsymbol{\theta}}_{or} = \left(\widehat{\mathbf{X}}'\mathbf{M}_{Z_{\mathcal{A}}}\widehat{\mathbf{X}}\right)^{-1}\widehat{\mathbf{X}}'\mathbf{M}_{Z_{\mathcal{A}}}\mathbf{y}. \tag{2.6}$$

Under standard assumptions, the limiting distribution of $\widehat{\boldsymbol{\theta}}_{or}$ is

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_{or} - \boldsymbol{\theta}\right) \xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{\sigma}^2_{\theta_{or}}\right), \tag{2.7}$$

where

$$\boldsymbol{\sigma}^2_{\theta_{or}} = \sigma^2_u\left(E\left[\mathbf{Z}_i\mathbf{X}_i\right]' E\left[\mathbf{Z}_i\mathbf{Z}_i'\right]^{-1} E\left[\mathbf{Z}_i\mathbf{X}_i\right] - E\left[\mathbf{Z}_{\mathcal{A},i}\mathbf{X}_i\right]' E\left[\mathbf{Z}_{\mathcal{A},i}\mathbf{Z}'_{\mathcal{A},i}\right]^{-1} E\left[\mathbf{Z}_{\mathcal{A},i}\mathbf{X}_i\right]\right)^{-1}.$$

The derivation is similar to Appendix A.2 of Windmeijer et al. (2021).

2.3 The Adaptive Lasso Method for IV Selection and Estimation

2.3.1 The Adaptive Lasso with the Median-of-medians Estimator

Based on the adaptive Lasso method for IV selection in Windmeijer et al. (2019), we extend the method to allow for two endogenous variables. Let $\boldsymbol{\theta} = (\beta, \tau)'$, where β and τ are coefficients of the two endogenous treatment/exposure variables. Consider again the model

$$\mathbf{y} = \mathbf{X}_1\beta + \mathbf{X}_2\tau + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{u}. \tag{2.8}$$

Under model (2.5) and (2.8), we assume throughout that the following assumptions 2.1 to 2.5 hold.

Assumption 2.1. *Let $\boldsymbol{\Pi}_{jk}$ be the matrix containing the j-th and i-th row of $\boldsymbol{\Pi}$. rank$(\boldsymbol{\Pi}_{jk}) = 2$ for any $j, k = 1, ..., k_z, j \neq k$.*

This pair-wise rank condition is necessary for the identification of the casual parameters in each of the just-identified models. The resulting just-identified estimators are key inputs to the adaptive Lasso IV selection method, as explained below.

Assumption 2.2. *The number of invalid instruments satisfies: $k_{\mathcal{A}} < (k_z - 1)/2$*

This is a sufficient condition for identification. Under the majority rule $k_{\mathcal{A}} < k_z/2$, Windmeijer et al. (2019) propose the adaptive Lasso IV selection method allowing for a single endogenous variable. Similarly, we need the modified majority rule $k_{\mathcal{A}} < (k_z - 1)/2$, as stated in Assumption (2.2), to guarantee consistent IV selection and estimation.

Assumption 2.3. *$E[\mathbf{Z}_i\mathbf{Z}_i'] = \mathbf{Q}$, with $\mathbf{Q}$ a finite and full rank matrix.*

Assumption 2.4. *Let $\mathbf{w}_i = (u_i \ \boldsymbol{v}_{xi}')'$. Then $E[\mathbf{w}_i] = \mathbf{0}$;*

Conditional homoskedasticity: $E\left[\mathbf{w}_i\mathbf{w}_i'|\mathbf{Z}_i\right] = \begin{bmatrix} \sigma_u^2 & \sigma_{uv_{x1}} & \sigma_{uv_{x2}} \\ \sigma_{uv_{x1}} & \sigma_{v_{x1}}^2 & \sigma_{v_{x1}v_{x2}} \\ \sigma_{uv_{x2}} & \sigma_{v_{x1}v_{x2}} & \sigma_{v_{x2}}^2 \end{bmatrix} = \mathbf{\Sigma}$. *The elements of* $\mathbf{\Sigma}$ *are finite.*

For ease of exposition, we assume conditional homoskedasticity. We can easily relax this assumption to allow for heteroskedasticity and/or clustering by replacing the Sargan test by the Hansen J-test, and doing post-selection two-step GMM instead of 2SLS or by using the robust 2SLS standard errors and.

Assumption 2.5. $plim\left(n^{-1}\mathbf{Z}'\mathbf{Z}\right) = E\left[\mathbf{Z}_i\mathbf{Z}_i'\right] = \mathbf{Q}$; $plim\left(n^{-1}\mathbf{Z}'\mathbf{X}\right) = E\left[\mathbf{Z}_i\mathbf{X}_i'\right]$; $plim\left(n^{-1}\mathbf{Z}'\mathbf{u}\right) = E\left[\mathbf{Z}_i u_i\right] = 0$; $plim\left(n^{-1}\mathbf{Z}'\boldsymbol{v}_x\right) = E\left[\mathbf{Z}_i\boldsymbol{v}_{xi}'\right] = \mathbf{0}$.

Based on the definition of a valid instrument, IV selection is equivalent to identifying which entries in $\boldsymbol{\alpha}$ that are zero. For this purpose, we consider using the adaptive Lasso to estimate $\boldsymbol{\alpha}$, as the Lasso-type methods will shrink some entries in $\boldsymbol{\alpha}$ to exactly zero. Hence, we can obtain estimators for $\mathcal{V}$ and $\mathcal{A}$ from the adaptive Lasso estimator for $\boldsymbol{\alpha}$, which we denote by $\widehat{\boldsymbol{\alpha}}_{ad}$. The estimators for $\mathcal{V}$ and $\mathcal{A}$ are then $\widehat{\mathcal{V}} = \{j : \widehat{\alpha}_{ad,j} = 0\}$ and $\widehat{\mathcal{A}} = \{j : \widehat{\alpha}_{ad,j} \neq 0\}$. Following Windmeijer et al. (2019), the adaptive Lasso estimator for $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ is defined as

$$(\widehat{\boldsymbol{\alpha}}_{ad}, \widehat{\boldsymbol{\theta}}_{ad}) = \underset{\boldsymbol{\alpha},\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{2} \left\|\mathbf{P}_Z(\mathbf{y} - \mathbf{Z}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\theta})\right\|_2^2 + \lambda_n \sum_{j=1}^{k_z} \frac{\alpha_j}{\hat{\alpha}_{w,j}}, \tag{2.9}$$

where $\hat{\alpha}_{w,j}$, the penalty weight for α_j, is a pre-specified estimator for α_j, which we will discuss in detail below. The l_2 norm is $(\mathbf{y} - \mathbf{Z}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\theta})'\mathbf{P}_Z(\mathbf{y} - \mathbf{Z}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\theta})$. As $\boldsymbol{\theta}$ is not penalized, according to Kang et al. (2016) and Windmeijer et al. (2019), we can rewrite the adaptive Lasso in (2.9) as:

$$\widehat{\boldsymbol{\alpha}}_{ad} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \frac{1}{2} \left\|\mathbf{y} - \widetilde{\mathbf{Z}}\boldsymbol{\alpha}\right\|_2^2 + \lambda_n \sum_{j=1}^{k_z} \frac{\alpha_j}{\hat{\alpha}_{w,j}}, \tag{2.10}$$

where $\widetilde{\mathbf{Z}} = \mathbf{M}_{\widehat{\mathbf{X}}}\mathbf{Z}$, $\mathbf{M}_{\widehat{\mathbf{X}}} = \mathbf{I}_n - \widehat{\mathbf{X}}(\widehat{\mathbf{X}}'\widehat{\mathbf{X}})^{-1}\widehat{\mathbf{X}}'$. λ_n is the tuning parameter controlling the strength of the penalization. A larger λ_n leads to more entries in $\boldsymbol{\alpha}$ being shrunk to zero, which implies that the adaptive Lasso selects more instruments as valid. According to Theorem 1 and Remark 1 in Zou (2006), the adaptive Lasso

estimator for $\boldsymbol{\alpha}$, as defined in (2.10), can select the valid instruments consistently if:

- the penalty weight $\hat{\alpha}_{w,j}$ is a $\sqrt{n}$-consistent estimator for α_j,
- $\lambda_n \to \infty$, $\lambda_n = o(\sqrt{n})$.

The intuition of these conditions is as follows. If $\hat{\alpha}_{w,j}$ is a $\sqrt{n}$-consistent estimator for α_j, then $\hat{\alpha}_{w,j}$ will be close to zero when $\alpha_j = 0$. Since $\hat{\alpha}_{w,j}$ enters in the denominator in (2.10), a value close to zero will produce a large penalty weight, and, thus, make it more likely that α_j is shrunk to zero.

To implement the adaptive Lasso, we need to find a $\sqrt{n}$-consistent estimator for $\boldsymbol{\alpha}$ that we can use for the penalty weights. We obtain such an estimator in the following way. The reduced form specification for $\mathbf{y}$ is

$$\mathbf{y} \;=\; \mathbf{Z}\boldsymbol{\Gamma} + \boldsymbol{v}_y,$$

where

$$\boldsymbol{\Gamma} = \boldsymbol{\Pi}\boldsymbol{\theta} + \boldsymbol{\alpha}. \tag{2.11}$$

Using OLS, we can estimate $\boldsymbol{\Gamma}$ and $\boldsymbol{\Pi}$ by $\widehat{\boldsymbol{\Gamma}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$ and $\widehat{\boldsymbol{\Pi}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$, both of which are $\sqrt{n}$-consistent estimators. If we can find a $\sqrt{n}$-consistent estimator for $\boldsymbol{\theta}$, then we can obtain a $\sqrt{n}$-consistent estimator for $\boldsymbol{\alpha}$ from (2.11). Under the majority rule $k_{\mathcal{A}} < \frac{1}{2}k_z$, and with a single endogenous variable, Windmeijer et al. (2019) propose a $\sqrt{n}$-consistent median estimator for θ. In their case, where θ is a scalar, there would be k_z just-identified estimators for θ. Hence, if Z_j is valid, then the just-identified estimator $\widehat{\theta}_j$ using only Z_j as the valid instrument is consistent and, thus, it converges to θ. However, for invalid instruments, their just-identified estimators converge to values different from θ. When Windmeijer et al. (2019) impose the majority rule $k_{\mathcal{A}} < \frac{1}{2}k_z$, more than half of the just-identified estimators are consistent, and, therefore, their median must also be a consistent estimator for θ. In our case, where there are two endogenous variables, a natural extension of Windmeijer et al. (2019) would be to assume $\binom{k_{\mathcal{V}}}{2} > \frac{1}{2}\binom{k_z}{2}$ such that the median of the $\binom{k_z}{2}$ just-identified estimators is consistent for $\boldsymbol{\theta}$. However, instead of this straightforward extension, we propose a median-of-medians estimator

under the milder condition $k_{\mathcal{A}} < \frac{k_z-1}{2}$, which allows for more invalid instruments than $\binom{k_{\mathcal{V}}}{2} > \frac{1}{2}\binom{k_z}{2}$. As an illustration, suppose $k_z = 100$, then the maximum number of invalid instruments allowed for in the simple extension is 29, while for the median-of-medians estimator, this number can be 49. Similar to Windmeijer et al. (2019), the inputs for obtaining the median-of-medians estimator are all the just-identified estimators. Based on (2.11), for any pair of instruments $\{Z_j, Z_k\}$, $j, k = 1, ..., k_z, j \neq k$, we have

$$\begin{pmatrix} \Gamma_j \\ \Gamma_k \end{pmatrix} = \begin{pmatrix} \Pi_{j1} & \Pi_{j2} \\ \Pi_{k1} & \Pi_{k2} \end{pmatrix} \begin{pmatrix} \beta \\ \tau \end{pmatrix} + \begin{pmatrix} \alpha_j \\ \alpha_k \end{pmatrix},$$

which we write compactly as

$$\boldsymbol{\Gamma}_{jk} = \boldsymbol{\Pi}_{jk}\boldsymbol{\theta} + \boldsymbol{\alpha}_{jk},$$

where the vectors $\boldsymbol{\Gamma}_{jk}$ and $\boldsymbol{\alpha}_{jk}$ contain the j-th and k-th elements of $\boldsymbol{\Gamma}$ and $\boldsymbol{\alpha}$, respectively. Under Assumption 2.1, $\boldsymbol{\Pi}_{jk}^{-1}$ exists for any $j, k, j \neq k$. It follows that

$$\boldsymbol{\Pi}_{jk}^{-1}\boldsymbol{\Gamma}_{jk} = \boldsymbol{\theta} + \boldsymbol{\Pi}_{jk}^{-1}\boldsymbol{\alpha}_{jk}. \tag{2.12}$$

We let $\widehat{\boldsymbol{\theta}}_{jk}$ denote the just-identified estimator for $\boldsymbol{\theta}$ using Z_j and Z_k as the valid instruments. It is given by 2SLS on the following specification:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta}_{jk} + \mathbf{Z}_{-jk}\boldsymbol{\alpha}_{-jk} + \mathbf{u}_{jk}, \tag{2.13}$$

where $\mathbf{Z}_{-jk}$ denotes all the candidate instruments except for Z_j and Z_k. Following Proposition A1 in Windmeijer et al. (2019), it can be shown that the just-identified estimator generated from specification (2.13) using 2SLS is equivalent to

$$\widehat{\boldsymbol{\theta}}_{jk} = \widehat{\boldsymbol{\Pi}}_{jk}^{-1}\widehat{\boldsymbol{\Gamma}}_{jk}. \tag{2.14}$$

As we prove in Appendix 3.A.2 in Chapter 3, it follows that

$$plim\left(\widehat{\boldsymbol{\theta}}_{jk}\right) = \boldsymbol{\Pi}_{jk}^{-1}\boldsymbol{\Gamma}_{jk} = \boldsymbol{\theta} + \boldsymbol{\Pi}_{jk}^{-1}\boldsymbol{\alpha}_{jk}. \tag{2.15}$$

From (2.15), if Z_k and Z_j are any valid pair of instruments, i.e. it holds that $\boldsymbol{\alpha}_{jk} = \mathbf{0}$, then $\widehat{\boldsymbol{\theta}}_{jk}$ is a consistent estimator for $\boldsymbol{\theta}$. However, if at least one of Z_j and Z_k is invalid, i.e. $\boldsymbol{\alpha}_{jk} \neq \mathbf{0}$, then the resulting $\widehat{\boldsymbol{\theta}}_{jk}$ is inconsistent.

Using the just-identified estimators, we construct the $\sqrt{n}$-consistent median-of-medians estimator for $\boldsymbol{\theta}$ as follows. Focusing on β, which is the first element of $\boldsymbol{\theta}$, we have $k_z - 1$ just-identified estimators that use Z_j as one of the two instruments. We take the median of these $k_z - 1$ estimators, and denote this median by $\widehat{\beta}_j^m$. We repeat this procedure for all the candidate instruments to get k_z median estimators for β. We again take the median of these median estimators to get a $\sqrt{n}$-consistent median-of-medians estimator for β, which we denote by $\widehat{\beta}^{mm}$. We repeat the same procedure for τ, which is the second element of $\boldsymbol{\theta}$, to obtain a $\sqrt{n}$-consistent estimator $\widehat{\tau}^{mm}$ for τ. The median-of-medians estimator of $\boldsymbol{\theta}$ is then:

$$\widehat{\boldsymbol{\theta}}^{mm} = (\widehat{\beta}^{mm}, \widehat{\tau}^{mm})'. \tag{2.16}$$

Using an example, we illustrate the algorithm for obtaining the median-of-medians estimator. Suppose we have six candidate instruments with instruments Z_1 and Z_2 being invalid. Table 2.1 lists the just-identified estimators for β and they are estimated using each IV pair. We color all the invalid instruments and inconsistent estimators with red. For the general case, the just-identified estimator $\hat{\beta}_{jk}$ is a consistent estimator for β if and only if both Z_j and Z_k are valid. Hence, all the estimators in Column (1) and Column (2) are inconsistent as at least one

	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6
	(1)	(2)	(3)	(4)	(5)	(6)
Z_1		$\widehat{\beta}_{21}$	$\widehat{\beta}_{31}$	$\widehat{\beta}_{41}$	$\widehat{\beta}_{51}$	$\widehat{\beta}_{61}$
Z_2	$\widehat{\beta}_{12}$		$\widehat{\beta}_{32}$	$\widehat{\beta}_{42}$	$\widehat{\beta}_{52}$	$\widehat{\beta}_{62}$
Z_3	$\widehat{\beta}_{13}$	$\widehat{\beta}_{23}$		$\widehat{\beta}_{43}$	$\widehat{\beta}_{53}$	$\widehat{\beta}_{63}$
Z_4	$\widehat{\beta}_{14}$	$\widehat{\beta}_{24}$	$\widehat{\beta}_{34}$		$\widehat{\beta}_{54}$	$\widehat{\beta}_{64}$
Z_5	$\widehat{\beta}_{15}$	$\widehat{\beta}_{25}$	$\widehat{\beta}_{35}$	$\widehat{\beta}_{45}$		$\widehat{\beta}_{65}$
Z_6	$\widehat{\beta}_{16}$	$\widehat{\beta}_{26}$	$\widehat{\beta}_{36}$	$\widehat{\beta}_{46}$	$\widehat{\beta}_{56}$	
median	$\widehat{\beta}_1^m$	$\widehat{\beta}_2^m$	$\widehat{\beta}_3^m$	$\widehat{\beta}_4^m$	$\widehat{\beta}_5^m$	$\widehat{\beta}_6^m$

Table 2.1: Illustration of the Median-of-Medians Estimator

of the invalid instruments Z_1 and Z_2 is involved in the estimation. If we take the median of the estimators in each of Column (1) and Column (2), then the resulting median estimators $\widehat{\beta}_1^m$ and $\widehat{\beta}_2^m$ would also be inconsistent. In Column (3) to (6), more than half of the $k_z - 1$ estimators in each column are consistent as we assume $k_{\mathcal{A}} < \frac{k_z-1}{2}$. Hence, the median estimators for these columns, which we refer to as the "column median estimators", i.e. $\widehat{\beta}_3^m$ to $\widehat{\beta}_3^m$, are all consistent. Now, we take the median of all these column median estimators (as shown in the last row of Table 2.1), i.e. $\widehat{\beta}^{mm} = median(\widehat{\beta}_1^m, ..., \widehat{\beta}_6^m)$. The assumption $k_{\mathcal{A}} < \frac{k_z-1}{2}$ implies $k_{\mathcal{A}} < \frac{k_z}{2}$. Thus, more than half of the column median estimators $\widehat{\beta}_1^m, ..., \widehat{\beta}_6^m$ are consistent. Therefore, the median of these column median estimators $\widehat{\beta}^{mm}$ is also consistent. In this way, under the assumption $k_{\mathcal{A}} < \frac{k_z-1}{2}$, even if we have no knowledge about which of the instruments are valid, the median-of-medians estimator is always consistent. We repeat the procedure for the second entry in $\widehat{\boldsymbol{\theta}}_{jk}$, to get a consistent estimator $\hat{\tau}_{mm}$ for τ. In this way, we get a consistent estimator $\widehat{\boldsymbol{\theta}}^{mm} = (\widehat{\beta}^{mm}, \hat{\tau}^{mm})'$ for $\boldsymbol{\theta}$.

The following theorem establishes the $\sqrt{n}$-consistency of the median-of-medians estimator defined in (2.16). See Appendix 2.A for the proof.

Theorem 2.1. *Under model specification (2.5) and (2.8) and Assumption 2.1 - 2.5, for $j = 1, ..., k_z$, let $\widehat{\boldsymbol{\beta}}_j$ be the $(k_z - 1)$-vector with each element defined as the first element of $\widehat{\boldsymbol{\theta}}_{jk} = \widehat{\boldsymbol{\Pi}}_{jk}^{-1}\widehat{\boldsymbol{\Gamma}}_{jk}$, $k = 1, ...k_z, k \neq j$. Define the median estimator $\widehat{\beta}_j^m$ as*

$$\widehat{\beta}_j^m = median(\widehat{\boldsymbol{\beta}}_j)$$

and the median-of-medians estimator $\widehat{\beta}^{mm}$ as

$$\widehat{\beta}^{mm} = median(\widehat{\beta}_1^m, ..., \widehat{\beta}_{k_z}^m).$$

If $k_{\mathcal{A}} < (k_z - 1)/2$, then $\widehat{\beta}^{mm}$ is a consistent estimator for β

$$plim(\widehat{\beta}^{mm}) = \beta.$$

Let $\boldsymbol{\delta}_{jk} = \boldsymbol{\Pi}_{jk}^{-1}\boldsymbol{\alpha}_{jk}$ and $\boldsymbol{\delta}_j$ be the $(k_z - 1)$-vector with each element defined as the

first element of $\boldsymbol{\delta}_{jk}$. *Let* $\widehat{\delta}_j^m = median(\boldsymbol{\delta}_j)$. *The limiting distribution of* $\widehat{\beta}^{mm}$ *is*

$$\sqrt{n}\left(\widehat{\beta}^{mm} - \beta\right) \xrightarrow{d} Q_{[L],k_z,k_{\mathcal{A}}},$$

where $Q_{[L],k_z,k_{\mathcal{A}}}$ *is the* L*-th order statistic of the limiting distribution of* $\widehat{\beta}_j^m$ *for* $j \in \mathcal{V}$. L *is determined by* k_z, $k_{\mathcal{A}}$ *and the signs of* $\widehat{\delta}_j^m$ *for* $j \in \mathcal{A}$. *For* $k_{\mathcal{V}}$ *even, the* L*-th order statistic is defined as either the average of the* L*-th and* $L+1$*-th order statistics or the average of the* $L-1$*-th and* L*-th order statistics.*
For $j \in \mathcal{V}$, *the limiting distribution of* $\widehat{\beta}_j^m$ *is*

$$\sqrt{n}\left(\widehat{\beta}_j^m - \beta\right) \xrightarrow{d} q_{j,[l],k_z,k_{\mathcal{A}}},$$

where $q_{j,[l],k_z,k_{\mathcal{A}}}$ *is the* l*-th order statistics of the limiting normal distribution of* $\sqrt{n}\left(\widehat{\beta}_{jk} - \beta\right)$ *with* $k \in \mathcal{V}$. l *is determined by* k_z, $k_{\mathcal{A}}$ *and the signs of* $\widehat{\delta}_{jk}$ *for* $j \in \mathcal{A}$. *The results stated above also hold for* τ, *the second entry of* $\boldsymbol{\theta}$. *Therefore, the median-of-medians estimator defined in (2.16) is a* $\sqrt{n}$*-consistent estimator for* $\boldsymbol{\theta}$.

Using $\widehat{\boldsymbol{\theta}}^{mm}$, we obtain a $\sqrt{n}$-consistent estimator for $\boldsymbol{\alpha}$ by

$$\widehat{\boldsymbol{\alpha}}^{mm} = \widehat{\boldsymbol{\Gamma}} - \widehat{\boldsymbol{\Pi}}\widehat{\boldsymbol{\theta}}^{mm}. \tag{2.17}$$

Following Theorem 1 and Remark 1 in Zou (2006), if we plug the estimator in (2.17) into the adaptive Lasso defined in (2.10), then the resulting estimated set of valid instruments $\widehat{\mathcal{V}}_n = \{j : \widehat{\alpha}_{ad,j} = 0\}$ satisfies $P(\widehat{\mathcal{V}}_n = \mathcal{V}) \to 1$ with $\lambda_n \to \infty$, $\lambda_n = o(\sqrt{n})$. Hence, we select valid instruments consistently with the adaptive Lasso.

Similar to Kang et al. (2016) and Windmeijer et al. (2019), we can obtain the adaptive Lasso estimator for $\boldsymbol{\theta}$ as follows:

$$\hat{\boldsymbol{\theta}}_{ad} = \begin{bmatrix} \widehat{\mathbf{X}}_1'\widehat{\mathbf{X}}_1 & \widehat{\mathbf{X}}_1'\widehat{\mathbf{X}}_2 \\ \widehat{\mathbf{X}}_1'\widehat{\mathbf{X}}_2 & \widehat{\mathbf{X}}_2'\widehat{\mathbf{X}}_2 \end{bmatrix}^{-1} \begin{bmatrix} \widehat{\mathbf{X}}_1'(\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\alpha}}_{ad}) \\ \widehat{\mathbf{X}}_2'(\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\alpha}}_{ad}) \end{bmatrix}, \tag{2.18}$$

and it has the oracle distribution as in (2.7). As an alternative to obtaining the causal estimator directly from the adaptive Lasso as in (2.18), we can also estimate $\boldsymbol{\theta}$ by post-selection 2SLS using the estimated set of valid $\widehat{\mathcal{V}}_n$ in the following

specification:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{Z}_{\widehat{\mathcal{A}}_n}\boldsymbol{\alpha}_{\widehat{\mathcal{A}}_n} + \mathbf{u}, \tag{2.19}$$

where $\widehat{\mathcal{A}_n} = \{j : \widehat{\alpha}_{ad,j} \neq 0\}$. Note that we include the selected invalid instruments in $\widehat{\mathcal{A}_n}$ as additional explanatory variables in the above model. Following Theorem 2 in Guo et al. (2018), the next theorem states the oracle property of the post-selection 2SLS estimator obtained from (2.19) using $\mathbf{Z}_{\widehat{\mathcal{V}}_n}$ as the valid instruments.

Theorem 2.2. *Let $\widehat{\boldsymbol{\theta}}_n$ be the post-selection 2SLS estimator obtained from (2.19), which is given by*

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

Under the conditions of Theorem 2.1 and $\lambda_n \to \infty$, $\lambda_n = o(\sqrt{n})$, it follows that

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

The proof of Theorem 2 follows directly from Theorem 2 in Guo et al. (2018) under $P(\widehat{\mathcal{V}}_n = \mathcal{V}) \to 1$.

Inference for $\widehat{\boldsymbol{\theta}}_n$ can be implemented as an analogue of that for the oracle estimator defined in (2.6), as they have the same limiting distribution. From (2.7), the variance estimator is

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

where $\widehat{\sigma}^2_u = \widehat{\mathbf{u}}'\widehat{\mathbf{u}}/n$ and $\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\theta}}_n - \mathbf{Z}_{\widehat{\mathcal{A}}_n}\widehat{\boldsymbol{\alpha}}_{\widehat{\mathcal{A}}_n}$.

2.3.2 The Downward Testing Procedure

Consistent IV selection using the adaptive Lasso depends on the choice of the tuning parameter λ_n which controls the strength of penalization. While it is clear in theory that λ_n needs to satisfy $\lambda_n \to \infty$, $\lambda_n = o(\sqrt{n})$, it can be challenging to pick a specific value of λ_n for a given sample. A common practice of choosing the tuning parameter is k-fold cross-validation. However, it is well known that cross-validation works better for prediction rather than model selection (Bühlmann and Van De Geer, 2011), and cross-validation almost always results in inconsistent

variable selection, as stated in Chand (2012).

As an alternative, and similar to Windmeijer et al. (2019), we combine the adaptive Lasso with the downward testing procedure for moment selection as proposed by Andrews (1999) which uses the Sargan test statistic as the selection criterion. The downward testing procedure starts with a model treating all k_z instruments as valid. If the Sargan test rejects the model, then the procedure moves to models with $k_z - 1$ valid instruments and tests all such models. If the Sargan test rejects them all, then it moves to evaluate all models with $k_z - 2$ valid instruments, and so on, until it finds the model with $k_{\mathcal{V}}$ valid instruments that is not rejected by the Sargan test. This procedure quickly becomes computationally infeasible, since, for each number of instruments $(k_z, k_z - 1, ..., k_{\mathcal{V}})$, we need to exhaustively test models corresponding to all possible combinations of instruments.

The adaptive Lasso can mitigate the computational challenges in the downward testing procedure. When Lasso is implemented using the Least-Angle Regression (LARS) algorithm (Efron et al., 2004), it generates a selection path starting with a model with k_z valid instruments, and, for each LARS step, the number of valid instruments decreases by one. This means that, for each number of valid instruments $(k_z, k_z - 1, ..., k_{\mathcal{V}})$, we only need to evaluate a single model, i.e. the one on the LARS selection path, instead of all combinations of instruments. Suppose that, for a certain number of valid instruments k_n, the model on the LARS selection path has the set of valid instruments $\widehat{\mathcal{V}}_{k_n}$ and the set of invalid instruments $\widehat{\mathcal{A}}_{k_n}$. Now, the model on the selection path is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{Z}_{\widehat{\mathcal{A}}_{k_n}} \boldsymbol{\alpha}_{\widehat{\mathcal{A}}_{k_n}} + \mathbf{u}_{\widehat{\mathcal{A}}_{k_n}}. \tag{2.20}$$

The Sargan statistic is given by

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

where $\widehat{\boldsymbol{\theta}}_{k_n}$ is the 2SLS estimator using $Z_{\widehat{\mathcal{V}}_{k_n}}$ as instruments in model (2.20), and $\mathbf{u}(\widehat{\boldsymbol{\theta}}_{k_n})$ is the residual. We compare $\widehat{\mathcal{S}}\left(\widehat{\boldsymbol{\theta}}_{k_n}\right)$ with a critical value for the Sargan test, which we denote by ζ_{n,k_n-2}. We select, as the valid set of instruments, the $\widehat{\mathcal{V}}_{k_n}$ with the largest k_n that satisfies $\widehat{\mathcal{S}}\left(\widehat{\boldsymbol{\theta}}_{k_n}\right) < \zeta_{n,k_n-2}$. If there are multiple such

models, we select the one with the smallest Sargan statistic.

According to Andrews (1999), if the critical value ζ_{n,k_n-2} from the $\chi^2_{k_n-2}$ distribution satisfies

$$\zeta_{n,k_n-2} \to \infty \text{ for } n \to \infty, \text{ and } \zeta_{n,k_n-2} = o(n), \tag{2.21}$$

then the consistent selection and oracle post-selection estimation results summarized in Theorem 2.1 and 2.2 hold. In practice, following Windmeijer et al. (2019) and Windmeijer et al. (2021), instead of a critical value ζ_{n,k_n-2} for the Sargan test, we use a p-value p_n. If p_n satisfies $p_n \to 0$ and $log(p_n) = o(n)$, then condition (2.21) is satisfied (Windmeijer et al., 2019). As in Windmeijer et al. (2019) and Windmeijer et al. (2021), for a given sample, we set $p_n = 0.1/log(n)$, as suggested by Belloni et al. (2012). After IV selection, we obtain the causal estimator using post-selection 2SLS in the same way as in (2.19). Here Theorem 2.2 holds for the post-selection 2SLS estimator as IV selection is consistent.

2.4 The Block Structure for Obtaining the Median-of-medians Estimator

In the previous sections, we maintained Assumption 2.1. However, in practical applications, this assumption is likely to be violated because the candidate instruments might be identified for each endogenous variable separately, such that a given instrument may only be relevant for one of them. In this case, the pairs of instruments that are relevant for a given exposure variable (and only this variable) would violate the pair-wise full rank assumption. To mitigate this, we propose a variation of our method which uses a block structure to obtain the $\sqrt{n}$-consistent median-of-medians estimator for the adaptive Lasso. The block structure relaxes Assumption 2.1 as it only requires each instrument to be relevant for at least one endogenous variable:

$$\mathbf{\Pi}_j \neq \mathbf{0} \text{ for } j = 1, ..., k_z.$$

Under the block structure, the majority assumption 2.2 also needs to be adjusted. Suppose that the number of instruments selected for each endogenous

variable is, respectively, k_{z_1} and k_{z_2}, and that the corresponding number of invalid instruments is, respectively, $k_{\mathcal{A}_1}$ and $k_{\mathcal{A}_2}$. Note that the sets of instruments for the two endogenous variables can overlap. In general, the block structure requires that the majority assumption 2.2 holds for each set of instruments:

$$k_{\mathcal{A}_1} < (k_{z_1} - 1)/2 \text{ and } k_{\mathcal{A}_2} < (k_{z_2} - 1)/2. \tag{2.22}$$

We now give two examples to illustrate how we obtain the median-of-medians estimator under the block-structure. In the first example, suppose we have 8 candidate instruments, $Z_1, ..., Z_8$. Instruments $Z_1, ..., Z_5$ are only relevant for the first endogenous variable, while the remaining instruments $Z_6, ..., Z_8$ are only relevant for the second one. Hence, the two sets of instruments do not overlap. In this case, the majority assumption can be relaxed to

$$k_{\mathcal{A}_1} < k_{z_1}/2 \text{ and } k_{\mathcal{A}_2} < k_{z_2}/2. \tag{2.23}$$

For the sake of example, let instruments Z_1, Z_2 and Z_6 be invalid, such that our setup satisfies (2.23). Here we obtain just-identified estimators only from pairs that contain instruments from two different sets, see Table 2.2. As in Table 2.1, invalid instruments and inconsistent estimators are marked as red. Condition (2.23) guarantees that the medians of the estimators in Column (3) to (5) are consistent, and that the median of all the column medians $(med_1(\hat{\beta}), ..., med_5(\hat{\beta}))$ are consistent as well. In this way, we obtain a consistent median-of-medians estimator, i.e. $median_\beta = median(med_1(\hat{\beta}), ..., med_5(\hat{\beta})))$.

	Z_1	Z_2	Z_3	Z_4	Z_5
	(1)	(2)	(3)	(4)	(5)
Z_6	$\widehat{\beta}_{16}$	$\widehat{\beta}_{26}$	$\widehat{\beta}_{36}$	$\widehat{\beta}_{46}$	$\widehat{\beta}_{56}$
Z_7	$\widehat{\beta}_{17}$	$\widehat{\beta}_{27}$	$\widehat{\beta}_{37}$	$\widehat{\beta}_{47}$	$\widehat{\beta}_{57}$
Z_8	$\widehat{\beta}_{18}$	$\widehat{\beta}_{28}$	$\widehat{\beta}_{38}$	$\widehat{\beta}_{48}$	$\widehat{\beta}_{58}$
$median_\beta$	$med_1(\hat{\beta})$	$med_2(\hat{\beta})$	$med_3(\hat{\beta})$	$med_4(\hat{\beta})$	$med_5(\hat{\beta})$

Table 2.2: Illustration of the Block Structure with no overlapping instruments.

In the second example, we consider the case where two sets of instruments are

still identified for each endogenous variable separately, but they are overlapping with each other, i.e. some of the instruments are relevant for both variables. Suppose we have 7 instruments where instruments $Z_1, ..., Z_5$ are relevant for the first endogenous variable, $Z_3, ..., Z_7$ are relevant for the second endogenous variable, and, thus, $Z_3, ..., Z_5$ are relevant for both variables. In this setup, condition (2.22) holds. Again, the just-identified estimators are estimated only with the pairs of instruments that are from two different sets, unless at least one of them is relevant for both, see Table 2.3. As condition (2.22) is satisfied, the median-of-medians estimator is consistent.

	Z_1	Z_2	Z_3	Z_4	Z_5
Z_3	$\widehat{\beta}_{13}$	$\widehat{\beta}_{23}$	-	$\widehat{\beta}_{34}$	$\widehat{\beta}_{35}$
Z_4	$\widehat{\beta}_{14}$	$\widehat{\beta}_{24}$	$\widehat{\beta}_{34}$	-	$\widehat{\beta}_{45}$
Z_5	$\widehat{\beta}_{15}$	$\widehat{\beta}_{25}$	$\widehat{\beta}_{35}$	$\widehat{\beta}_{45}$	-
Z_6	$\widehat{\beta}_{16}$	$\widehat{\beta}_{26}$	$\widehat{\beta}_{36}$	$\widehat{\beta}_{46}$	$\widehat{\beta}_{56}$
Z_7	$\widehat{\beta}_{17}$	$\widehat{\beta}_{27}$	$\widehat{\beta}_{37}$	$\widehat{\beta}_{47}$	$\widehat{\beta}_{57}$
$\boldsymbol{median}_\beta$	$med_1(\hat{\beta})$	$med_2(\hat{\beta})$	$med_3(\hat{\beta})$	$med_4(\hat{\beta})$	$med_5(\hat{\beta})$

Table 2.3: Illustration of the Block Structure with overlapping instruments.

2.5 Monte Carlo Simulations

We conduct Monte Carlo simulations to evaluate the performance of our method in three settings. In the first setting, all instruments are relevant for both of the endogenous variables, while in the other two settings, some of the instruments are relevant for only one of the variables. We run the simulations for $1,000$ replications, and we implement the adaptive Lasso using the `Lars` package (Hastie and Efron, 2013) in `R`. We generate the simulation data from

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{u}$$
$$\mathbf{X}_1 = \mathbf{Z}\boldsymbol{\Pi}_{x_1} + \boldsymbol{\epsilon}_1$$
$$\mathbf{X}_2 = \mathbf{Z}\boldsymbol{\Pi}_{x_2} + \boldsymbol{\epsilon}_2$$

where

$$\begin{pmatrix} u_i \\ \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & 0 \\ \rho_2 & 0 & 1 \end{pmatrix} \right);$$

$$\mathbf{Z}_i \sim N\left(\mathbf{0}, \mathbf{\Sigma}_z\right);$$

with $\boldsymbol{\theta} = c_\theta(0.5, 1)'$, $c_\theta = 0.6$; $k_z = 21$; $\rho_1 = 0.25$, $\rho_2 = 0.3$; $k_{\mathcal{A}} = 9$, $\boldsymbol{\alpha} = c_\alpha \left(\boldsymbol{\iota}_9', \mathbf{0}_{12}'\right)'$, $c_\alpha = 0.4$. We follow the simulation setup in Windmeijer et al. (2021). We generate the elements of $\mathbf{\Pi}_{x_1}$ and $\mathbf{\Pi}_{x_2}$ from a uniform distribution on the interval $[1.5, 2.5]$, and we set the elements of $\mathbf{\Sigma}_z$ to be $\mathbf{\Sigma}_{\boldsymbol{z,jk}} = 0.5^{|j-k|}$. In this setup, all the instruments are relevant for both endogenous variables, and both the pair-wise full rank assumption 2.1 and the majority assumption 2.2 are satisfied.

First, in Table 2.4, we present the IV selection and estimation results of the adaptive Lasso method with penalty parameters chosen by 10-fold cross-validation. The first two columns of Table 2.4 report statistics related to estimation, and in both of these columns, we average the statistics over the two entries in $\boldsymbol{\theta}$. Column 1 presents the averaged median absolute error (MAE), while Column 2 shows the averaged standard deviation (SD). The remaining three columns in Table 2.4 report statistics related to IV selection. Column 3 is the number of instruments selected as invalid, Column 4 is the frequency with which all invalid instruments have been selected as invalid, and Column 5 is the frequency with which the oracle model has been selected. The three panels in Table 2.4 correspond to the sample sizes $N = 500, 1000, 2000$. In each panel, the first row, denoted "Oracle 2SLS", shows the results for the oracle 2SLS estimator, which is the 2SLS estimator that uses the true set of valid instruments, while it controls for the remaining invalid ones. The second row, denoted "Naive 2SLS", reports the results for the naive 2SLS estimates, which is the 2SLS that considers all candidate instruments to be valid. The third row, denoted "$\widehat{\boldsymbol{\theta}}^{mm}$", shows the results for the median-in-medians estimator $\widehat{\boldsymbol{\theta}}^{mm}$, as defined in (2.16). The fourth and fifth rows, denoted "Post-ALasso", report the results for the post-selection 2SLS estimators, which are the 2SLS estimators that use the instruments selected as valid, and include the invalid instruments as control variables. The last two rows, denoted "ALasso", show the results for our adaptive Lasso estimators, as defined in (2.18). We present results

for the adaptive Lasso and Post-Selection 2SLS estimators using two different types of cross-validation. First, as denoted with the "CV" subscript, we show cross-validation using the tuning parameter that gives the minimum cross-validation Sargan statistics. Second, as denoted with the "$CVSE$" subscript, we show cross-validation using the tuning parameter chosen by the one-standard-error rule.

In terms of IV selection, in all three sample sizes, the CV-procedure dominates the CVSE-procedure, especially for the smallest sample with $N = 500$. Both methods improve as the sample size increases. The frequencies of selecting the oracle model are both almost equal to 1 at $N = 2,000$ with 0.992 for CV and 0.956 for CVSE. In line with the selection performance, the post-selection 2SLS estimates are close to the oracle model at $N = 2,000$. In all three sample sizes, the post-selection 2SLS estimates outperform the adaptive Lasso estimates.

	MAE (1)	SD (2)	# invalid (3)	p allinv (4)	p oracle (5)
Panel (a), $N = 500$					
Oracle 2SLS	0.0624	0.0912	9	1	1
Naive 2SLS	0.2856	0.2685	0	0	0
$\widehat{\boldsymbol{\theta}}^{mm}$	0.1263	0.1517			
Post-ALasso$_{CV}$	0.1295	0.3060			
Post-ALasso$_{CVSE}$	0.2912	0.4368			
ALasso$_{CV}$	0.3460	0.4212	8.095	0.440	0.440
ALasso$_{CVSE}$	0.4393	0.4542	6.908	0.115	0.115
Panel (b), $N = 1,000$					
Oracle 2SLS	0.0439	0.0681	9	1	1
Naive 2SLS	0.2889	0.2037	0	0	0
$\widehat{\boldsymbol{\theta}}^{mm}$	0.0892	0.1268			
Post-ALasso$_{CV}$	0.0513	0.1627			
Post-ALasso$_{CVSE}$	0.0716	0.2332			
ALasso$_{CV}$	0.2047	0.2843	8.857	0.882	0.882
ALasso$_{CVSE}$	0.2895	0.3446	8.509	0.634	0.634
Panel (c), $N = 2,000$					
Oracle 2SLS	0.0305	0.0473	9	1	1
Naive 2SLS	0.2796	0.1448	0	0	0
$\widehat{\boldsymbol{\theta}}^{mm}$	0.0618	0.0941			
Post-ALasso$_{CV}$	0.0307	0.0574			
Post-ALasso$_{CVSE}$	0.0319	0.0881			
ALasso$_{CV}$	0.1341	0.1795	8.991	0.992	0.992
ALasso$_{CVSE}$	0.1753	0.2244	8.949	0.956	0.956

This table reports IV selection and estimation results of the adaptive Lasso method with 10-fold cross validation. The reported statistics include median absolute error (column 1), standard deviation (column 2), number of IVs selected as invalid (column 3), frequency with which all invalid IVs have been selected as invalid (column 4), and frequency with which oracle model has been selected (column 5). The simulations are based on 1,000 repetitions.

Table 2.4: Adaptive Lasso with 10-fold Cross Validation

Using the same simulation setup, we also conduct simulations where, instead of cross-validation, we rely on the downward testing procedure described in Section 2.3.2. Table 2.5 reports the simulation results for the adaptive Lasso with

downward testing, and each row shows one of the sample sizes $N = 500, 1000, 2000$. We see that the adaptive Lasso with the downward testing procedure outperforms the methods based on cross-validation. In particular, we find that the adaptive Lasso reaches a high frequency, 0.947, of selecting the oracle model even at the smallest sample size $N = 500$. This is much better than the corresponding frequencies for the CV (0.440) and CVSE (0.115) that we reported in Table 2.4.

	MAE (1)	SD (2)	# invalid (3)	p allinv (4)	p oracle (5)
$N = 500$	0.0853	0.1446	9.09	0.987	0.947
$N = 1000$	0.0544	0.0690	9.02	1	0.983
$N = 2000$	0.0375	0.0478	9.018	1	0.984

This table reports IV selection and estimation results of the adaptive Lasso method with downward testing stopping rule. The reported statistics include median absolute error (column 1), standard deviation (column 2), number of IVs selected as invalid (column 3), frequency with which all invalid IVs have been selected as invalid (column 4), and frequency with which oracle model has been selected (column 5). The simulations are based on $1,000$ repetitions.

Table 2.5: Adaptive Lasso with Downward Testing

Next, we consider two simulation settings where some of the instruments are relevant only for one of the endogenous variables, and, thus, the pair-wise full rank assumption 2.1 may not hold. We conduct simulations both with and without the block structure for these settings.

In simulation design (1), which we call "partial overlap", we consider the case where some of the instruments are relevant for both endogenous variables. We set $\mathbf{\Pi}_{x_1} = (\boldsymbol{\gamma}_{x_1}, \mathbf{0}_8')'$ and $\mathbf{\Pi}_{x_2} = (\mathbf{0}_{10}', \boldsymbol{\gamma}_{x_2})'$ where $\boldsymbol{\gamma}_{x_1}$ is a vector with 13 elements and $\boldsymbol{\gamma}_{x_2}$ a vector with 11 elements that are all generated from a uniform distribution on the interval $[1.5, 2.5]$. In this case, three instruments are relevant for both endogenous variables, while the remaining instruments are only relevant for one. We set $\boldsymbol{\alpha} = (\mathbf{0}_6', \boldsymbol{\iota}_5', \mathbf{0}_2', \boldsymbol{\iota}_3', \mathbf{0}_5')'$, such that 5 out of 13 relevant instruments for $\mathbf{X}_1$ are invalid, and 4 out of 11 for $\mathbf{X}_2$ are invalid. Note that the adjusted majority assumption (2.22) is satisfied in this setup.

In simulation design (2), which we call "no overlap", we consider the case where

the two sets of instruments are completely separate, such that no instrument is relevant for both endogenous variables. We set $\mathbf{\Pi}_{x_1} = (\boldsymbol{\gamma}_{x_1}, \mathbf{0}'_{11})'$ and $\mathbf{\Pi}_{x_2} = (\mathbf{0}'_{10}, \boldsymbol{\gamma}_{x_2})'$, where $\boldsymbol{\gamma}_{x_1}$ has length 10 and $\boldsymbol{\gamma}_{x_2}$ has length 11. We let $\boldsymbol{\alpha} = (\boldsymbol{\iota}'_4, \mathbf{0}'_6, \boldsymbol{\iota}'_5, \mathbf{0}'_6)'$ such that 4 out of 10 relevant instruments for $\mathbf{X}_1$ are invalid, and 5 out of 11 for $\mathbf{X}_2$ are invalid. All the other parameters are identical to simulation design (1). Again, note that the adjusted majority assumption 2.23 is satisfied.

Table 2.6 reports the results for simulation design (1), where some instruments are relevant for both endogenous variables (i.e., partial overlap). Table 2.7 shows the results for simulation design (2), where none of the instruments are relevant for both endogenous variables (i.e. no overlap). In all cases, we conduct IV selection and estimation using adaptive Lasso with the downward testing procedure, both with and without the block structure to obtain the median-of-medians estimators, as described in Section 2.4. Both tables report the same statistics as earlier, and the panels correspond to different choices of sample size, $N = 500, 1000, 2000$. In each panel, the first row, denoted "Oracle 2SLS", shows the results for the oracle 2SLS estimator, which is the 2SLS estimator that uses the true set of valid instruments, while it controls for the remaining invalid ones. The second row, denoted "Naive 2SLS", reports the results for the naive 2SLS estimates, which is the 2SLS that considers all candidate instruments to be valid. The third row, denoted "$\widehat{\boldsymbol{\theta}}^{mm}$", shows the results for the median-in-medians estimator obtained without the block structure. The fourth row, denoted "Post-ALasso", reports IV selection and estimation results using the adaptive Lasso downward testing procedure without the block structure. The fifth row, denoted "$\widehat{\boldsymbol{\theta}}^{mm}_{block}$", shows the median-in-medians estimator obtained with the block structure. The sixth row, denoted "Post-ALasso$_{block}$", presents IV selection and estimation results using the adaptive Lasso downward testing procedure with the block structure.

For all three sample sizes, results with the block structures dominates the ones without. This can be seen from the fact that $\widehat{\boldsymbol{\theta}}^{mm}_{block}$ has smaller MAE than $\widehat{\boldsymbol{\theta}}^{mm}$. Also the frequencies of selecting the oracle model with the block structure are also larger than these without the block structure. The edge of the block structure is even more significant in the "no overlap" setup. Here the selection performance without the block structure does not improve with the sample size anymore, and the frequencies of selecting the oracle model remains around 0.5. But for the block

structure, the oracle selection frequency is very close to 1 even at $N = 500$ (0.971).

	MAE (1)	SD (2)	# invalid (3)	p allinv (4)	p oracle (5)
Panel (a), $N = 500$					
Oracle 2SLS	0.0106	0.0152	8	1	1
Naive 2SLS	0.2578	0.0198	0	0	0
$\widehat{\boldsymbol{\theta}}^{mm}$	0.0532	0.0580			
Post-ALasso	0.0109	0.0780	8.129	0.997	0.931
$\widehat{\boldsymbol{\theta}}^{mm}_{block}$	0.0470	0.0318			
Post-ALasso$_{block}$	0.0107	0.0157	8.025	1.000	0.983
Panel (b), $N = 1,000$					
Oracle 2SLS	0.0073	0.0107	8	1	1
Naive 2SLS	0.2579	0.0141	0	0	0
$\widehat{\boldsymbol{\theta}}^{mm}$	0.0377	0.0553			
Post-ALasso	0.0074	0.0274	8.085	0.996	0.950
$\widehat{\boldsymbol{\theta}}^{mm}_{block}$	0.0324	0.0222			
Post-ALasso$_{block}$	0.0073	0.0109	8.017	1.000	0.986
Panel (c), $N = 2,000$					
Oracle 2SLS	0.0051	0.0078	8	1	1
Naive 2SLS	0.2577	0.0098	0	0	0
$\widehat{\boldsymbol{\theta}}^{mm}$	0.0281	0.0594			
Post-ALasso	0.0052	0.1315	8.094	0.996	0.952
$\widehat{\boldsymbol{\theta}}^{mm}_{block}$	0.0242	0.0155			
Post-ALasso$_{block}$	0.0051	0.0077	8.013	1.000	0.989

This table reports IV selection and estimation results of the adaptive Lasso method with the block structure in simulation design (1) with partial overlap. The reported statistics include median absolute error (column 1), standard deviation (column 2), number of IVs selected as invalid (column 3), frequency with which all invalid IVs have been selected as invalid (column 4), and frequency with which oracle model has been selected (column 5). The simulations are based on 1,000 repetitions.

Table 2.6: IV selection with the block structure – Simulation design (1) with partial overlap.

	MAE (1)	SD (2)	# invalid (3)	p allinv (4)	p oracle (5)
Panel (a), $N = 500$					
Oracle 2SLS	0.0109	0.0161	9	1	1
Naive 2SLS	0.3285	0.0209	0	0	0
$\widehat{\boldsymbol{\theta}}^{mm}$	0.1124	0.1472			
Post-ALasso	0.0159	0.1779	10.241	0.791	0.515
$\widehat{\boldsymbol{\theta}}^{mm}_{block}$	0.0839	0.0394			
Post-ALasso$_{block}$	0.0111	0.0192	9.044	0.999	0.971
Panel (b), $N = 1,000$					
Oracle 2SLS	0.0075	0.0111	9	1	1
Naive 2SLS	0.3288	0.0149	0	0	0
$\widehat{\boldsymbol{\theta}}^{mm}$	0.0892	0.1629			
Post-ALasso	0.0102	0.2413	10.052	0.786	0.565
$\widehat{\boldsymbol{\theta}}^{mm}_{block}$	0.0599	0.0283			
Post-ALasso$_{block}$	0.0076	0.0115	9.019	1.000	0.987
Panel (c), $N = 2,000$					
Oracle 2SLS	0.0054	0.0080	9	1	1
Naive 2SLS	0.3286	0.0107	0	0	0
$\widehat{\boldsymbol{\theta}}^{mm}$	0.0703	0.1667			
Post-ALasso	0.0071	0.1847	10.086	0.803	0.544
$\widehat{\boldsymbol{\theta}}^{mm}_{block}$	0.0411	0.0196			
Post-ALasso$_{block}$	0.0054	0.0084	9.020	1.000	0.987

This table reports IV selection and estimation results of the adaptive Lasso method with the block structure in simulation design (2) with no overlap. The reported statistics include median absolute error (column 1), standard deviation (column 2), number of IVs selected as invalid (column 3), frequency with which all invalid IVs have been selected as invalid (column 4), and frequency with which oracle model has been selected (column 5). The simulations are based on $1,000$ repetitions.

Table 2.7: IV Selection with the block structure – Simulation design (2) with no overlap

2.6 Application: The Effects of Educational Attainment and Cognitive Ability on BMI

We apply our IV selection method to a multivariable Mendelian randomization (MVMR) study. We estimate the effects of educational attainment and cognitive ability on Body Mass Index (BMI), as in Sanderson et al. (2019). Both educational attainment and cognitive ability have been found to be negatively correlated with BMI (Sanderson et al., 2019). However, as educational attainment and cognitive ability are highly correlated, it is unclear to what extent each of them have a direct effect on BMI. In this application, we account for both variables in order to disentangle their direct effects. We use 74 SNPs as instruments for educational attainment and 18 SNPs for cognitive ability, and one SNPs overlaps between the two sets of candidate instruments. These SNPs have previously been identified in independent Genome-Wide Association Studies (GWAS), see Okbay et al. (2016) for educational attainment, and Sniekers et al. (2017) for cognitive ability. We use data on $107,371$ individuals from the UK Biobank. Educational attainment is measured in years of completed education, and it is imputed based on the individuals' qualifications, which is standard in the literature, see, e.g., Okbay et al. (2016). Cognitive ability is measured as a unitless fluid intelligence score that the UK biobank constructs from a series of tests completed by the individuals during assessment. We standardise the cognitive ability to mean zero and variance one. BMI is the ratio of weight to height, both of which were measured for all individuals during assessment, and we log-transform it due to skewness. Hence, we interpret our estimates as the percentage change in BMI that is associated with a one unit increase in the relevant explanatory variable. We also include additional covariates that control for age at assessment, sex, and the first 10 genetic principal components, all of which are available from the UK biobank. See Sanderson et al. (2019) for a detailed definition of the variables and presentation of the data.

Table 2.8 reports the results of our analysis. Columns (1) and (2) show, respectively, the point estimates and their standard errors. Column (3) is the number of instruments selected as invalid, and column (4) shows the p-value of the Sargan test. Panel (a) presents the estimates from a naive 2SLS regression where we treat

	Estimate (1)	Std. error (2)	# Invalid (3)	p-value, Sargan (4)
Panel (a) – 2SLS				
Educational attainment	−0.035	0.004	0	1.69e-13
Cognitive ability	0.031	0.011		
Panel (b) – Post-ALasso				
Educational attainment	−0.029	0.005	10	0.011
Cognitive ability	0.021	0.012		
med_{edu}	−0.031			
med_{cog}	0.017			
Panel (c) – Post-ALasso$_{block}$				
Educational attainment	−0.029	0.005	10	0.011
Cognitive ability	0.021	0.012		
med_{edu}	−0.034			
med_{cog}	0.031			

This table reports the estimation results of the effects of educational attainment and cognitive ability on $ln(BMI)$. The sample size is $n = 107371$. The number of instruments for educational attainment is $k_{edu} = 74$. The number of instruments for cognitive ability is $k_{cog} = 18$. There is one instrument identified for both educational attainment and cognitive ability.

Table 2.8: The impacts of educational attainment and cognitive ability on $ln(BMI)$

.

all the candidate instruments as valid. Both estimates are statistically significant at the 1% level. However, these results are from the naive 2SLS regression, and they might be biased due to the presence of invalid instruments. This is supported by the small p-value of the Sargan test (1.69e-13). In practice, SNPs can exhibit so-called pleiotropic effects, which would make them invalid instruments. In our setting, pleiotrophy would mean that some of the SNPs, either for educational attainment or cognitive ability, have direct effects on BMI.

Instead of the naive 2SLS, we now conduct IV selection using the adaptive Lasso with the downward testing procedure, as described in Section 2.3.2, and we obtain post-selection 2SLS estimates. In Panel (b) and (c) of Table 2.8, we

report the results for the direct effects of educational attainment and cognitive ability using our method, and we show the estimates both with (Panel (c)) and without (Panel (b)) the block structure. We also present the associated median-of-medians estimates, denoted med_{edu} and med_{cog} for, respectively, educational attainment and cognitive ability. The threshold p-value for the Sargan test is $0.1/log(n) = 0.0086$.

For educational attainment, the median-of-medians estimates are insensitive to the use of the block structure, and we find that $med_{edu} = -0.0314$ with the block structure and $med_{edu} = -0.034$ without. For cognitive ability, the estimates differ significantly, and the estimate with the block structure is $med_{cog} = 0.017$, while it is $med_{cog} = 0.031$ without. Consider that the two sets of SNPs for educational attainment and cognitive ability are identified in separate GWAS studies, it may be preferable to estimate with the block structure, although in this case the adaptive Lasso selects the same 10 instruments as invalid with two methods.

The method selects the same invalid instruments with and without the block structure, as shown in Column (3). Therefore, the estimates of the direct effects and their standard errors are the same in Panel (b) and (c). We find that our method selects 10 instruments as invalid. Six of these are for educational attainment, three are for cognitive ability, and one is for both variables. As seen in Column (4), the Sargan statistic for the selected model is 0.011, which is a significant improvement over the Sargan statistic reported for the naive 2SLS in Panel (a). We find that both post-selection 2SLS estimates are closer to zero compared to the estimates for the naive 2SLS. The post-selection estimate for educational attainment is -0.029, while, for cognitive ability, it is 0.021.

Compared to the results for the naive 2SLS, the effect of educational attainment on BMI is still significant at the 1% level, while the effect of cognitive ability is now insignificant, even at the 5% level. These results indicate that higher educational attainment lowers BMI, although, the naive 2SLS regression may have exaggerated the magnitude of this effect. We find limited evidence of a direct effect of cognitive ability on BMI, as the estimate becomes smaller and statistically insignificant when we conduct IV selection.

For the results in Table 2.8, we assume conditional homoskedasticity. However, a robust version of our method, i.e. using the two-step Hansen-J test and the post-

selection two-step GMM estimator, produces almost identical results. We use the Sanderson-Windmeijer conditional F-statistic (Sanderson and Windmeijer, 2016) to evaluate the power of the instruments in predicting educational attainment and cognitive ability jointly. When we include the instruments in the naive 2SLS, the conditional F statistics are 2.57 for educational attainment and 2.65 for cognitive ability. Both of them are significantly lower than the rule-of-thumb value of 10, showing that the joint prediction power of the instruments is relatively weak. One way to deal with this weak IV problem to create a weighted score of all the instruments, that is, one score for each of educational attainment and cognitive ability, and then use these two scores as the instruments in the regression. In the naive 2SLS, when we use the weighted scores, the conditional F statistics are 67.73 for educational attainment and 68.65 for cognitive ability, which are much larger than the rule-of-thumb value of 10. For the post-selection 2SLS, we create the weighted scores using only the selected valid instruments. The estimate for educational attainment is -0.042 (se 0.009) and for cognitive ability it is 0.041 (se 0.024). This maintains the conclusion that educational attainment has a significant negative effect on BMI, while the direct effect of cognitive ability is insignificant.

2.7 Conclusion

We investigate the use of the adaptive Lasso method for selecting valid instrumental variables from a set of candidate instruments when some of the instruments may be invalid. While existing work has focused on a single endogenous variable, our method contributes to the literature by allowing for two endogenous variables. Under the modified majority rule, we show that the adaptive Lasso method can achieve consistent selection and oracle estimation. In this work, we consider the number of candidate instruments to be fixed, but in some settings it may grow with the sample size (or even at a quicker rate), and, therefore, future research will focus on extending the method to handle such cases.

2.A Appendix

Proof. The proof for Theorem 2.1 follows a similar way as the proof for Theorem 1 in Windmeijer et al. (2019). Under the model setup (2.5) and (2.8), and stated assumptions 2.1-2.5, for the first entry in the just-identified estimator $\widehat{\boldsymbol{\theta}}_{jk} = (\widehat{\beta}_{jk}, \widehat{\tau}_{jk})'$ as defined in (2.14) using Z_j and Z_k as valid instruments, following (2.15), we have

$$plim(\widehat{\beta}_{jk}) = \beta + (\boldsymbol{\Pi}_{jk}^{-1}\boldsymbol{\alpha}_{jk})_1 = \beta + \delta_{jk}$$

where $j, k = 1, ..., k_z, j \neq k$. From the definition of the valid instrument, it follows that if $j, k \in \mathcal{V}$, then $\delta_{jk} = 0$ as $\boldsymbol{\alpha}_{jk} = \mathbf{0}$. If $j \in \mathcal{V}$ while $k \in \mathcal{A}$, depending on the value of $\boldsymbol{\Pi}_{jk}^{-1}$, in some special cases, it is possible that δ_{jk} equals to 0 as $\alpha_j = 0$. Here we focus on the general case where $\delta_{jk} \neq 0$. By the continuous mapping theorem, it follows that

$$plim(\widehat{\beta}_j^m) = median(plim(\widehat{\boldsymbol{\beta}}_j)) = median(\beta\boldsymbol{\iota}_{k_z-1} + \boldsymbol{\delta}_j) = \beta + \widehat{\delta}_j^m \tag{2.A.1}$$

If $j \in \mathcal{V}$, as $k_{\mathcal{A}} < (k_z - 1)/2$, then more than 50% of the entries in $\boldsymbol{\delta}_j$ are equal to zero. It follows that $\widehat{\delta}_j^m$ is equal to zero, thus $\widehat{\beta}_j^m$ is consistent for β. If $j \in \mathcal{A}$, $\boldsymbol{\delta}_{jk}$ are nonzero for all k, then δ_j^m is nonzero. From (2.A.1) it follows that $\widehat{\beta}_j^m$ is inconsistent for β.
Similar to (2.A.1), we have

$$plim(\widehat{\beta}^{mm}) = median(plim(\widehat{\beta}_1^m), ..., plim(\widehat{\beta}_{k_z}^m)) \tag{2.A.2}$$

$k_{\mathcal{A}} < (k_z - 1)/2$ indicates that $k_{\mathcal{A}} < k_z/2$. Therefore, the majority of $\widehat{\beta}_j^m$ satisfies $plim(\widehat{\beta}_j^m) = \beta$. It follows that from (2.A.2) that $\widehat{\beta}^{mm}$ is a consistent estimator for β

$$plim(\widehat{\beta}^{mm}) = median(plim(\widehat{\beta}_1^m), ..., plim(\widehat{\beta}_{k_z}^m)) = \beta$$

By the delta method, the limiting distribution of $\widehat{\boldsymbol{\beta}}_j$ can be obtained as

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_j - (\beta\boldsymbol{\iota}_{k_z-1} + \boldsymbol{\delta}_j)\right) \xrightarrow{d} N(0, \boldsymbol{\Sigma}_j) \tag{2.A.3}$$

For $j \in \mathcal{V}$, let $\boldsymbol{\delta}_j = (\boldsymbol{\delta}_{j,\mathcal{A}}', \mathbf{0}_{k_{\mathcal{V}-1}}')'$ where $\boldsymbol{\delta}_{j,\mathcal{A}}$ includes all $\delta_{jk} \neq 0$. Partition $\widehat{\boldsymbol{\beta}}_j$

accordingly as $\widehat{\boldsymbol{\beta}}_j = (\widehat{\boldsymbol{\beta}}'_{j,\mathcal{A}}, \widehat{\boldsymbol{\beta}}'_{j,\mathcal{V}})'$. Then for $\widehat{\boldsymbol{\beta}}_{j,\mathcal{V}}$, we have

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{j,\mathcal{V}} - \beta\boldsymbol{\iota}_{k_{\mathcal{V}}-1})\right) \xrightarrow{d} N\left(0, \boldsymbol{\Sigma}_{j,\mathcal{V}}\right) \tag{2.A.4}$$

As

$$\sqrt{n}(\widehat{\beta}_j^m - \beta) = median(\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_j - \beta\boldsymbol{\iota}_{k_z-1}\right)) = median\begin{pmatrix}\sqrt{n}(\widehat{\boldsymbol{\beta}}_{j,\mathcal{A}} - (\beta\boldsymbol{\iota}_{k_{\mathcal{A}}} + \widehat{\boldsymbol{\delta}}_{j,\mathcal{A}})) + \sqrt{n}\widehat{\boldsymbol{\delta}}_{j,\mathcal{A}} \\ \sqrt{n}(\widehat{\boldsymbol{\beta}}_{j,\mathcal{V}} - \beta\boldsymbol{\iota}_{k_{\mathcal{V}_0}-1})\end{pmatrix},$$

using the continuous mapping theorem, the limiting distribution of $\sqrt{n}(\widehat{\beta}_j^m - \beta)$ is the median of the limiting distribution of $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_j - \beta\boldsymbol{\iota}_{k_z-1}\right)$. It follows from (2.A.3) and (2.A.4) that $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{j,\mathcal{V}} - \beta\boldsymbol{\iota}_{k_{\mathcal{V}_0}-1}) = Op(1)$ while $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{j,\mathcal{A}} - (\beta\boldsymbol{\iota}_{k_{\mathcal{A}}} + \widehat{\boldsymbol{\delta}}_{j,\mathcal{A}})) + \sqrt{n}\widehat{\boldsymbol{\delta}}_{j,\mathcal{A}} \longrightarrow \infty$. From the previous result, we know that the majority of the entries of $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_j - \beta\boldsymbol{\iota}_{k_z-1}\right)$ are in $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{j,\mathcal{V}} - \beta\boldsymbol{\iota}_{k_{\mathcal{V}_0}-1})$. Therefore, the limiting distribution of $\widehat{\beta}_j^m$ is

$$\sqrt{n}(\widehat{\beta}_j^m - \beta) = median(\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_j - \beta\boldsymbol{\iota}_{k_z-1}\right)) \xrightarrow{d} q_{j,[l],k_z,k_{\mathcal{A}}} \tag{2.A.5}$$

For $j \in \mathcal{A}$, we have

$$\sqrt{n}(\widehat{\beta}_j^m - \beta) = median(\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_j - \beta\boldsymbol{\iota}_{k_z-1}\right)) = median(\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_j - (\beta\boldsymbol{\iota}_{k_z-1} + \boldsymbol{\delta}_j)\right) + \sqrt{n}\boldsymbol{\delta}_j)$$

As $\boldsymbol{\delta}_j \neq \mathbf{0}$, it follows from (2.A.3) that $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_j - (\beta\boldsymbol{\iota}_{k_z-1} + \boldsymbol{\delta}_j)\right) + \sqrt{n}\boldsymbol{\delta}_j \longrightarrow \infty$, then

$$\sqrt{n}(\widehat{\beta}_j^m - \beta) = median(\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_j - \beta\boldsymbol{\iota}_{k_z-1}\right)) \longrightarrow \infty \tag{2.A.6}$$

Let $\widehat{\boldsymbol{\beta}}_{\mathcal{V}}^m$ and $\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^m$ be the vectors containing $\widehat{\beta}_j^m$ for $j \in \mathcal{V}$ and $j \in \mathcal{A}$ respectively. Then we have

$$\sqrt{n}\left(\widehat{\beta}^{mm} - \beta\right) = median\begin{pmatrix}\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^m - \beta\boldsymbol{\iota}\right) \\ \sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\mathcal{V}}^m - \beta\boldsymbol{\iota}\right)\end{pmatrix}$$

As $k_{\mathcal{A}} < k_z/2$, it follows from (2.A.5) and (2.A.6) that

$$\sqrt{n}\left(\widehat{\beta}^{mm} - \beta\right) \xrightarrow{d} Q_{[L],k_z,k_{\mathcal{A}}}.$$

□

Chapter 3

Agglomerative Hierarchical Clustering for Selecting Valid Instrumental Variables

Abstract

We propose a procedure, which combines hierarchical clustering with a test of overidentifying restrictions for selecting valid instrumental variables (IV) from a large set of IVs. Some of these may be invalid in the sense that they may fail the exclusion restriction. We show that if the largest group of IVs is valid, our method achieves oracle properties. The advantage of our method is that it addresses weak instruments, multiple endogenous regressors and heterogeneous treatment effects. In simulations our procedure outperforms the Hard Thresholding and the Confidence Interval method. The method is applied to estimating the effect of immigration on wages.

This chapter is co-authored with Nicolas Apfel, who was a PhD student at the time of writing this paper. Xiaoran and Nicolas made equal and major contributions to all aspects of the paper. This paper is available on arXiv.

3.1 Introduction

Instrumental variables estimation is a widely used statistical method for analysing the causal effects of treatment variables on an outcome when the causal relationship between them is confounded. For consistent IV estimation, all instruments must be valid, which requires that

(a) The instruments are associated with the endogenous variables (relevance condition)

(b) The instruments do not affect the outcome directly or through unobserved factors (exclusion restriction)

In practice, a main challenge in IV estimation is that when there are many candidate instruments, some of them may be invalid in the sense that they fail the exclusion restriction. Hence, the key task is to estimate the causal effect in situations where many IVs are invalid.

In this paper, we propose a new method to select the valid instruments and to estimate the causal effect. The method combines the agglomerative hierarchical clustering (AHC) algorithm, a statistical learning algorithm typically employed in cluster analysis, with the Sargan test for overidentifying restrictions. Our estimator relies on the plurality rule (Guo et al., 2018), which states that the largest group of IVs consists of valid instruments. Instruments are said to form a group if their instrument-specific just-identified estimators converge to the same value. Under the plurality rule, our method achieves oracle selection. This means that the estimator works as well as if the true set of valid instruments were known; valid instruments can be selected consistently, and the two-stage least squares (2SLS) estimator, using the instruments selected as valid, has the same limiting distribution as the ideal estimator that uses the true set of valid instruments. Our method improves the existing methods as it (1) allows for multiple endogenous regressors, (2) deals effectively with weak instruments, and (3) accommodates heterogeneous effects. In simulations, we also show that it outperforms the existing methods.

An example of a setting with many IVs is the estimation of the effect of immigration on wages in labor economics. To identify the causal effect, researchers

often use the lagged origin-country specific immigration pattern, as measured by previous shares of immigrants. If all the IVs are valid, that is, if none of the previous shares by origin-country are directly or indirectly correlated with wages, the causal effect can be consistently estimated. This assumption is often invoked in the literature.[1] However, some of the shares may violate the exclusion restrictions, as they may affect the wage variable directly through long-term dynamic adjustment processes, or be correlated with unobserved demand shocks.

Another field that makes use of many instruments, some of which may be invalid, is Mendelian Randomization (MR). In MR, researchers use genetic variation to estimate the causal effect of an exposure on an outcome (Von Hinke et al., 2016). This field has also inspired much of the initial invalid IV selection literature. An example is the estimation of the effect of C-reactive protein on coronary heart disease (Wensley et al., 2011).

In the applied literature, the common solutions is to either (1) select the valid instruments from the set of potential instruments based on economic intuition, or (2) directly include all the candidate instruments in IV estimation. These approaches can be problematic because including invalid instruments often leads to severely biased results. Therefore, under incomplete knowledge of the instruments validity, it is important to develop data-driven IV selection methods to select the valid instruments.

Previous work focus on the IV selection problem in the case of a single endogenous variable. Kang et al. (2016) propose a selection method based on the least absolute shrinkage and selection operator (LASSO). Windmeijer et al. (2019) make improvements by proposing an adaptive Lasso based method that has oracle properties under the assumption that more than half of the candidate instruments are valid (the *majority* rule). Guo et al. (2018) propose the Hard Thresholding with Voting method (HT), which has oracle properties under the sufficient and necessary identification condition that the largest group is formed by all the valid instruments (the *plurality* rule). This is a relaxation of the majority rule. Under the same identification condition, Windmeijer et al. (2021) propose the Confidence Interval method (CIM), which has better finite sample performance.
Our research adds to the literature in five ways:

[1] See Table 6 in Apfel (2021) for a non-exhaustive list of papers in this literature.

1. We combine agglomerative hierarchical clustering with a traditional statistical test, the Sargan over-identification test, to yield a novel downward testing algorithm for IV selection. Under the plurality rule, our method has the theoretical guarantee that it consistently selects the true set of valid instruments in a computationally feasible manner.

2. We extend our method to settings with multiple endogenous regressors. Previous methods do not allow for this, but, in our setting, it is a straightforward extension.

3. Our method performs well in the presence of weak instruments, whether they are valid or invalid, which is an advantage over existing methods.

4. We also discuss the application of our method to a setting with heterogeneous treatment effects. We show that we can retrieve and inspect the entire group structure, which is not possible with existing methods.

5. Computationally, our algorithm has lower complexity than the CI and HT methods. Also, the only pre-specified parameter for our algorithm is the critical value for the Sargan test. The optimal choice of this parameter, such that consistent selection is guaranteed, is well studied in the literature.

We use Monte Carlo simulations to examine the performance of our method, and compare it with the HT and CI methods. We benchmark against these methods as they also rely on the plurality rule. The simulation results show that our method achieves oracle performance in large sample settings, both with single and multiple endogenous regressors, when all the instruments are strong. Also, our method outperforms HT and CIM when some of the candidate instruments are weak. We apply our method to estimate the short- and long-run effects of immigration on wages in the US. We also provide an R-package that implements our method.

The remainder of this paper is structured as follows. In Section 3.2, we state the model and assumptions and illustrate some of the well-established properties of the 2SLS just-identified estimator. In Section 3.3, we describe the basic method and the algorithm when there is a single endogenous variable, and investigate its asymptotic properties. In Section 3.4, we present extensions to settings with

multiple endogenous regressors and weak instruments, and discuss our method in presence of heterogeneous treatment effects. In Section 3.5, we provide Monte Carlo simulation results. In Section 3.6, we apply our method to estimate the effects of immigration on wages. Section 3.7 concludes.

3.2 Model and Assumptions

3.2.1 Model Setup

In the following, we introduce the notation used in the paper. Matrices are in upper case and bold. Vectors are in lower case and bold. Scalars are in lower case and not in bold. Let $\mathbf{y}$ be an $n \times 1$-vector of the observed outcome, $\mathbf{d}_1$, ..., $\mathbf{d}_P$ be P endogenous regressor vectors (each $n \times 1$), which can be subsumed in an $n \times P$ -matrix $\mathbf{D}$, $\mathbf{z}_1$, ..., $\mathbf{z}_J$ be J instrument vectors, which can be subsumed in an $n \times J$ -matrix $\mathbf{Z}$. Let error terms be $\mathbf{u}$ and $\boldsymbol{\varepsilon_p}$ for $p \in \{1, ..., P\}$, which are all $n \times 1$ error-vectors and are correlated with $\sigma_{up} := cov(\mathbf{u}, \boldsymbol{\varepsilon}_p)$. The latter covariances measure the endogeneity of the regressors in $\mathbf{D}$. The $P \times 1$ coefficient vector of interest is $\boldsymbol{\beta}$. The $J \times P$ matrix $\boldsymbol{\gamma}$ contains the first-stage coefficients. Let s be the number of instruments in the set of invalid instruments, $\mathcal{I}$, g be the number of instruments in the set of valid instruments, $\mathcal{V}$, and $J = g + s$ be the total number of instruments in the overall set of instruments, $\mathcal{J}$. The arithmetic mean of a variable x is defined as $\mu_x = \frac{\Sigma x}{n}$, the mean of a vector is the vector of dimension-wise arithmetic means, $\|\cdot\|$ denotes the L2-norm, and $|\cdot|$ denotes cardinality when used around a set and an absolute value when used around a quantity. The symbol & denotes the logical conjunction, *and*. The $n \times n$ projection matrix is $\mathbf{P}_X = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, and the annihilator matrix is $\mathbf{M}_X = \mathbf{I} - \mathbf{P}_X$ and $\hat{\mathbf{D}} = \mathbf{P}_Z\mathbf{D}$ are the fitted values.

In Section 2 and Section 3, we consider a model with a single endogenous regressor, i.e. $P = 1$. The extension of our method to the case with multiple endogenous regressors can be found in Section 4.1. All the proofs in the Appendix are for a general P. Following the literature on invalid IV selection (Kang et al., 2016), we adopt the following observed data model, which takes the potentially

invalid instruments into account:

$$\mathbf{y} = \mathbf{d}\beta + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{u}, \tag{3.1}$$

with $\mathbf{E}[u_i|\mathbf{z}_i] = 0$. The linear projection of $\mathbf{d}$ on $\mathbf{Z}$ is

$$\mathbf{d} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{3.2}$$

The vector $\boldsymbol{\alpha}$ is $J \times 1$ and has entries α_j, each of which is associated with an individual instrument. Each entry indicates which of the instruments that has a direct effect on the outcome variable and hence is invalid. Following Definition 1 in Guo et al. (2018), we define a valid instrument as:

Definition 3.1. *For $j = 1, ..., J$, instrument $\mathbf{z}_j$ is valid if $\alpha_j = 0$. If $\alpha_j \neq 0$, then $\mathbf{z}_j$ is an invalid instrument.*

The ideal model, which selects the truly valid instruments as valid and controls for the set of invalid instruments, is the oracle model, defined as follows:

$$\mathbf{y} = \mathbf{d}\beta + \mathbf{Z}_{\mathcal{I}}\boldsymbol{\alpha}_{\mathcal{I}} + \mathbf{u} = \mathbf{X}_{\mathcal{I}}\boldsymbol{\theta}_{\mathcal{I}} + \mathbf{u}. \tag{3.3}$$

where $\mathbf{X} = (\mathbf{d} \quad \mathbf{Z}_{\mathcal{I}})$ and $\boldsymbol{\theta}_{\mathcal{I}} = (\beta \quad \boldsymbol{\alpha}'_{\mathcal{I}})'$.

3.2.2 Assumptions

The assumptions that follow are the same as in Windmeijer et al. (2021). The first assumption makes sure that all the just-identified estimators exist.

Assumption 3.1. *Identification of just-identified models.*

$$\boldsymbol{\gamma} = (E[\mathbf{z}_i\mathbf{z}'_i])^{-1}E[\mathbf{z}_i d_i]), \gamma_j \neq 0 \quad j = 1, ..., J.$$

Assumption 3.2. *Rank assumption.*

$$E(\mathbf{z}_i\mathbf{z}'_i) = \mathbf{Q} \text{ with } \mathbf{Q} \text{ a finite and full rank matrix.}$$

Assumption 3.3. *Error structure.*
Let $\mathbf{w}_i = (u_i \quad \varepsilon_i)'$. *Then,* $E(\mathbf{w}_i) = 0$ *and* $E[\mathbf{w}_i\mathbf{w}_i'] = \begin{pmatrix} \sigma_u^2 & \sigma_{u,\varepsilon} \\ \sigma_{u,\varepsilon} & \sigma_\varepsilon^2 \end{pmatrix} = \mathbf{\Sigma}$ *with* $Var(u_i) = \sigma_u^2, \quad Var(\varepsilon_i) = \sigma_\varepsilon^2 \quad Cov(u_i, \varepsilon_i) = \sigma_{u,\varepsilon}$ *and the elements of* $\mathbf{\Sigma}$ *are finite.*

Assumption 3.4.

$$\begin{aligned} plim(n^{-1}\mathbf{Z}'\mathbf{Z}) = E(\mathbf{z}_i\mathbf{z}_i') = \mathbf{Q} \quad &; \quad plim(n^{-1}\mathbf{Z}'\mathbf{d}) = E(\mathbf{z}_i d_i) \\ plim(n^{-1}\mathbf{Z}'\boldsymbol{u}) = E(\mathbf{z}_i u_i) = 0 \quad &; \quad plim(n^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}) = E(\mathbf{z}_i\varepsilon_i) = 0 \\ plim(n^{-1}\sum_{i=1}^{n}\mathbf{w}_i) = 0 \quad &; \quad plim(n^{-1}\mathbf{w}_i\mathbf{w}_i') = \mathbf{\Sigma}. \end{aligned}$$

Assumption 3.5. $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} vec(\mathbf{z}_i\mathbf{w}_i') \xrightarrow{d} N(0, \mathbf{\Sigma}\otimes\mathbf{Q})$ *as* $n \to \infty$.

We modify the assumptions above when there are more than one endogenous regressor. From (3.1) and (3.2), we have the outcome-instrument reduced form

$$\mathbf{y} = \mathbf{Z}\mathbf{\Gamma} + \boldsymbol{\epsilon}$$

where $\mathbf{\Gamma} = \boldsymbol{\gamma}\beta + \boldsymbol{\alpha}$. Each individual instrument $\mathbf{z}_j$ is associated with a just-identified estimator for β, denoted by $\hat{\beta}_j$, which is defined as the two-stage least squares (2SLS) estimator using $\mathbf{z}_j$ as the single valid instruments, and treating the remaining IVs as controls. There are J just-identified IV estimators. We write these estimators as in Windmeijer et al. (2021).

$$\hat{\beta}_j = \frac{\hat{\Gamma}_j}{\hat{\gamma}_j}$$

where $\hat{\Gamma}_j$ and $\hat{\gamma}_j$ are the OLS estimators for Γ_j and γ_j respectively. Then we have

Property 3.1. *Properties of just-identified estimates.*
Under Assumptions 3.1 to 3.5 it holds that

$$plim(\hat{\beta}_j) = plim\left(\frac{\hat{\Gamma}_j}{\hat{\gamma}_j}\right) = \beta + \frac{\alpha_j}{\gamma_j}$$

Hence, the inconsistency of $\hat{\beta}_j$ is $plim(\hat{\beta}_j) - \beta = \frac{\alpha_j}{\gamma_j} = q$. We define a group following the definition in Guo et al. (2018) as:

Definition 3.2. *A group $\mathcal{G}_q$ is a set of IVs that has the same estimand $\beta_j = \beta + q$.*

$$\mathcal{G}_q = \{j : \beta_j = \beta + q\}$$

The group consisting of all valid instruments is

$$\mathcal{G}_0 = \{j : q = 0\}$$

Let the number of groups be Q.

The next assumption is the key assumption for identification. It states that among the Q groups formed by $\mathbf{z}_1, ..., \mathbf{z}_J$, the largest group is composed by all the valid IVs. A group is defined as above, as a set of instruments whose just-identified estimators converge to the same value $\beta + q$.

Assumption 3.6. *Plurality Rule.*

$$g > \max_{q \neq 0} |\mathcal{G}_q|$$

3.3 IV Selection and Estimation Method

Based on the definition of groups, and the plurality rule, a natural strategy for IV selection is to find the Q IV groups, and then select the largest group as the set of valid instruments. In this paper, we explore clustering methods to discover the IV groups. First, we adapt the general clustering framework to the IV selection problem, which is summarized in the minimisation problem in 3.4. This general method needs a pre-specified parameter K, which is the number of clusters. We show that when K equals the number of groups, there is a unique solution to this minimization problem. This solution coincides with the true underlying partition. However, the fact that consistent selection depends on K makes it difficult to implement in practice, as we do not have prior knowledge about the number of groups. If K is too large (larger than the number of groups), then the largest

group will be split. If K is too small, then the largest group might be in a cluster with some other group. To tackle this problem, we propose a downward testing procedure that combines the agglomerative hierarchical clustering method (Ward's method) with the Sargan test for overidentifying restrictions to select the valid instruments. This procedure allows us to select the valid instruments without pre-specifying K.

3.3.1 Clustering Method for IV Selection

Let $\mathcal{S} = \{\mathcal{S}_1, ..., \mathcal{S}_K\}$ be a partition of J just-identified estimators $\hat{\beta}_j$ into K cluster cells. The clustering result is the solution to the following minimization problem:

$$\hat{\mathcal{S}}(K) = \underset{\mathcal{S}}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{\hat{\beta}_j \in \mathcal{S}_k} ||\hat{\beta}_j - \bar{\mathcal{S}}_k||^2, \tag{3.4}$$

where $\bar{\mathcal{S}}_k$ is the arithmetic mean of all just-identified estimators in cluster $\mathcal{S}_k$.
Let the clustering result $\hat{\mathcal{S}}(K)$ be an estimator of sets containing IV-estimators $\hat{\beta}_j$. The IV-estimators in a cluster $\hat{\mathcal{S}}_k$ are selected to belong to a certain group:

$$\hat{\mathcal{G}}_k = \{j : \hat{\beta}_j \in \hat{\mathcal{S}}_k\}$$

Based on Assumption 3.6, the cluster that consists of estimators that use valid IVs is estimated as the cluster that contains the largest number of just-identified estimators:

$$\hat{\mathcal{S}}_m(K) = \{\mathcal{S}(K) : |\hat{\mathcal{S}}(K)| = \underset{k}{max}|\hat{\mathcal{S}}_k(K)|\}$$

The valid IVs are selected as those IVs that are used to estimate the largest cluster $\hat{\mathcal{S}}_m(K)$

$$\hat{\mathcal{V}}(K) = \{j : \hat{\beta}_j \in \hat{\mathcal{S}}_m(K)\}$$

Then, the remaining IVs are selected as invalid

$$\hat{\mathcal{I}}(K) = \mathcal{J} \quad \setminus \quad \hat{\mathcal{V}}(K).$$

When the number of clusters K is equal to the number of groups Q, then there is a partition minimizing the sum in Equation 3.4. This occurs, when the grouping is such that $\hat{\mathcal{G}}_k = \mathcal{G}_q$, i.e. each selected group $\hat{\mathcal{G}}_k$ is in fact formed by a true group, $\mathcal{G}_q$. Define the partition leading to this grouping of IVs as the true partition $\mathcal{S}_0 = \{\mathcal{S}_{01}, ..., \mathcal{S}_{0Q}\}$.

To see that, first note that if the partition is such that $\hat{\mathcal{S}}_k = \mathcal{S}_{0q}\ \forall k, q$, i.e. $\hat{\mathcal{S}}(K) = \mathcal{S}_0$,

$$g(\hat{\mathcal{S}}(K)) = g(\mathcal{S}_0) = plim\{\sum_{k=1}^{K} \sum_{\hat{\beta}_j \in \mathcal{S}_k} ||\hat{\beta}_j - \bar{\mathcal{S}}_k||^2\} = 0.$$

For all $\hat{\beta}_j \in \mathcal{S}_k$, we have $plim\, \hat{\beta}_j = plim\, \bar{\mathcal{S}}_k$, and $plim\{||\hat{\beta}_j - \bar{\mathcal{S}}_k||^2\} = 0$. This is the case for all $k \in 1, ...K$, hence $g(\mathcal{S}_0) = 0$. Second, if the partition is such that some $\mathcal{S}_k \neq \mathcal{S}_{0q}$, i.e. $\mathcal{S} \neq \mathcal{S}_0$, then $plim\, \hat{\beta}_j \neq plim\, \bar{\mathcal{S}}_k$ for some $\hat{\beta}_j \in \mathcal{S}_k$ and $g(\mathcal{S}) > 0$. This means that when $n \to \infty$ there is a unique solution for Equation 3.4, which is such that $\mathcal{S} = \mathcal{S}_0$. A necessary condition for this to hold is that $K = Q$.

3.3.2 Ward's Algorithm for IV Selection

To choose the correct value of K without prior knowledge of the number of groups, we propose a selection method which combines Ward's algorithm, a general agglomerative hierarchical clustering procedure proposed by Ward (1963), with the Sargan test of overidentifying restrictions. Our selection algorithm has two parts, and the set of instruments selected as valid by the algorithm is denoted by $\hat{\mathcal{V}}^{dts}$.

The first part is Ward's algorithm, as described in Algorithm 3.1 below. Ward's algorithm aims to minimize the total within-cluster sum of squared error. This is achieved by minimizing the increase in within-cluster sum of squared error at each step of the algorithm. The method generates a path of cluster assignments with K clusters at each step so that $K \in \{1, ..., J\}$. After obtaining the clusters for each K, we use a downward testing procedure based on the Sargan-test to select the set of valid instruments (Algorithm 3.2).

Ward's Algorithm works as follows

Algorithm 3.1. *Ward's algorithm*

1. ***Input:*** *Each just-identified point estimate is calculated. The Euclidean distance between all of these estimates is calculated and written as a dissimilarity matrix.*

2. ***Initialization:*** *Each just-identified estimate has its own cluster. Hence, initially, the total number of clusters is J.*

3. ***Joining:*** *The two clusters that are closest as measured by their weighted squared Euclidean distance $\frac{|\mathcal{S}_k||\mathcal{S}_l|}{|\mathcal{S}_k|+|\mathcal{S}_l|}||\bar{\mathcal{S}}_k - \bar{\mathcal{S}}_l||^2$ are joined to a new cluster. $|\mathcal{S}_k|$ is the number of estimates in cluster k. $\bar{\mathcal{S}}_k$ denotes the mean of cluster k, which is the arithmetic mean of all the just-identified estimates in $\mathcal{S}_k$.*

4. ***Iteration:*** *The joining step is repeated until all just-identified point-estimates are in one cluster.*

This yields a path of $S = J - 1$ steps, on which there are clusters of size $K \in \{1, ..., J\}$. Ward (1963) also allows for alternative objective functions. These are associated with different dissimilarity metrics and different ways to define the distance between clusters. We discuss alternative choices of these so-called linkage methods and dissimilarity metrics in Section 3.4.4.

After generating the clustering path using Algorithm 3.1, we select the set of valid instruments following Algorithm 3.2:

Algorithm 3.2. *Downward testing procedure*

1. *Starting from $K = 1$, find the cluster that contains the largest number of just-identified estimators.*

2. *Do the Sargan test on the instruments associated with the largest cluster, using the rest of IVs as controls. If there are multiple such clusters, select the one with the smallest Sargan statistic.*

3. *Repeat the procedure for each $K = 2, ..., J - 1$.*

4. *Stop the first time that the model selected by the largest cluster at some K does not get rejected by the Sargan test.*

5. *Select the instruments associated with the cluster from Step 4 as valid instruments.*

The Sargan statistic in Step 4 is given by

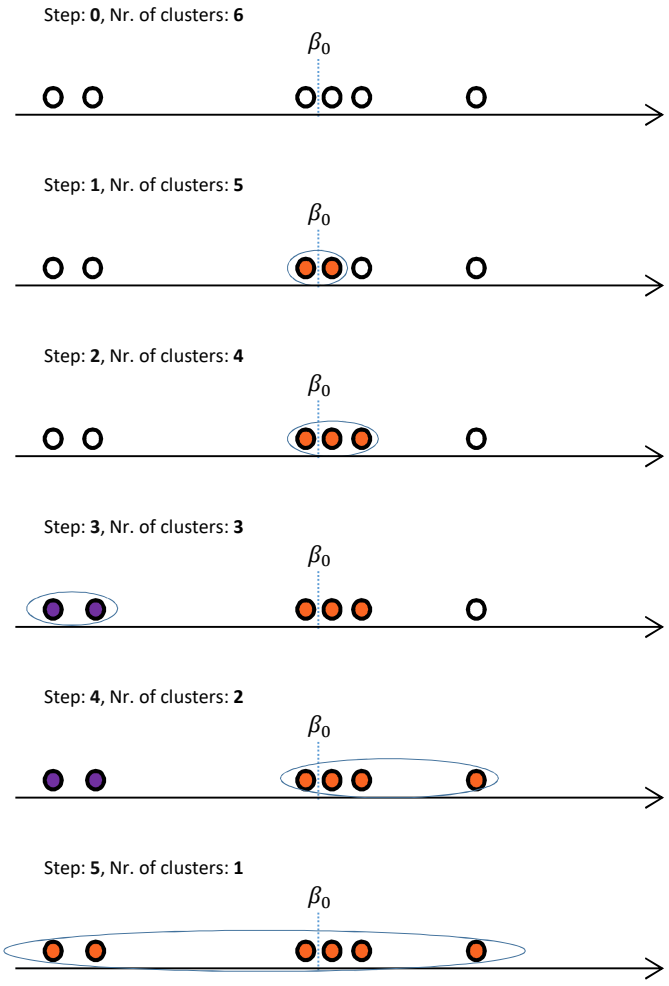
$$Sar\,(K) = \frac{\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_K)'\mathbf{Z}\,(\mathbf{Z}'\mathbf{Z})^{-1}\,\mathbf{Z}'\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_K)}{\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_K)'\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_K)/n}$$

where $\hat{\boldsymbol{\theta}}_K$ is the 2SLS estimator using the instruments associated with the largest cluster for each K as valid instruments and controlling for the rest of the instruments, and $\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_K)$ is the 2SLS residual. Later, we show that, to guarantee consistent selection, the critical value for the Sargan test, denoted by $\xi_{n,J-|\hat{\mathcal{I}}|-P}$, should satisfy $\xi_{n,J-|\hat{\mathcal{I}}|-P} \to \infty$ and $\xi_{n,J-|\hat{\mathcal{I}}|-P} = o(n)$. In practice, we choose the significance level $\frac{0.1}{log(n)}$ following Windmeijer et al. (2021).

The procedure is illustrated in Figure 3.1, which shows a situation with six instruments. Three of them are valid as they affect the outcome variable only through the endogenous regressor, while it is not the case for the other three invalid instruments. In the figure, the circles above the real line denote the just-identified estimate for the coefficient β_0 estimated by each of the six instruments. In the explanation below, we refer to these estimates and their corresponding instruments by No. 1 to No. 6, from left to right.

In the initial Step (0) of the clustering process, each just-identified estimate has its own cluster. In Step (1), we join the two estimates which are closest in terms of their weighted Euclidean distance, i.e. those estimated with instrument No. 3 and No. 4 (the two orange circles). These two estimates now form one cluster and we only have five clusters. We re-calculate the distances with the new cluster and merge the closest two into a new cluster. We continue with this procedure, until there is only one cluster left in the bottom right graph. We continue with Algorithm 3.2 and evaluate the Sargan test at each step, using the instruments contained in the largest cluster. When the p-value is larger than a certain threshold, say $0.1/log(n)$, we stop the procedure. Ideally this will be the case at Step (3) of the algorithm, because here the largest group (in orange) is formed only by valid IVs (No. 2, No. 3 and No. 4). If this is the case, only the valid IVs are selected as valid.

Figure 3.1: Illustration of the Algorithm with One Regressor



3.3.3 Oracle Selection and Estimation Property

In this section, we state the theoretical properties of the IV selection results obtained by Algorithm 3.1 and Algorithm 3.2, and of the post-selection estimators. See Section 3.4 for detailed theoretical results developed for the general case $P \geq 1$. We establish that our method can achieve oracle properties in the sense that it can select the valid instruments consistently, and that the post-selection IV estimator has the same limiting distribution as if we knew the true set of valid instruments.

Theorem 3.1. *Consistent selection*
Let ξ_n be the critical value for the Sargan test in Algorithm 3.2. Let $\hat{\mathcal{V}}^{dts}$ be the set of instruments selected from Algorithm 3.1 and Algorithm 3.2. Under Assumptions 3.1 - 3.6, for $\xi_n \to \infty$ and $\xi_n = o(n)$,

$$\lim_{n\to\infty} P(\hat{\mathcal{V}}^{dts} = \mathcal{V}) = 1.$$

The post-selection 2SLS estimator using the selected valid instruments and controlling for the selected invalid instruments has the same asymptotic distribution as the oracle estimator:

Theorem 3.2. *Asymptotic oracle distribution*
Let $\mathbf{Z}_{\hat{\mathcal{I}}} = \mathbf{Z} \setminus \mathbf{Z}_{\hat{\mathcal{V}}^{dts}}$ with $\mathbf{Z}_{\hat{\mathcal{I}}}$, $\mathbf{Z}_{\hat{\mathcal{V}}^{dts}}$ being the selected invalid and valid instruments respectively. Let $\hat{\beta}_{\hat{\mathcal{V}}^{dts}}$ be the 2SLS estimator given by

$$\hat{\beta}_{\hat{\mathcal{V}}^{dts}} = (\hat{\mathbf{d}}'\mathbf{M}_{\mathbf{Z}_{\hat{\mathcal{I}}}}\hat{\mathbf{d}})^{-1}\hat{\mathbf{d}}'\mathbf{M}_{\mathbf{Z}_{\hat{\mathcal{I}}}}\mathbf{y}$$

Under Assumptions 3.1 - 3.6, the limiting distribution of $\hat{\beta}_{\hat{\mathcal{V}}^{dts}}$ is

$$\sqrt{n}(\hat{\beta}_{\hat{\mathcal{V}}^{dts}} - \beta) \xrightarrow{d} N(0, \sigma^2_{or})$$

where σ^2_{or} is the asymptotic variance for the oracle 2SLS estimator given by

$$\sigma^2_{or} = \sigma^2_u \left(E\left[\mathbf{z}_{i.}d_i\right]' E\left[\mathbf{z}_{i.}\mathbf{z}'_{i.}\right]^{-1} E\left[\mathbf{z}_{i.}d_i\right] - E\left[\mathbf{z}_{\mathcal{I},i.}d_i\right]' E\left[\mathbf{z}_{\mathcal{I},i.}\mathbf{z}'_{\mathcal{I},i.}\right]^{-1} E\left[\mathbf{z}_{\mathcal{I},i.}d_i\right] \right)^{-1}.$$

with $\mathcal{I}$ being the true set of invalid instruments.

The proof of Theorem 3.2 follows from the proof of Guo et al. (2018, Consistent selection leads to oracle properties, Theorem 2)

3.3.4 Computational Complexity

Recent implementations of the hierarchical agglomerative clustering algorithm have a computational cost of $O(J^2)$ (Amorim et al., 2016). In the downward testing procedure, a maximum of $J-1$ different models need to be tested. Therefore, the computational cost of the downward testing algorithm is $O(J^2)$. This is

an improvement over the CIM, which has a time complexity of $O(J^2 log(J))$ and where the maximal number of tests is $J(J-1)/2$.

3.4 Extensions

In this section, we propose an extension of our method to a setting with multiple endogenous regressors, and we discuss the performance of our method in the presence of weak instruments as compared with the HT and CI method. We also discuss a setting with heterogeneous treatment effects.

3.4.1 Multiple Endogenous Regressors

One shortcoming of previous IV selection methods is that they only allow for one endogenous regressor. Therefore, in this section, we show how our method can be extended to select invalid instruments when $P > 1$. First of all, the inputs to our method, all the just-identified estimators, are estimated by all the P-combinations from $\mathbf{z}_1, ..., \mathbf{z}_J$. Hence, we now have $\binom{J}{P}$ instead of J just-identified estimators. Let $[j]$ be a set of indices of any P instruments such that the model is exactly identified with these P instruments. Let $\mathbf{Z}_{[j]}$ denote the corresponding $n \times P$ instrument matrix. To guarantee that all the $\binom{J}{P}$ just-identified estimators exist, we modify Assumption 3.1 as follows:

Assumption 3.1.a. *Existence of just-identified estimators*
For all possible values of $[j]$*, let* $\boldsymbol{\gamma}_{[j]}$ *be the combinations of the* k_{th}*-rows of* $\boldsymbol{\gamma}$ *for all* $k \in [j]$*. Then we assume*

$$rank(\boldsymbol{\gamma}_{[j]}) = P.$$

The plurality assumption must also be modified for $P > 1$. For $P = 1$, Assumption 3.6 states that the valid instruments form the largest group, where instruments form a group if their just-identified estimators converge to the same value. If we find the largest set of just-identified estimators that converge to the same value, then this set is automatically the largest group of instruments, as each just-identified estimator is estimated by a single instrument. However, when $P > 1$, each just-identified estimator is estimated by multiple instruments, hence

the equivalence between the largest set of just-identified estimators and the largest group of instruments may not hold. In this case, we modify the plurality rule so it is based on the combinations of P instruments instead of individual instruments. This modification starts with revisiting the asymptotics of the just-identified estimators for $P > 1$. The details can be found in Appendix 3.A.2.

Let $\hat{\boldsymbol{\beta}}_{[j]}$ be the just-identified 2SLS estimator estimated with $\mathbf{Z}_{[j]}$, then analogously to the case with one regressor, we have the following property of the just-identified estimates:

Property 3.2. *Properties of the just-identified estimates with $P \geq 1$*
Under Assumptions 3.1.a to 3.5 it holds that

$$plim\hat{\boldsymbol{\beta}}_{[j]} = \boldsymbol{\beta} + \boldsymbol{\gamma}_{[j]}^{-1}\boldsymbol{\alpha}_{[j]} = \boldsymbol{\beta} + \mathbf{q}$$

where the inconsistency of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}_{[j]} - \boldsymbol{\beta} = \boldsymbol{\gamma}_{[j]}^{-1}\boldsymbol{\alpha}_{[j]} = \mathbf{q}$ and there are $\binom{J}{P}$ inconsistency terms $\mathbf{q}$. Note that $\mathbf{q}$ is a $P \times 1$ vector. When $P > 1$, not every IV is associated with a single scalar q, so we introduce the concept of a *family*:

Definition 3.3. *A family is a set of just-identifying IV combinations that is associated with just-identified estimators which converge to the same value.*

$$\mathcal{F}_q = \{[j] : \boldsymbol{\beta}_{[j]} = \boldsymbol{\beta}_0 + \mathbf{q}\}$$

Note that each element of a family is itself a set of P IVs, such that a model is just-identified. By definition, the family that consists of IV combinations that generate consistent estimators is

$$\mathcal{F}_0 = \{[j] : \mathbf{q} = \mathbf{0}\}.$$

Let there be Q families. Note that when $P = 1$ a group of IVs is automatically a family.

Analogously to Assumption 3.6, we assume that $\mathcal{F}_0$ is the largest family:

$$|\mathcal{F}_0| > \max_{q \neq 0} |\mathcal{F}_q|$$

We show in Appendix 3.A.3, that a combination of IVs is an element of $\mathcal{F}_0$ if and

only if all of the P IVs in the combination are in fact valid. This means that the family of valid IVs consists of all combinations that use P IVs from the set of valid instruments, $\mathcal{V}$, and hence $|\mathcal{F}_0| = \binom{g}{P}$. Therefore, the plurality assumption can be modified to

Assumption 3.6.a. *Family plurality*

$$\binom{g}{P} > \underset{\mathbf{q}\neq\mathbf{0}}{max}|\mathcal{F}_q|$$

The inconsistency term of elements in $\mathcal{F}_{\mathbf{q}}$ with $\mathbf{q} \neq \mathbf{0}$ depends on the first-stage coefficient vectors and hence there is no direct relation from $\boldsymbol{\alpha}_{[j]}$ to $\mathbf{q}$. One way in which this new plurality can be fulfilled, is when the largest set of IVs has zero direct effects $\alpha_j = 0$. Moreover, the vectors $\boldsymbol{\gamma}_{[j]}^{-\mathbf{1}}\boldsymbol{\alpha}_{[j]}$ constituted by P-sets with $\boldsymbol{\alpha}_{[j]} \neq 0$ are sufficiently dispersed. Strictly speaking, the family plurality assumption can also hold when the largest group of IVs has some direct effect $\alpha_j = c$. If the dispersion of $\boldsymbol{\gamma}_{[j]}^{-\mathbf{1}}\boldsymbol{\alpha}_{[j]}$ is large enough, the largest family will still be constituted by valid IVs only.

The procedure to estimate $\mathcal{V}$ is analogous to the one in the preceding section (see Appendix 3.A.1 for an illustration) except that we now need to account for the presence of families. First, for a certain number of clusters, K, a unique cluster is selected by the algorithm. This works as follows: The algorithm selects the cluster which contains the largest number of point estimates, $\hat{\beta}_{[j]}$, as the cluster potentially associated with the valid instruments at K. Again, this largest cluster is $\hat{\mathcal{S}}_m(K)$.

$$\hat{\mathcal{S}}_m(K) = \{\hat{\mathcal{S}}(K) : |\hat{\mathcal{S}}(K)| = \underset{k}{max}|\hat{\mathcal{S}}_k(K)|\}$$

The cluster $\hat{\mathcal{S}}_m(K)$ denotes a cluster of just-identified estimates. This needs to be translated to the *family* associated with the largest cluster, i.e. the set of IV-combinations, $\hat{\mathcal{F}}(K)$, used for the estimates that end up in the largest cluster.

$$\hat{\mathcal{F}}_m(K) = \{[j] : \hat{\boldsymbol{\beta}}_{[j]} \in \hat{\mathcal{S}}_m(K)\}$$

In the case with one regressor, each cluster is directly associated with a group of IVs. Now, the families need to be translated to sets of IVs to be tested. To achieve this, for each K, the potentially valid IVs are selected as those that are in

combinations contained in the largest family.

$$\hat{\mathcal{V}}_m(K) = \{j : [j] \in \hat{\mathcal{F}}(K)\}$$

The remaining IVs are then selected as invalid.

$$\hat{\mathcal{I}}(K) = \mathcal{J} \quad \setminus \quad \hat{\mathcal{V}}_m(K)$$

For each K, there might be cases where there are multiple maximal clusters $\hat{\mathcal{S}}_m(K)$, and then there are multiple associated $\hat{\mathcal{V}}_m(K)$. Let $\hat{\mathcal{V}}^M(K)$ denote the set of the multiple $\hat{\mathcal{V}}_m(K)$. In such a case, we select the cluster in which the most IVs are involved. If there are multiple clusters with the maximal number of estimates *and* of IVs, we select the set of IVs which leads to a lower Sargan test. Then for each K, the unique set of instruments to be checked by the Sargan test is:

$$\hat{\mathcal{V}}^{Sar}(K) = \{\hat{\mathcal{V}}_m(K) : \hat{\mathcal{V}}_m(K) = max|\hat{\mathcal{V}}^M(K)| \ \ \& \ \ min\, Sar(\hat{\mathcal{V}}^M(K))\} \tag{3.5}$$

The downward testing procedure considers the selection via $\hat{\mathcal{V}}^{Sar}(K)$, for each number of clusters $K \in \{1, ..., \binom{J}{P} - 1\}$, and chooses the smallest K such that the selected group of IVs passes the Sargan test:

$$\hat{\mathcal{V}}^{dts} = \{\hat{\mathcal{V}}^{Sar}(K),\ K = min(1, ..., \binom{J}{P} - 1) : Sar(\hat{\mathcal{V}}^{Sar}(K)) < \xi_{n,J-|\hat{\mathcal{I}}|-P}\} \tag{3.6}$$

The method has oracle properties as stated in Theorem 3.1 and Theorem 3.2. Here, we formally establish the theoretical results for the general case with an arbitrary number of regressors, $P \geq 1$. See Appendix 3.A.4 for proofs of all theorems. Suppose Algorithm 3.1 decides whether to merge two of the three clusters $\mathcal{S}_j$, $\mathcal{S}_k$ and $\mathcal{S}_l$, where all the IV combinations associated with the just-identified estimators in $\mathcal{S}_j$ and $\mathcal{S}_k$ are from the same true cluster $\mathcal{S}_{0q}$. For $\mathcal{S}_l$, however, it contains at least one estimator such that the corresponding IV combination is from a family other than $\mathcal{F}_q$. The following Lemma establishes asymptotically that Algorithm 3.1 merges $\mathcal{S}_j$ and $\mathcal{S}_k$.

Lemma 3.1. *Let $\mathcal{S}_j$ and $\mathcal{S}_k$ be two clusters such that any just-identified estimator*

$\hat{\boldsymbol{\beta}}_{[j]}$ that is contained in $\mathcal{S}_j$ and $\mathcal{S}_k$ satisfies $[j] \in \mathcal{F}_q$. Let $\mathcal{S}_l$ be a cluster such that $\exists \hat{\boldsymbol{\beta}}_{[l]} : \hat{\boldsymbol{\beta}}_{[l]} \in \mathcal{S}_l$ and $[l] \in \mathcal{F}_r$ with $r \neq q$. Under assumptions 3.1.a, 3.2, 3.3, 3.4, 3.5, 3.6.a in Algorithm 3.1, if merging two of $\mathcal{S}_j$, $\mathcal{S}_k$ and $\mathcal{S}_l$, then $\mathcal{S}_j$ and $\mathcal{S}_k$ are merged with probability converging to 1.

In Algorithm 3.1, we start from the number of clusters $K = \binom{J}{P}$. For each step onward, according to Step 3 in Algorithm 3.1, there would be two clusters merging with each other and forming a new cluster. Based on Lemma 3.1, along the path of Algorithm 3.1, members of different families will not be merged with each other until all the members from the same family have been merged into one family. If for each family, all the just-identified estimators associated with the IV combinations in the family have been merged into the same cluster, then we know that the total number of clusters is $K = Q$. This implies that when the number of clusters is smaller than Q, then at least one cluster contains estimators that use IV-combinations from different families. If the number of clusters is larger than Q, then the estimated families are subsets of a family.

Corollary 3.1. *Under assumptions 3.1.a, 3.2 to 3.5, and 3.6.a, in steps 3 and 4 of Algorithm 3.1:*

$$When \binom{J}{P} \geq K \geq Q, \quad \forall k : \quad lim\, P(\hat{\mathcal{F}}_k \subseteq \mathcal{F}_q) = 1$$

To better understand why this is the case, consider the following analogy. There are N guests ($\binom{J}{P}$ just-identified estimates) which belong to Q families. These N people live in a hotel, which has N rooms (clusters). Each day, one room disappears, and one of the people needs to move into the room of some other guest. The people in a family have closer ties, so the person whose room disappears will move into the room of somebody from their own family. This goes on until each family is living respectively in one crowded room. The hotel now continues to shrink. Only now are people from different families merged together into the same rooms. The largest family can be detected when all people from the same family have been merged into one room, but people from other families have not been merged into one room completely or have just been all merged into one room respectively).

In Algorithm 3.1, the number of clusters starts with $K = \binom{J}{P}$ and ends with $K = 1$. For each step in-between, the number of clusters decreases by 1, hence there must be a step where $K = Q$. Based on Lemma 3.1 and Corollary 3.1, estimators from different families are merged only when all elements of their own family have been completely merged to their clusters. This implies that when $K = Q$, there would be a cluster such that all the just-identified estimators in this cluster are estimated by all the valid instruments. Therefore, the path generated by Algorithm 3.1 contains the true family with probability going to 1 as there must be one step such that $K = Q$.

Corollary 3.2. *When* $K = Q$, $lim\, P(\hat{\mathcal{F}}_k = \mathcal{F}_q) = 1 \quad \forall k, q.$

The theoretical results above establish that the selection path generated by Algorithm 3.1 covers the family which uses only valid IVs, $\mathcal{F}_0$. In Appendix 3.A.4 we show that by Algorithm 3.2, we can locate this $\mathcal{F}_0$ and select the valid instruments consistently. This consistent selection property is summarized in Theorem 3.1 which holds for $P \geq 1$ under Assumption 3.1 (3.1.a) to Assumption 3.6 (3.6.a). These assumptions must also hold for Theorem 3.2 to hold.

3.4.2 The Weak Instruments Problem

In previous sections, we assumed that all the candidate instruments (or all the $\binom{J}{P}$ IV combinations when $P > 1$) are relevant for the endogenous variables by Assumption 3.1 and Assumption 3.1.a. However, in practice, these assumptions might not be satisfied in the sense that some of the candidate instruments are only weakly correlated with the endogenous variables. We now relax these assumptions and allow for individually weak instruments among the candidates. To be specific, we model the weak instruments as local to zero following Staiger and Stock (1997), i.e. an instrument Z_j is defined as weak if $\gamma_j = C/\sqrt{n}$, where C is a fixed scalar and $C \neq 0$. For consistent IV selection, we maintain the plurality assumption 3.6 for *strong and valid* instruments as in Guo et al. (2018): The group formed by all the strong and valid instruments is the largest group. Note that the largest group now also needs to be strong, while IVs in other groups can be weak.[2]

[2]The equivalent holds for the largest family when there are multiple regressors.

Inherently, our method can rule out weak and invalid instruments. This is because, for these instruments, under Model 3.1 and 3.2, it can be shown that their just-identified estimators tend to infinity.[3] Therefore, they can be separated from the just-identified estimators of the strong and valid instruments by the algorithm as the latter converge to the true value of the causal effect.

As for weak valid instruments, where the Wald ratio estimator is strongly biased, they are dropped from the set of selected valid instruments by the algorithm, because they do not pass the Sargan test, even if they cluster with the strong and valid IVs. Unlike the HT method, which uses first-stage hard thresholding and selects all weak valid instruments as invalid, our method is more flexible and instead uses the algorithm to decide which of the weak valid instruments that should be classified as invalid.

This mechanism has two advantages for valid weak instruments selection. First, compare with the HT method which drops all such instruments. Our method can avoid loss of information as the individually weak instruments can be informative all together. Second, it can limit the impact of including the selected weak instruments on IV estimation. By the algorithm, it can be seen that if the weak valid instruments are classified as valid, then this indicates that their just-identified estimators are not biased too much from the true value. Also, Windmeijer (2019) shows that the 2SLS estimator is a weighted average of all the just-identified estimates. The weights for each IV-specific estimate increase with the strength of each IV. By the plurality assumption, there are already strong valid instruments for post-selection IV estimation. In this case, the bias in the 2SLS estimator of including additional weak valid instruments would be small as their weights of contribution to the 2SLS estimator are small.

In comparison, the CIM can be problematic in the presence of weak instruments as it tends to select weak invalid instruments as valid, causing bias of the post-selection estimator. This is because the confidence intervals of the weak invalid instruments tend to have large ranges. Thus, most of them will be overlapping

[3]Consider $P = 1$. Let Z_j be a weak and invalid instrument, i.e. $\gamma_j = C/\sqrt{n}$ and $\alpha_j \neq 0$. Following Appendix A.5 in Windmeijer et al. (2021), for the just-identified estimator of Z_j, denoted by $\hat{\beta}_j$, we have $plim(\hat{\beta}_j) = plim(\beta_j) = plim(\beta + \frac{\alpha_j}{\gamma_j}) = \beta + plim(\sqrt{n}\frac{\alpha_j}{C})$ with $\alpha_j \neq 0$. Therefore $\hat{\beta}_j \to \infty$ as $n \to \infty$.

with all other confidence intervals, and the resulting largest group (which would be the selected set of valid instruments) will always contain some of the weak invalid instruments. As for the HT method, except for the disadvantage that there can be a potential loss of information by dropping all the weak valid instruments, it is also not clear how it chooses the optimal value of the threshold for any given sample, as noted in Windmeijer et al. (2021). In Section 3.5.2, we provide a detailed comparison via Monte Carlo experiments.

To summarize, our method can select all invalid instruments as invalid regardless of their strength, which is the key for consistent estimation of the causal effect. It treats weak valid instruments in a flexible way to avoid information loss and at the same time limits the bias-inducing effect of including weak instruments in the IV estimation.

3.4.3 Heterogeneous Treatment Effects

The instrumental variable estimator also has a local average treatment effect (LATE) interpretation, as it estimates the average treatment effect of a subpopulation, whose treatment can be changed by the instrument (Imbens and Angrist, 1994). Hence, LATEs will naturally vary with the instruments. For example, an increase in minimum school-leaving age versus proximity to school will see different populations increase their schooling. In this section, we show such a setting and argue that our method can retrieve the largest group associated with a given LATE, or the whole set of different LATEs.

For simplicity, we look at a setting with a binary treatment d_i, a binary instrument z_i, and potential outcomes y_{1i} and y_{0i}. The outcome and the treatment can be written as

$$y_i = y_{0i}(1 - d_i) + y_{1i}d_i$$
$$d_i = d_{0i}(1 - z_i) + d_{1i}z_i$$

Assumption 3.7. *Independence* $\{y_{0i}, y_{1i}, d_{0i}, d_{1i}\} \perp\!\!\!\perp z_i$

Assumption 3.8. *First Stage* $P(d_i = 1|z_i = 1) \neq P(d_i = 1|z_i = 0)$

Assumption 3.9. *Monotonicity* $d_{1i} > d_{0i}$

If the last three assumptions are fulfilled, Imbens and Angrist (1994) show that the IV estimand is the average treatment effect of compliers:

$$\frac{E(y_i|z_i = 1) - E(y_i|z_i = 0)}{E(d_i|z_i = 1) - E(d_i|z_i = 0)} = E(y_{1i} - y_{0i}|d_{1i} > d_{0i}) \tag{3.7}$$

In the following, we show a setting in which the LATEs are dependent on a potentially unobserved variable u. For this, we make use of the setting in Angrist and Fernandez-Val (2010). The treatment is determined by the following latent-index assignment mechanism

$$d_i = 1(h^z(u_i, z_i) > \eta_i) \tag{3.8}$$

where $h^z(u_i, 1) \geq h^z(u_i, 0)$ and the potential outcomes depend on the variable u:

$$y_{0i} = g_0(u_i) + \epsilon_{0i}$$
$$y_{1i} = g_1(u_i) + \epsilon_{1i}$$

where the errors are $E(\epsilon_i|u_i, z_i) = 0$. Angrist and Fernandez-Val (2010) then assume

Assumption 3.10. *Conditional Effect Ignorability:* $E(y_{1i} - y_{0i}|d^z_{1i}, d^z_{0i}, u_i) = E(y_{1i} - y_{0i}|u_i)$

Angrist and Fernandez-Val (2010) then show that under this assumption the LATE can be written as a function of u:

$$\beta_j = E(y_{1i} - y_{0i}|u, d_{1i} > d_{0i}) = g_1(u_i) - g_0(u_i) \tag{3.9}$$

We are interested in a setting where the by-IV treatment effects form groups:

$$\mathcal{G}_q = \{j : \beta_j = q\} \tag{3.10}$$

This might be the case, when different compliant populations have the same u or different u lead to the same β_j. Keep in mind that the number of groups is Q.

Note that Lemma 3.1 and Corollaries 3.1 and 3.2 also hold in the heterogeneous effects setting. In this case, the algorithm can find groups of heterogeneous treatment effects. Now, Algorithms 3.1 and 3.2 are altered. Instead of Steps 4 and 5 in Algorithm 3.1, which select the largest cluster and run post-selection 2SLS, we still do the downward testing procedure, but now do the Sargan-test for all clusters and stop at the step where none of the Sargan-tests reject. Finally, all cluster centers are reported.
In the same way as before:

Theorem 3.3. *Consistent selection of LATE groups*
Let ξ_n be the critical value for the Sargan test in Algorithm 3.2. Under Assumptions 3.7 - 3.10 and Lemma 3.1, for $\xi_n \to \infty$ and $\xi_n = o(n)$,

$$lim(\hat{\mathcal{G}}_k = \mathcal{G}_q) = 1 \quad \forall k, q.$$

The proof is in the Appendix. This theorem states that we can retrieve all heterogeneous treatment effect groups, when the heterogeneity is structured in groups. The difference to the setting with invalid IVs is that in the LATE-setting not only the largest cluster contains valuable information, but also the smaller clusters contain coefficient estimates obtained with valid instruments.

3.4.4 Different Proximity Measures

In Algorithm 3.1 we have made use of the Euclidean distance to assess the proximity of the clusters. One might be worried that the results are sensible to the choice of proximity measure. However, in practice this choice does not seem to play a big role.

Especially in settings with multiple regressors, there might be better choices to assess proximity. Aggarwal et al. (2001) discuss that the difference between the maximum and minimum distances to a given point becomes zero as the number of dimensions increases. This problem is exacerbated for higher-order norms, that is with $||\cdot||_k$-norms, where k is large. Therefore, in high dimensions, Aggarwal et al. (2001) suggest to rely on the Manhattan distance instead of the Euclidean distance. Going further than this, fractional norms of the shape $\sum_{d=1}^{D}[(x_1^d - x_2^d)^f]^{1/f}$

are introduced. It is shown that these fractional distance metrics preserve the contrast better than integral distance metrics.

Therefore, we also allow to use alternative distances in Algorithm 3.1. We consider the Manhattan and the Minkowski distance, which is similar to the fractional distance as proposed in Aggarwal et al. (2001), with the difference that the absolute value of the distances is taken.

Furthermore, Algorithm 3.1 computes the weighted Euclidean norm to evaluate the distance between clusters. The choice of linkage and distance definition is associated with a specific choice of the objective function, as discussed in Ward (1963). The latter aims to minimize the sum of within-cluster variation. In complete linkage, the two most distant elements of two clusters define the distance between the clusters. Alternative ways to assess proximity would be to use the medians or centroids of each cluster. We allow for alternative distance definitions and linkage methods in the R-package we provide.

In additional simulations, we considered these variants of the agglomerative hierarchical clustering algorithm, and the results are very similar to those obtained by using the Euclidean distance and the Ward-linkage function. The results of these simulations are available on request.

3.5 Monte Carlo Simulations

3.5.1 All Candidate Instruments are Strong

We conduct Monte Carlo simulations to illustrate the performance of our AHC method, and compare with that of the existing Confidence Interval Method and the Two-Stage Hard Thresholding Method in situations where Assumption 3.1 and Assumption 3.1.a are satisfied. In this set of simulations, we find that our method works as well as the CIM in terms of bias and it outperforms HT in small-sample settings. When there are multiple regressors, the summed bias is very close to the oracle bias and is only a fraction of the bias of the naive estimator.

We follow closely the setting in Windmeijer et al. (2021): There are 21 candidate instruments, 12 of which are invalid, while 9 are valid with $\boldsymbol{\alpha} = c_\alpha \left(\boldsymbol{\iota}_6', \ 0.5\boldsymbol{\iota}_6', \ \mathbf{0}_9'\right)'$ where $\mathbf{0}_r$ is an $r \times 1$ vector of zeros and $\boldsymbol{\iota}_r$ is an $r \times 1$ vector of ones. The first-stage

parameters are given by $\boldsymbol{\gamma} = c_\gamma \times \boldsymbol{\iota}_{21}$. We set $c_\alpha = 1$ and $c_\gamma = 0.4$. The true β is 0 and $\mathbf{z}_i \sim N(0, \boldsymbol{\Sigma}_{\boldsymbol{z}})$ with $\boldsymbol{\Sigma}_{\boldsymbol{z},\boldsymbol{jk}} = 0.5^{|j-k|}$. Errors are generated from

$$\begin{pmatrix} u_i \\ \varepsilon_i \end{pmatrix} \sim N\left(\mathbf{0}, \begin{pmatrix} 1 & 0.25 \\ 0.25 & 1 \end{pmatrix}\right).$$

The IV selection and estimation results are presented in Table 3.1 for sample sizes $N = 500,\ 1000, 2000$ for $1,000$ Monte Carlo replications. We report the median absolute error (column "MAE") and the standard deviation (column "SD") of the IV estimators, and the coverage rate of the 95% confidence intervals (column "*Coverage*"). For the IV selection results, we report three statistics: The number of selected invalid instruments (column "*# invalid*"), the frequency of selecting all invalid instruments as invalid (column "*p allinv*"), and the frequency of selecting the oracle model (column "*p oracle*").

For $N = 500$, the oracle 2SLS estimator (row "*oracle*"), which uses only the valid IVs and controls for the truly invalid ones, has the lowest MAE at 0.016, and the coverage rate of the 95 % confidence interval is 0.929. The naive 2SLS estimator (row "naive"), which treats all candidates instruments as valid irrespective of their validity, has a much larger median absolute error of about 1.056, and its 95 % confidence interval never covers the true value. As expected, even when increasing the sample size to 2000, this does not change. When using the HT method (row "HT") with 500 observations, the MAE is even larger than that of the naive 2SLS estimator, the method never chooses the oracle model, and none of the confidence intervals cover the true value. This is in line with the IV selection results; both the frequency of including all invalid instruments as invalid, and the frequency of selecting the oracle model, are 0. When using CIM (row "CIM"), the MAE is already low with sample size $N = 500$, the number of IVs chosen as invalid is close to 12, the frequency with which the oracle model is selected is at 0.966, and the coverage rate is 0.906. The results are very similar for our AHC method (row "AHC"). When increasing the sample size, the performance improves for all three selection methods. For CIM and AHC, the MAE is equal to that of the oracle estimator both at $N = 1000$ and $N = 2000$, and the probabilities of selecting the oracle model are close to one. For HT this probability is lower, which shows that

Table 3.1: Simulation Results with One Regressor

	MAE	SD	# invalid	p allinv	Coverage	p oracle
N=500						
oracle	0.016	0.025	12	1	0.929	1
naive	1.056	0.049	0	0	0	0
HT	1.165	0.127	12.696	0	0	0
CIM	0.017	0.267	12.023	0.987	0.906	0.966
AHC	0.016	0.179	12.049	0.989	0.912	0.983
N=1000						
oracle	0.012	0.017	12	1	0.953	1
naive	1.058	0.034	0	0	0	0
HT	1.374	0.114	18.205	0	0.001	0
CIM	0.012	0.017	12.015	1	0.948	0.986
AHC	0.012	0.135	12.052	0.991	0.936	0.980
N=2000						
oracle	0.008	0.012	12	1	0.943	1
naive	1.059	0.025	0	0	0	0
HT	0.010	0.384	12.679	0.885	0.864	0.708
CIM	0.008	0.012	12.013	1	0.938	0.988
AHC	0.008	0.160	12.039	0.993	0.931	0.984

This table reports median absolute error standard deviation, number of IVs selected as invalid, frequency with which all invalid IVs have been selected as invalid, coverage rate of the 95 % confidence interval and frequency with which oracle model has been selected. The true coefficient is $\beta = 0$. WLHB setting and invalid weaker setting are described in the text. 1000 repetitions per setting.

CIM and AHC have better finite sample performance.

We also inspect the performance of our method when there are multiple endogenous regressors. The existing selection methods do not allow for such an extension. Again, we draw 21 IVs with $\boldsymbol{\alpha} = c_\alpha \left(\boldsymbol{\iota}_6', \ 0.5\boldsymbol{\iota}_6', \ \mathbf{0}_9'\right)'$. The first-stage parameters are drawn from uniform distributions as $\boldsymbol{\gamma}_1 = unif(1,2)$, $\boldsymbol{\gamma}_2 = unif(3,4)$, and, when there is a third endogenous regressor, $\boldsymbol{\gamma}_3 = unif(5,6)$. The rest of the parameters are the same as the earlier simulation design. In this setting, we estimate $\boldsymbol{\beta} = \mathbf{0}$ for $m = 1,000$ replications. We report the results in Table 3.2. It can be seen that the performance of our method approaches that of the oracle estimator as the sample size grows large. Although, when the number of endogenous variables

Table 3.2: Simulation Results with More Than One Regressor

	MAE	SD	# invalid	p allinv	Coverage	p oracle
P=2						
N=500						
Oracle	0.049	0.085	12	1	0.965	1
Naive	0.597	0.377	0	0	0.032	0
AC	0.080	0.583	12.215	0.930	0.879	0.750
N=1000						
Oracle	0.044	0.068	12	1	0.952	1
Naive	0.658	0.272	0	0	0	0
AC	0.055	0.343	12.202	0.982	0.919	0.827
N=5000						
Oracle	0.021	0.033	12	1	0.949	1
Naive	0.755	0.138	0	0	0	0
AC	0.024	0.037	12.109	1	0.938	0.909
P=3						
N=500						
Oracle	0.063	0.099	12	1	0.952	1
Naive	0.880	0.372	0	0	0.002	0
AC	0.121	0.804	12.190	0.794	0.725	0.520
N=1000						
Oracle	0.050	0.078	12	1	0.934	1
Naive	0.915	0.279	0	0	0	0
AC	0.073	0.416	12.367	0.948	0.844	0.696
N=5000						
Oracle	0.037	0.058	12	1	0.919	1
Naive	0.941	0.211	0	0	0	0
AC	0.049	0.307	12.261	0.976	0.853	0.797

This table reports median absolute error, standard deviation, number of IVs selected as invalid, frequency with which all invalid IVs have been selected as invalid, coverage rate of the 95 % confidence interval and frequency with which oracle model has been selected. For the first two, means over the statistic for each regressor are taken. The true coefficient is $\boldsymbol{\beta} = \mathbf{0}$. Settings are described in the text. 1000 repetitions per setting.

increases from 1 to 3, it needs a larger sample size to achieve oracle selection.

3.5.2 Some Weak Instruments Among the Candidate Instruments

Next, we investigate the performance of the previously mentioned methods when Assumption 3.1 and Assumption 3.1.a are violated, i.e there are weak instruments among the candidates. Overall, we find that our AHC method outperforms the CIM, and in the case where the largest group does not consist of strong and valid IVs, also the HT. Moreover, with two endogenous regressors, AHC is still very close to oracle performance.

For individually weak instruments, we consider the local to zero setup and we set their first stage parameters as $\gamma_j = C/\sqrt{n}$ with $C = 0.1$. First, consider the same setting as in Section 3.5.1 with one endogenous variable, but with the following variations:

- Design 1: All the 12 invalid instruments are irrelevant, and all the 9 valid instruments are relevant: $\boldsymbol{\gamma} = c_\gamma \left(\boldsymbol{\iota}_{12}' C/\sqrt{n},\ \ \boldsymbol{\iota}_9'\right)'$.
- Design 2: All the 12 invalid instruments are irrelevant, and almost half of the valid instruments are irrelevant (4 out of 9): $\gamma = c_\gamma \left(\boldsymbol{\iota}_{16}' C/\sqrt{n},\ \ \boldsymbol{\iota}_5'\right)'$.
- Design 3: Both the valid and invalid instruments are mixtures of irrelevant and relevant instruments.
  - a). Relevant and valid instruments still form the largest group: $\boldsymbol{\gamma} = c_\gamma \left(\boldsymbol{\iota}_6',\ \ \boldsymbol{\iota}_7' C/\sqrt{n},\ \ \boldsymbol{\iota}_8'\right)'$.
  - b). Relevant and valid instruments do not form the (strictly) largest group: $\boldsymbol{\gamma} = c_\gamma \left(\boldsymbol{\iota}_6',\ \ \boldsymbol{\iota}_9' C/\sqrt{n},\ \ \boldsymbol{\iota}_6'\right)'$.

All the other parameters are the same as in Section 3.5.1. We focus on the large sample performance in the presence of weak instruments, and we fix the sample size to $N = 2000$. The simulation results are calculated based on $1,000$ Monte Carlo replications. We present the results in Table 3.3, where MAE, *# invalid*

and *p allinv* are defined in the same way as in Section 3.5.1. Here we also report three new IV selection statistics: the frequency of selecting all valid and strong instruments as valid (column "*strongvalid*"), the frequency of selecting all weak invalid instruments as invalid (column "*weakin*"), and the frequency of selecting all weak valid instruments as invalid (column "*weakva*"). In these designs, we let the oracle models include only the strong and valid instruments as valid. Our primary focus is the selection of the invalid instruments. It is crucial that all the invalid instruments (whether strong or weak) are selected as invalid, because including *any* invalid instruments in IV estimation can cause severe bias.

In Table 3.3, we can see that, in the presence of weak instruments, the CI method can be problematic; the frequencies of selecting all invalid instruments as invalid are low in all settings (lowest at 0.024 in Design 1 and highest at 0.351 in Design 3a). This means that the CI method almost always includes invalid instruments as valid. Consequently, the MAE of the post-selection estimator is very large (and much larger than those of the oracle, and the HT and AHC methods).

The HT method performs well in almost all designs. It selects all weak instruments (both valid and invalid) as invalid with probability almost equal to 1. Also, it has high frequencies of selecting all strong and valid instruments as valid. It can be seen that if the strong and valid instruments form the largest group, then the voting mechanism of the HT method can select the oracle model.

In line with the selection performance, the MAEs of HT are identical to those of the oracle models. In Design 3b, however, the plurality rule does not hold anymore; there is a tie between the group of strong and valid instruments, and strong and invalid instruments. In this situation, the voting mechanism does not perform well, and we see that *p allinv* is only at 0.053. This results in a significantly larger MAE than the oracle model.

In general, the AHC performs well and it has similar MAE as the oracle model in all settings. For Design 1, 2 and 3a, it guarantees that all the invalid instruments are selected as invalid with *p allinv* and *weakin* close to 1. In terms of valid instruments, all the strong valid instruments are included as valid with high frequencies (*strongvalid* close to 1). For weak valid instruments, some of them are selected as valid. This is because the just-identified estimators of the weak valid instruments may not be too far away from those of the strong and valid instruments, and,

Table 3.3: Some Weak Instruments with One Regressor

	MAE	# invalid	p allinv	strongvalid	weakin	weakva
Design 1						
oracle	0.008	12	1	1	1	-
HT	0.008	12.000	1	1	1	-
CIM	35.112	13.289	0.024	0	0.024	-
AHC	0.008	12.028	1	0.988	1	-
Design 2						
oracle	0.013	16	1	1	1	1
HT	0.013	15.951	1	1	1	0.952
CIM	33.646	12.806	0.027	0	0.027	0.527
AHC	0.012	12.445	0.999	0.997	0.999	0.002
Design 3a						
oracle	0.008	13	1	1	1	1
HT	0.008	13.164	1	0.842	1	0.984
CIM	14.497	16.772	0.351	0.002	0.467	0.691
AHC	0.008	12.323	0.998	0.992	1	0.306
Design 3b						
oracle	0.011	15	1	1	1	1
HT	0.929	10.511	0.053	0.870	0.999	0.961
CIM	13.636	16.500	0.277	0.008	0.462	0.421
AHC	0.013	12.766	0.847	0.847	1	0.002

This table reports median absolute error, number of IVs selected as invalid, frequency of all invalid IVs selected as invalid, frequency of all valid and strong instruments selected as valid, frequency of all weak invalid instruments selected as invalid, and frequency of all weak valid instruments as invalid. 1000 repetitions per setting.

Design 1				Design 2				Design 3			
IV	γ_1	γ_2	α	IV	γ_1	γ_2	α	IV	γ_1	γ_2	α
$\mathbf{z}_1$	1	$C/\sqrt{n}$	0	$\mathbf{z}_1$	1	$C/\sqrt{n}$	1	$\mathbf{z}_1$	1	$C/\sqrt{n}$	0
$\mathbf{z}_2$	2	$C/\sqrt{n}$	0	$\mathbf{z}_2$	2	$C/\sqrt{n}$	1	$\mathbf{z}_2$	2	$C/\sqrt{n}$	0
$\mathbf{z}_3$	3	$C/\sqrt{n}$	0	$\mathbf{z}_3$	3	$C/\sqrt{n}$	1	$\mathbf{z}_3$	3	$C/\sqrt{n}$	1
$\mathbf{z}_4$	4	$C/\sqrt{n}$	0	$\mathbf{z}_4$	4	$C/\sqrt{n}$	0	$\mathbf{z}_4$	$C/\sqrt{n}$	$C/\sqrt{n}$	1
$\mathbf{z}_5$	$C/\sqrt{n}$	$unif(1,2)$	0	$\mathbf{z}_5$	$C/\sqrt{n}$	$unif(1,2)$	0	$\mathbf{z}_5$	$C/\sqrt{n}$	$C/\sqrt{n}$	1
$\mathbf{z}_6$	$C/\sqrt{n}$	$unif(1,2)$	0	$\mathbf{z}_6$	$C/\sqrt{n}$	$unif(1,2)$	0	$\mathbf{z}_6$	$C/\sqrt{n}$	$C/\sqrt{n}$	0
$\mathbf{z}_7$	$C/\sqrt{n}$	$unif(1,2)$	0	$\mathbf{z}_7$	$C/\sqrt{n}$	$unif(1,2)$	0	$\mathbf{z}_7$	$C/\sqrt{n}$	$unif(3,4)$	1
$\mathbf{z}_8$	$C/\sqrt{n}$	$unif(1,2)$	0	$\mathbf{z}_8$	$C/\sqrt{n}$	$unif(1,2)$	1	$\mathbf{z}_8$	$C/\sqrt{n}$	$unif(3,4)$	0
$\mathbf{z}_9$	$C/\sqrt{n}$	$unif(1,2)$	0	$\mathbf{z}_9$	$C/\sqrt{n}$	$unif(1,2)$	1	$\mathbf{z}_9$	$C/\sqrt{n}$	$unif(3,4)$	0

Table 3.4: Weak IV Simulation Designs with Two Endogenous Regressors

thus, in some cases they are not totally separated by the algorithm. This is not a major concern as, for weak valid instruments, the algorithm would only keep those whose Wald ratio estimators are not severely distorted. Hence, the effect of the selected weak instruments on the resulting post-selection IV estimator is limited (the MAEs of AHC are very close to those of the oracle models). It is notable, that in Design 3b where there are two largest groups, AHC outperforms HT with a frequency of 0.847 of including all the invalid instruments as invalid. Moreover, AHC can, alternatively, report both groups.

We also investigate the large-sample performance of AHC in the presence of weak IVs with two endogenous variables, and we fix the sample size at $N = 5000$. The simulations are conducted in four designs with 9 candidate instruments, see Table 3.4. In Design 1, each instrument is valid but only strong for one endogenous variable, respectively, which violates Assumption 3.1.a. We are interested to see if the AHC method can include all the instruments as valid. In Design 2, all the candidate instruments are strong for only one treatment variable, but some of them are invalid. In Design 3, we make some of the instruments weak for both variables, and use a mix of valid and invalid instruments. Table 3.5 presents the results. In all designs, AHC achieves selection results close to the oracle model, and hence also has similar MAEs. This shows that, even in settings where the usual 2SLS estimator would fail because the first-stage coefficient matrix is near rank-reduced, we can still obtain useful estimates. This is because some of the

Table 3.5: Some Weak Instruments with Two Endogenous Regressors

	MAE	# invalid	p allinv	strongvalid	weakin	weakva
Design 1						
oracle	0.003	0	1	1	-	-
AHC	0.003	0.018	1	0.991	-	-
Design 2						
oracle	0.006	5	1	1	-	-
AHC	0.006	5.006	0.867	0.867	-	-
Design 3						
oracle	0.007	5	1	1	1	1
AHC	0.007	4.215	0.929	0.904	0.997	0.122

This table reports median absolute error, number of IVs selected as invalid, frequency of all invalid IVs selected as invalid, frequency of all valid and strong instruments selected as valid, frequency of all weak invalid instruments selected as invalid, and frequency of all weak valid instruments as invalid. 1000 repetitions per setting.

just-identified estimates use combinations of IVs that are strong, which can provide sufficient information for selecting valid instruments, and, thus, also for delivering consistent estimates.

3.6 Application: Effect of Immigration on Wages

In this section, we apply our method to the estimation of the effects of immigration on wages in the US. We first describe the setting and then discuss the results.

Many recent studies have tried to estimate the causal effects of immigration on labor market outcomes.[4] Most papers in the literature only estimate the contemporaneous effects of immigration on labor market outcomes. Jaeger et al. (2020) point out that there might be long-term adjustments that affect wages in the long run, for example, because local workers and firms react to the inflow of migrants in the long-term. This calls for including lagged immigration into the regression equation.

[4] An overview of the literature can be found in Dustmann et al. (2016).

To illustrate our method, we estimate the following linear model:

$$\Delta y_{lt} = \beta^{short} \Delta immi_{l,t} + \beta^{long} \Delta immi_{l,t-10} + \psi_t + \varepsilon_{lt}, \tag{3.11}$$

as in Basso and Peri (2015).

There are three years $t \in \{1990, 2000, 2010\}$ and 722 commuting zones l. The dependent variable Δy_{lt} is the change in log weekly wages of high-skilled workers. The independent variables are $\Delta immi_{l,t}$, denoting the *current* change of immigrants in employment, and $\Delta immi_{l,t-10}$, denoting the same change ten years ago (note the lagged time subscript). The coefficients of interest are the short-term (contemporaneous) effect β^{short} and the long-term effect β^{long}. Decade fixed-effects are captured by ψ_t, and ε_{lt} is the error term. Commuting-zone fixed effects are eliminated through first-differencing as is standard with panel data (see e.g. Wooldridge, 2010, p. 315). This regression is canonical in migration economics. We use data from the Census Integrated Public Use Micro Samples (IPUMS) and the American Community Survey (Ruggles et al., 2015).

The key econometric challenge is that migrants select where to live endogenously. For example, migrants might choose where to live based on economic conditions in a region. This creates a bias in the estimates. A much-used estimation strategy to address this issue is to use historical settlement patterns of migrants from many countries of origin as instruments. When earlier migrants attract migrants at later points in time, the instruments are relevant. This identification strategy dates back to Altonji and Card (1991). The papers that use this type of instrument in this context are numerous (Jaeger et al., 2020).

We use all shares of foreign-born people (we call them migrants, analogously) in working age from a certain origin country j at a base period t_0 in region l. The share is measured relative to their total number in the country and is denoted by s_{jlt_0}. We use origin-specific shares from 19 origin country groups and base years 1970 and 1980 as separate IVs and obtain $L = 38$ IVs. It is usually expected that the reasons that attracted migrants in the past are quasi-random as compared with current migration. Validity is typically defended on these grounds.
However, these previous settlement patterns might be invalid. Jaeger et al. (2020) show that IV estimators that rely on this kind of exclusion restriction might be

inconsistent, first, because of correlation of the IVs with unobserved demand shocks and, second, because of dynamic adjustment processes. Hence, none of these two should play a role. However, it is plausible that some origin country groups did not locate randomly in the past or have had direct effects on the wages. The second challenge can be somewhat tackled by including lagged immigration as an additional regressor. Of course, this will also be subject to the same endogeneity problem as before, and, thus, should also be instrumented. To circumvent these problems, we apply our estimator, which allows for direct effects of many migrant settlement variables on wages by pre-selecting the valid instruments.

This approach is canonical and is also highly relevant in the current applied economic literature. In a recent paper, Goldsmith-Pinkham et al. (2020) discuss a class of IVs which are extensively used in labor economics.[5] A sufficient condition for this type of IVs to be valid is that all shares are valid. Therefore, the selection method proposed here can also be used to improve the construction of this class of instruments, as shown in Apfel (2021).

Results The results can be found in Table 3.6. The first column shows results for ordinary least squares: The contemporaneous effect is 0.586, while the lagged effect is lower and negative. When using all shares as valid IVs, both effects are higher in absolute terms but only the contemporaneous effect is marginally statistically significant. The Hansen-Sargan test for this model gives a p-value of 0.0126, which is lower than the proposed significance level of $0.1/log(n)$.

When using AHC with this significance level in the downward testing procedure, two origin country shares are selected as invalid: The shares of Mexicans in the US in 1970 and 1980. The coefficient estimates of the short- and long-term effects increase considerably in absolute terms. Now, both coefficient estimates are clearly statistically significant. This indicates that the use of AHC indeed makes a big difference. Moreover, the p-value of the Sargan test is pushed over the threshold of 0.013, used in the testing procedure.

The two IVs that are selected are similar a priori in that they are shares from the same origin country. These shares are likely to be invalid, because Mexican

[5]These so-called shift-share IVs combine the previous settlement shares, which we use in this application, with aggregate-level shocks, so-called shifts.

Table 3.6: Impact of Immigration on High-Skilled Wages

	OLS	2SLS	2SLS AHC
$\Delta immi_{lt}$	0.586	0.877	1.522
	(0.0935)	(0.460)	(0.292)
$\Delta immi_{lt-10}$	-0.197	-0.249	-0.771
	(0.0814)	(0.321)	(0.246)
Nr inv		0	2
P-value		.0126	.0447

N = 2166 (722 CZ × 3), L = 38. Standard errors in parentheses. Observations weighted by beginning-of-period population. Significance level in testing procedure: 0.013. "Nr inv" stands for the number of IVs selected as invalid.

migrants were attracted to border regions, such as Texas and California, by the good economic conditions in those states, both in the base year and in later periods. California's economy has a large agricultural sector, and both states are among the wealthiest in the US. It is therefore likely that wages, or unobserved productivity shocks that have driven the initial settlement, are correlated over time, thereby invalidating the initial shares. Moreover, in their application Goldsmith-Pinkham et al. (2020) find that Mexico has the highest sensitivity-to-misspecification weight, that is, the overall bias will be most sensible to any invalidity stemming from the Mexican share. This application has shown that our method can make a substantial difference in practical terms, because it can help researchers identify IVs which violate the exclusion restriction.

3.7 Conclusion

We have proposed a novel method to select valid instruments. This method can be particularly helpful in cases where there are many candidate instruments and the tests of overidentifying restrictions reject. The method is applied to the estimation of the effect of immigration on wages in the US. The method can also be easily applied to any other setting in which there are many candidate instruments. Another suitable example is Mendelian Randomization, which is the use

of instrumental variables in epidemiology.

The advantages of our method are that it extends to the setting with multiple endogenous regressors, and that it can also deal effectively with the problem of weak instruments. In fact, one might also use our method directly to select strong IVs. We also discuss a setting with heterogeneous treatment effects. It would be worth investigating how to retrieve causal effects when there are richer forms of heterogeneity. Another way to improve the method would be to account for the variance of each just-identified estimator in the selection algorithm, and to apply it in nonlinear models. We leave these as directions for future research.

3.A Appendix

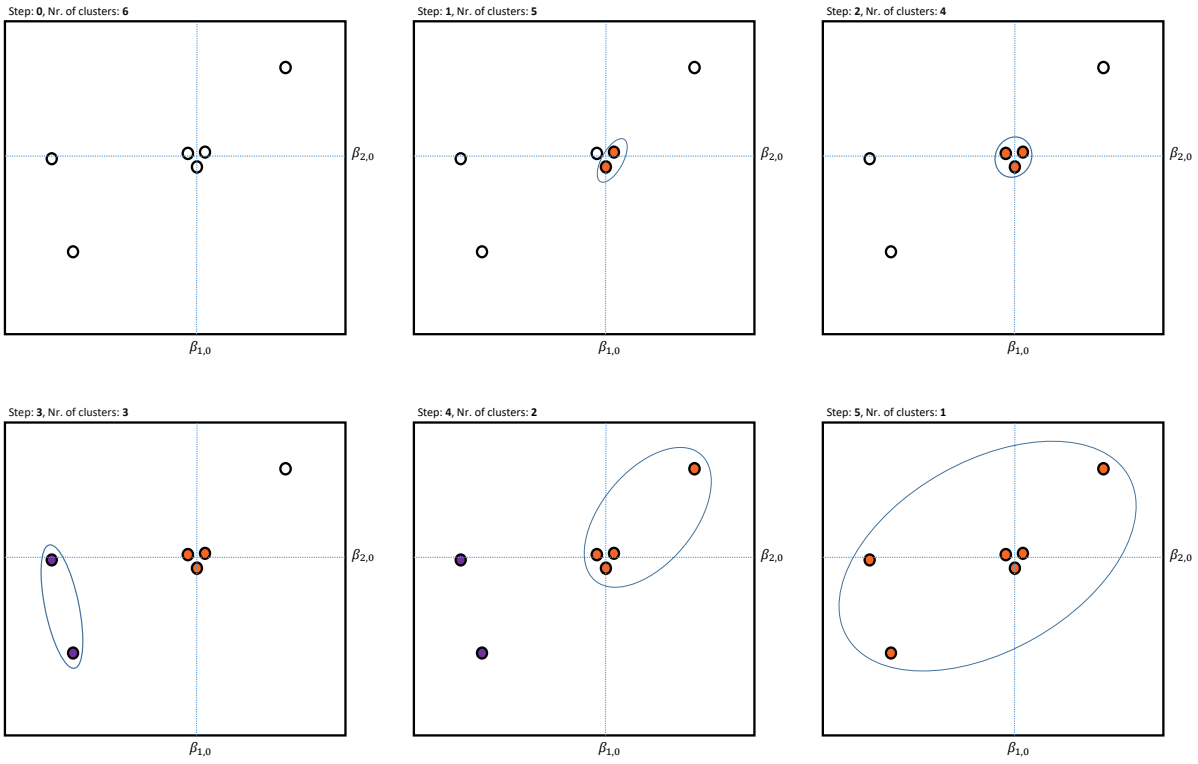
3.A.1 Illustration of the IV Selection Procedure for $P = 2$

In figure 3.A.1, the procedure is illustrated. Here, we have a situation with four IVs and two endogenous regressors. Instrument No. 1 is invalid, because it is directly correlated with the outcome, while the remaining three IVs (2, 3, 4) are related with the outcome only through the endogenous regressors and are hence valid.

In the first graph on the top left, we have plotted each just-identified estimate. The horizontal and vertical axes represent coefficient estimates of the effects of the first (β_1) and second regressor (β_2), respectively. Each point has been estimated with two IVs, in this case with IV pairs 1-2, 1-3, 1-4, 2-3, 2-4 and 3-4, because there are four candidate IVs.

In the initial Step (0), each just-identified estimate has its own cluster. In step 1, we join the estimates which are closest in terms of their Euclidean distance, e.g. those estimated with pairs 2-3 and 2-4. These two estimates now form one cluster and we only have five clusters. We re-estimate the distances to this new cluster and continue with this procedure, until there is only one cluster left in the bottom right graph. We evaluate the Sargan test at each step, using the IVs which are involved in the estimation of the largest group at each step. When the p-value is larger than a certain threshold, say 0.05, we stop the procedure. Ideally this will be the case at step 2 or 3 of the algorithm, because here the largest cluster (in orange) is formed only by valid IVs (2,3 and 4). If this is the case, only the valid IVs are selected as valid.

Figure 3.A.1: Illustration of the Algorithm with Two Regressors



3.A.2 Properties of just-identified estimates when $P \geq 1$

There are $\binom{J}{P}$ just-identified models. We write the corresponding just-identified estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ analogously to the proof of Proposition A.5 in Windmeijer et al. (2021) for the case $P = 1$. First, for an arbitrary $[j]$, partition the matrix $\mathbf{Z} = (\mathbf{Z}_1 \quad \mathbf{Z}_2)$, where $\mathbf{Z}_1$ is a $n \times P$ matrix containing the $[j]$-th columns of $\mathbf{Z}$, and $\mathbf{Z}_2$ is a $n \times (J - P)$ matrix containing the remaining columns of $\mathbf{Z}$. $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1' \quad \boldsymbol{\gamma}_2')'$ is the equivalent partition of the matrix of first-stage coefficients. $\mathbf{Z}^* = [\hat{\mathbf{D}} \quad \mathbf{Z}_2]$, then $\mathbf{Z}^* = \mathbf{Z}\hat{\mathbf{H}}$, with

$$\hat{\boldsymbol{H}} = \begin{pmatrix} \hat{\boldsymbol{\gamma}}_1 & 0 \\ \hat{\boldsymbol{\gamma}}_2 & \mathbf{I}_{J-P} \end{pmatrix}; \quad \hat{\boldsymbol{H}}^{-\mathbf{1}} = \begin{pmatrix} \hat{\boldsymbol{\gamma}}_1^{-1} & 0 \\ -\hat{\boldsymbol{\gamma}}_2\hat{\boldsymbol{\gamma}}_1^{-1} & \mathbf{I}_{J-P} \end{pmatrix}$$

The just-identified 2SLS estimators using $\mathbf{Z}_{[j]}$ as instruments and controlling for the remaining instruments can be written as

$$(\hat{\boldsymbol{\beta}}_{[j]} \quad \hat{\boldsymbol{\alpha}}_{[j]}')' = (\mathbf{Z}^{*\prime}\mathbf{Z}^*)^{-1}\mathbf{Z}^{*\prime}\mathbf{y} = \hat{\mathbf{H}}^{-1}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} = \hat{\mathbf{H}}^{-1}\hat{\boldsymbol{\Gamma}}$$

Then we have $\hat{\boldsymbol{\beta}}_{[j]} = \hat{\boldsymbol{\gamma}}_1^{-1}\hat{\boldsymbol{\Gamma}}_1$. It follows from $\boldsymbol{\Gamma} = \boldsymbol{\gamma}\boldsymbol{\beta} + \boldsymbol{\alpha}$ that $\boldsymbol{\gamma}_1^{-1}\boldsymbol{\Gamma}_1 = \boldsymbol{\beta} + \boldsymbol{\gamma}_1^{-1}\boldsymbol{\alpha}_1$. Therefore,

$$plim(\hat{\boldsymbol{\beta}}_{[j]}) = plim(\hat{\boldsymbol{\gamma}}_1^{-1}\hat{\boldsymbol{\Gamma}}_1) = \boldsymbol{\gamma}_1^{-1}\boldsymbol{\Gamma}_1 = \boldsymbol{\beta} + \boldsymbol{\gamma}_1^{-1}\boldsymbol{\alpha}_1$$

We denote the $\binom{J}{P}$ $P \times 1$-dimensional inconsistency terms as $plim(\hat{\boldsymbol{\beta}}_{[j]} - \boldsymbol{\beta}) = \boldsymbol{\gamma}_{[j]}^{-1}\boldsymbol{\alpha}_{[j]} = \mathbf{q}$.

3.A.3 $\mathcal{F}_0$ consists of valid IVs only

Next, we show that the family with $\mathbf{q} = \mathbf{0}$ is composed of valid IVs with $\boldsymbol{\alpha}_1 = \mathbf{0}$, only. Let $\boldsymbol{\gamma}$, $\mathbf{Z}$ and $\boldsymbol{\alpha}$ be partitioned the same way as in Appendix 3.A.2.

Remark 3.1. $\boldsymbol{\alpha}_1 = \mathbf{0}$ *is necessary and sufficient for* $\mathbf{q} = \mathbf{0}$.

Proof: First prove sufficiency: Direct proof: Assume $\boldsymbol{\alpha}_1 = \mathbf{0}$ holds. $\mathbf{q} = \boldsymbol{\gamma}_1^{-1}\boldsymbol{\alpha}_1 = \mathbf{0}$ follows directly.
Second, prove necessity: Proof by contraposition: Assume $\boldsymbol{\alpha}_1 \neq \mathbf{0}$, then $\boldsymbol{\gamma}_1^{-1}\boldsymbol{\alpha}_1 \neq \mathbf{0}$. The latter inequality holds, because otherwise the columns of $\boldsymbol{\gamma}_1^{-1}$ are linearly dependent, and $\boldsymbol{\gamma}_1^{-1}$ is not invertible and hence $(\boldsymbol{\gamma}_1^{-1})^{-1} = \boldsymbol{\gamma}_1$ does not exist, which it clearly does, by Assumption 1.a. □
This also implies that $\mathcal{F}_0$ consists of valid IVs only and all combinations $[j]$: $\boldsymbol{\gamma}_1^{-1}\boldsymbol{\alpha}_1 = \mathbf{0}$ are elements of $\mathcal{F}_0$. Hence, the following remark directly follows:

Remark 3.2. $|\mathcal{F}_0| = \binom{g}{P}$.

3.A.4 Proofs of the Oracle Properties

This section gives proofs for Lemma 3.1 and Theorems 3.1 and 3.3. All proofs apply for the general case that $P \geq 1$.

Proof of Lemma 3.1

The proof is structured as follows:

1. We note that the means of clusters which are formed by members from the same family converge to the same value as each estimator does in the cluster.

2. Merging two clusters which are associated only with elements from the same family is equivalent to the two clusters having minimal distance.

3. We show that clusters which are associated with members from the same family have distance zero and clusters which are associated with elements from different families have non-zero distance, with probability going to one.

Proof. Part 1: Consider

$$\begin{aligned} [j], [k] \in \mathcal{F}_q\,, \quad \mathbf{q} \in \mathbb{R}^P \\ [l] \in \mathcal{F}_r\,, \quad \mathbf{r} \in \mathbb{R}^P, \quad \mathbf{r} \neq \mathbf{q} \end{aligned}$$

Under Assumptions 3.1 (3.1.a) - 3.5:

$$\begin{aligned} plim(\hat{\boldsymbol{\beta}}_{[j]}) = plim(\hat{\boldsymbol{\beta}}_{[k]}) = \mathbf{q} \\ plim(\hat{\boldsymbol{\beta}}_{[l]}) = \mathbf{r} \end{aligned} \tag{3.A.1}$$

Let $\mathcal{S}_j$ and $\mathcal{S}_k$ be clusters associated with elements from the same family: $\mathcal{S}_j$, $\mathcal{S}_k \subset \mathcal{S}_{0q}$ and $\mathcal{S}_l \subset \mathcal{S}_{0r}$.

$$plim\ \bar{\mathcal{S}}_j = \frac{\sum\limits_{\hat{\beta}_{[j]} \in \mathcal{S}_j} \hat{\boldsymbol{\beta}}_{[j]}}{|\mathcal{S}_j|} = \frac{|\mathcal{S}_j|\mathbf{q}}{|\mathcal{S}_j|} \text{ where } \mathcal{S}_j \subset \mathcal{S}_{0q} \tag{3.A.2}$$

and hence

$$plim(\bar{\mathcal{S}}_j) = \mathbf{q}.$$

Part 2: Consider the case where the Algorithm decides whether to merge two clusters, $\mathcal{S}_j$ and $\mathcal{S}_k$, containing estimators using combinations from the same family, or to merge two clusters from different underlying partitions, $\mathcal{S}_j$ and $\mathcal{S}_l$. The two clusters which are closest in terms of their weighted Euclidean distance are merged first. Hence, we need to consider the distances between $\mathcal{S}_j$ and $\mathcal{S}_k$, $\mathcal{S}_j$ and $\mathcal{S}_l$, as well as $\mathcal{S}_k$ and $\mathcal{S}_l$.

$\mathcal{S}_j$ is merged with a cluster with elements of its own $\mathcal{S}_{0q}$ iff $\frac{|\mathcal{S}_j||\mathcal{S}_k|}{|\mathcal{S}_j|+|\mathcal{S}_k|}||\bar{\mathcal{S}}_j - \bar{\mathcal{S}}_k||^2 < \frac{|\mathcal{S}_j||\mathcal{S}_l|}{|\mathcal{S}_j|+|\mathcal{S}_l|}||\bar{\mathcal{S}}_j - \bar{\mathcal{S}}_l||^2$. The following two are hence equivalent

$$lim\, P(\mathcal{S}_j \cup \mathcal{S}_k = \mathcal{S}_{jk} \subseteq \mathcal{S}_{0q}) = 1$$

$$\Leftrightarrow \quad lim\, P(\frac{|\mathcal{S}_j||\mathcal{S}_k|}{|\mathcal{S}_j| + |\mathcal{S}_k|}||\bar{\mathcal{S}}_j - \bar{\mathcal{S}}_k||^2 < \frac{|\mathcal{S}_j||\mathcal{S}_l|}{|\mathcal{S}_j| + |\mathcal{S}_l|}||\bar{\mathcal{S}}_j - \bar{\mathcal{S}}_l||^2) = 1 \qquad (3.A.3)$$

where $\mathcal{S}_{jk}$ is the new merged cluster.

Part 3: We want to prove equation (3.A.3) in the following. We can then prove $lim\, P(\frac{|\mathcal{S}_j||\mathcal{S}_k|}{|\mathcal{S}_j|+|\mathcal{S}_k|}||\bar{\mathcal{S}}_k - \bar{\mathcal{S}}_j||^2 < \frac{|\mathcal{S}_k||\mathcal{S}_l|}{|\mathcal{S}_k|+|\mathcal{S}_l|}||\bar{\mathcal{S}}_k - \bar{\mathcal{S}}_l||^2) = 1$ by changing the subscripts.

First, define $a = \frac{|\mathcal{S}_j||\mathcal{S}_k|}{|\mathcal{S}_j|+|\mathcal{S}_k|}||\bar{\mathcal{S}}_j - \bar{\mathcal{S}}_k||^2$, $b = \frac{|\mathcal{S}_j||\mathcal{S}_l|}{|\mathcal{S}_j|+|\mathcal{S}_l|}||\bar{\mathcal{S}}_j - \bar{\mathcal{S}}_l||^2$ and $c = \frac{|\mathcal{S}_j||\mathcal{S}_l|}{|\mathcal{S}_j|+|\mathcal{S}_l|}(\mathbf{q} - \mathbf{r})'(\mathbf{q} - \mathbf{r})$.

Under (3.A.2)

$$plim(a) = \mathbf{0}$$

$$plim(b) = c$$

To show: $\lim_{n\to\infty} P(a < b) = 1$.

Proof by contradiction: Show that $\lim_{n\to\infty} P(b < a) \neq 0$ leads to a contradiction. Let lim imply $\lim_{n\to\infty}$ in the following. By the definitions of convergence in probability, it follows that

$$\lim\, P(a < \varepsilon) = 1 \qquad (3.A.4)$$

and

$$\lim\, P(|b - c| < \varepsilon) = 1. \qquad (3.A.5)$$

for any ε. Therefore, lim $P(a < b) \neq 0$ and lim $P(a < \varepsilon) = 1$ imply lim $P(b < \varepsilon) \neq 0$.

Now, consider $\varepsilon < \frac{1}{2}c$.

Then,

$$\lim P(b < \frac{1}{2}c) \neq 0 \tag{3.A.6}$$

Because of the absolute value $b - c$, consider two cases, $b < c$ and $b > c$. If $b < c$: $\lim P(c - b < \frac{1}{2}c) = 1 \Leftrightarrow \lim P(c - b > \frac{1}{2}c) = 0. \Rightarrow \lim P(b < \frac{1}{2}c) = 0$, a contradiction with (3.A.6). If $b \geq c$: $a < \varepsilon < \frac{1}{2}c < c \leq b$ and hence $\lim P(a < b) = 1 \Leftrightarrow \lim P(b \leq a) = 0$, again a contradiction. □

□

Proof of Theorem 3.1

Proof. The proof for Theorem 3.1 is structured as follows:

1. We show that asymptotically the selection path generated by Algorithm 3.1 contains $\mathcal{F}_0$, the family formed by all the valid instrumental variables.

2. We show that Algorithm 3.2 can recover $\mathcal{F}_0$ from the selection path from Algorithm 3.1.

Part 1 follows from Corollary 3.2 directly.
Part 2: Firstly, we establish the properties of the Sargan statistic. The following two equations can be also found in Windmeijer et al. (2021) (p.10). Let $\mathcal{I}$ be the true set of invalid instruments and $\mathcal{V}$ be the true set of valid instruments. The oracle model is

$$\mathbf{y} = \mathbf{D}\boldsymbol{\beta} + \mathbf{Z}_{\mathcal{I}}\boldsymbol{\alpha}_{\mathcal{I}} + \mathbf{u} = \mathbf{X}_{\mathcal{I}}\boldsymbol{\theta}_{\mathcal{I}} + \mathbf{u}$$

with $\mathbf{X}_{\mathcal{I}} = [\mathbf{D} \quad \mathbf{Z}_{\mathcal{I}}]$ and $\boldsymbol{\theta}_{\mathcal{I}} = [\boldsymbol{\beta} \quad \boldsymbol{\alpha}_{\mathcal{I}}']'$, the Sargan test statistic is given by

$$S(\hat{\boldsymbol{\theta}}_{\mathcal{I}}) = \frac{\hat{\mathbf{u}}(\hat{\boldsymbol{\theta}}_{\mathcal{I}})'\mathbf{Z}_{\mathcal{I}}(\mathbf{Z}_{\mathcal{I}}'\mathbf{Z}_{\mathcal{I}})^{-1}\mathbf{Z}_{\mathcal{I}}'\hat{\mathbf{u}}(\hat{\boldsymbol{\theta}}_{\mathcal{I}})}{\hat{\mathbf{u}}(\hat{\theta}_{\mathcal{I}})'\hat{\mathbf{u}}(\hat{\boldsymbol{\theta}}_{\mathcal{I}})/n} \tag{3.A.7}$$

where $\hat{\mathbf{u}}(\hat{\boldsymbol{\theta}}) = \mathbf{y} - \mathbf{X}_{\mathcal{I}}\hat{\boldsymbol{\theta}}_{\mathcal{I}}$, with $\hat{\boldsymbol{\theta}}_{\mathcal{I}}$ the 2SLS estimator of $\boldsymbol{\theta}_{\mathcal{I}}$.
Let $\hat{\mathcal{I}}$ be the estimated set of invalid instruments and $\hat{\mathcal{V}}$ be the estimated set of valid instruments where $\hat{\mathcal{I}} = \mathcal{J} \setminus \hat{\mathcal{V}}$. Following Proposition 3.2 in Windmeijer et al., 2021, the Sargan statistic has the following properties:

Property 3.3. *Properties of the Sargan statistic*

1. *For all the* $\binom{|\hat{\mathcal{V}}|}{P}$ *combinations of the instruments from* $\hat{\mathcal{V}}$*, if the IVs contained in them belong to the same family, then:* $S(\hat{\boldsymbol{\theta}}_{\hat{\mathcal{I}}}) \xrightarrow{d} \chi^2_{|\mathcal{J}|-|\hat{\mathcal{I}}|-P}$

2. *For all the* $\binom{|\hat{\mathcal{V}}|}{P}$ *combinations of the instruments from* $\hat{\mathcal{V}}$*, if the IVs contained in them belong to a mixture of families, then:* $S(\hat{\boldsymbol{\theta}}_{\hat{\mathcal{I}}}) = O_p(n)$.

With these properties we can show that the downward testing procedure described in Algorithm 3.2 selects valid instruments consistently with $\xi_{n,J-|\hat{\mathcal{I}}|-P} \to \infty$ for $n \to \infty$, and $\xi_{n,J-|\hat{\mathcal{I}}|-P} = o(n)$. Let the number of clusters formed in Algorithm 3.1 at some certain step be K, e.g. at Step 1, $K = \binom{J}{P}$ and at Step 2, $K = \binom{J}{P} - 1$ etc. Let the true number of families be Q. Consider applying the Sargan test to the model selected by the largest cluster at the each step under the following scenarios:

1. $1 \leq K < Q$. For each of these steps, the largest cluster is either associated with a mixture of different families, or with one family.

   - Consider the case where the largest cluster is associated with a mixture of different families. Then by Property 3.3 and $\xi_{n,J-|\hat{\mathcal{I}}|-P} = o(n)$, we have
   $$\lim_{n\to\infty} P(S(\hat{\boldsymbol{\theta}}_{\hat{\mathcal{I}}}) < \xi_{n,J-|\hat{\mathcal{I}}|-P}) = 0.$$
   In this case, asymptotically the Sargan test would be rejected and the downward testing procedure moves to the next step.
   - Consider the case where the largest cluster is associated with one family. Then this family must be $\mathcal{F}_0$ as by Assumption 3.6.a, $\mathcal{F}_0$ is the largest family among all Q families. Then following Property 3.3 and $\xi_{n,J-|\hat{\mathcal{I}}|-P} \to \infty$ for the Sargan test we have
   $$\lim_{n\to\infty} P(S(\hat{\boldsymbol{\theta}}_{\hat{\mathcal{I}}}) < \xi_{n,J-|\hat{\mathcal{I}}|-P}) = 1. \tag{3.A.8}$$
   indicating that $\mathcal{V}$ would be selected as the set of valid instruments asymptotically.

2. $K = Q$. By Corollary 3.2 we know that the K clusters are associated with the Q families respectively, and by Assumption 3.6.a, the cluster associated with $\mathcal{F}_0$ is the largest cluster. Then applying the Sargan test at this step would be testing all the valid instruments, hence we also have Equation (3.A.8) and Algorithm 3.2 selects $\mathcal{V}$ as the set of valid instruments.

To summarize, asymptotically, at steps $1 \leq K < Q$, Algorithm 3.2 only stops when $\mathcal{F}_0$ forms the largest cluster and hence selects the oracle model, otherwise it moves to step $K = Q$ and selects the oracle model. Combine *Part 1* and *Part 2*, we prove Theorem 3.1.

□

Proof of Theorem 3.3

The proof of Theorem 3.3 works in the same way as the proof of Theorem 3.1.

Proof. The proof for Theorem 3.3 is structured as follows:

1. We note that asymptotically the selection path generated from Algorithm 3.1 contains all groups $\mathcal{G}_q$.

2. We show that Algorithm 3.2 can recover all $\mathcal{G}_q$ from the selection path from Algorithm 3.1.

Part 1 again follows directly from Corollary 3.2.
Part 2: Firstly, we establish the properties of the Sargan statistic.

Property 3.4. *Properties of the Sargan statistic*

1. *For all combinations of instruments from $\hat{\mathcal{G}}_k$, if their just-identified estimators are associated with the same group, then: $S(\hat{\boldsymbol{\theta}}_{\hat{\mathcal{G}}_k}) \xrightarrow{d} \chi^2_{J-|\hat{\mathcal{G}}_k|-P}$*

2. *For all combinations of instruments from $\hat{\mathcal{G}}_k$, if their just-identified estimators are associated with a mixture of groups, then: $S(\hat{\boldsymbol{\theta}}_{\hat{\mathcal{G}}_k}) = O_p(n)$.*

As before, $\xi_{n,J-|\hat{\mathcal{I}}|-P} \to \infty$ for $n \to \infty$, and $\xi_{n,J-|\hat{\mathcal{I}}|-P} = o(n)$. Consider applying the Sargan test to each cluster separately at the following steps under the following scenarios:

1. $1 \leq K < Q$, i.e. the number of clusters is smaller than the number of groups. For each of these steps, at least one cluster is associated with a mixture of different groups.

   When one cluster is created by a mixture of different groups, by Property 3.4, we have

   $$\lim_{n\to\infty} P(S(\hat{\boldsymbol{\theta}}_{\mathcal{G}_q}) < \xi_{n,J-|\mathcal{G}_q|-P}) = 0. \tag{3.A.9}$$

   In this case, asymptotically at least one of the the Sargan tests would be rejected and the downward testing procedure moves to the next step.

2. $K = Q$. By Corollary 3.2 we know that the K clusters are formed by the Q groups respectively and $\hat{\mathcal{G}}_k = \mathcal{G}_q$ for all q. Then for each of the K tests we have

   $$S(\hat{\boldsymbol{\theta}}_{\hat{\mathcal{G}}_k}) = S(\hat{\boldsymbol{\theta}}_{\mathcal{G}_q}). \tag{3.A.10}$$

   By Property 3.4 and $\xi_{n,J-|\mathcal{G}_q|-P} = o(n)$, we have

   $$\lim_{n\to\infty} P(S(\hat{\boldsymbol{\theta}}_{\mathcal{G}_q}) < \xi_{n,J-|\hat{\mathcal{I}}|-P}) = 1.$$

In this case, Algorithm 3.2 stops.

Then applying the Sargan tests to each group at this step will be testing IVs from the same group each time, hence we also have Equation (3.A.9).

To summarize, asymptotically, at steps $1 \leq K < Q$, Algorithm 3.2 does not stop; then it moves to step $K = Q$ and selects the oracle model.
Combine *Part 1* and *Part 2*, we prove Theorem 3.1. □

Bibliography

Aggarwal, Charu C, Alexander Hinneburg, and Daniel A Keim (2001). "On the Surprising Behavior of Distance Metrics in High Dimensional Space". In: *International Conference on Database Theory*. Springer, pp. 420–434.

Altonji, Joseph G and David Card (1991). "The Effects of Immigration on the Labor Market Outcomes of Less-Skilled Natives". In: *Immigration, Trade, and the Labor Market*. University of Chicago Press, pp. 201–234.

Amorim, Renato Cordeiro de, Vladimir Makarenkov, and Boris Mirkin (2016). "A-Wardpβ: Effective hierarchical clustering using the Minkowski metric and a fast k-means initialisation". In: *Information Sciences* 370, pp. 343–354.

Andrews, Donald W. K. (1999). "Consistent Moment Selection Procedures for Generalized Method of Moments Estimation". In: *Econometrica* 67.3, pp. 543–563. DOI: 10.1111/1468-0262.00036.

Angrist, Joshua D and Ivan Fernandez-Val (2010). *Extrapolate-ing: External Validity and Overidentification in the LATE Framework*. Tech. rep. National Bureau of Economic Research.

Apfel, Nicolas (2021). "Relaxing the Exclusion Restriction in Shift-Share Instrumental Variable Estimation". In: *Available at SSRN 3408682*.

Athey, Susan and Guido W Imbens (2019). "Machine learning methods that economists should know about". In: *Annual Review of Economics* 11, pp. 685–725.

Basso, Gaetano and Giovanni Peri (2015). "The Association Between Immigration and Labor Market Outcomes in the United States". In: *IZA Discussion Paper 9436*.

Belloni, A. et al. (2012). "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain". In: *Econometrica* 80.6, pp. 2369–2429. DOI: 10.3982/ecta9626.

Bowden, J., G. Davey Smith, and S. Burgess (2015a). "Mendelian Randomization With Invalid Instruments: Effect Estimation and Bias Detection Through Egger Regression". In: *International Journal of Epidemiology* 44.2, pp. 512–525. DOI: 10.1093/ije/dyv080.

Bowden, Jack, George Davey Smith, and Stephen Burgess (2015b). "Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression". In: *International journal of epidemiology* 44.2, pp. 512–525.

Bowden, Jack et al. (2016). "Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator". In: *Genetic Epidemiology* 40.4, pp. 304–314. DOI: 10.1002/gepi.21965.

Bühlmann, Peter and Sara Van De Geer (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Burgess, Stephen, Dylan S Small, and Simon G Thompson (2017). "A Review of Instrumental Variable Estimators for Mendelian Randomization". In: *Statistical Methods in Medical Research* 26.5, pp. 2333–2355. DOI: 10.1177/0962280215597579.

Burgess, Stephen et al. (2018). *Assessing the Effectiveness of Robust Instrumental Variable Methods Using Multiple Candidate Instruments with Application to Mendelian Randomization*. arXiv:1606.03729.

Chand, Sohail (2012). "On tuning parameter selection of lasso-type methods-a monte carlo study". In: *Proceedings of 2012 9th international Bhurban conference on applied sciences & technology (IBCAST)*. IEEE, pp. 120–129.

Clarke, Paul S. and Frank Windmeijer (2012). "Instrumental Variable Estimators for Binary Outcomes". In: *Journal of the American Statistical Association* 107.500, pp. 1638–1652. DOI: 10.1080/01621459.2012.734171.

Davidson, Russell and James G. MacKinnon (2004). *Econometric Theory and Methods*. New York: Oxford University Press. ISBN: 9780195123722.

Davies, Neil M. et al. (2015). "The many weak instruments problem and Mendelian randomization". In: *Statistics in Medicine* 34.3, pp. 454–468. DOI: 10.1002/sim.6358.

Donoho, David L and Iain M Johnstone (1994). "Ideal Spatial Adaptation by Wavelet Shrinkage". In: *Biometrika* 81.3, pp. 425–455. DOI: 10.1093/biomet/81.3.425.

Dustmann, Christian, Uta Schönberg, and Jan Stuhler (2016). "The Impact of Immigration: Why Do Studies Reach Such Different Results?" In: *Journal of Economic Perspectives* 30.4, pp. 31–56.

Efron, Bradley et al. (2004). "Least angle regression". In: *The Annals of statistics* 32.2, pp. 407–499.

Fan, Jianqing and Runze Li (2001). "Variable selection via nonconcave penalized likelihood and its oracle properties". In: *Journal of the American statistical Association* 96.456, pp. 1348–1360.

Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift (2020). "Bartik Instruments: What, When, Why, and How". In: *American Economic Review*.

Guo, Zijian et al. (2018). "Confidence Intervals for Causal Effects with Invalid Instruments by Using Two-Stage Hard Thresholding with Voting". In: *Journal of the Royal Statistical Society: Series B* 80.4, pp. 793–815. DOI: 10.1111/rssb.12275.

Hansen, Lars Peter (1982). "Large Sample Properties of Generalized Method of Moments Estimators". In: *Econometrica* 50.4, pp. 1029–1054. DOI: 10.2307/1912775.

Hartwig, Fernando Pires, George Davey Smith, and Jack Bowden (2017). "Robust Inference in Summary Data Mendelian Randomization Via The Zero Modal Pleiotropy Assumption". In: *International Journal of Epidemiology* 46.6, pp. 1985–1998. DOI: 10.1093/ije/dyx102.

Hastie, Trevor and Brad Efron (2013). *lars: Least Angle Regression, Lasso and Forward Stagewise*. R package version 1.2.

Hinke, Stephanie von et al. (2016). "Genetic Markers as Instrumental Variables". In: *Journal of Health Economics* 45, pp. 131–148. DOI: 10.1016/j.jhealeco.2015.10.007.

Holland, Paul W. (1988). "Causal Inference, Path Analysis, and Recursive Structural Equations Models". In: *Sociological Methodology* 18, pp. 449–484. DOI: 10.2307/271055.

Imbens, Guido W and Joshua D Angrist (1994). "Estimation and Identification of Local Average Treatment Effects". In: *Econometrica* 62, pp. 467–475.

Imbens, Guido W. (2014). "Instrumental Variables: An Econometrician's Perspective". In: *Statistical Science* 29.3, pp. 323–358. DOI: 10.1214/14-sts480.

Jaeger, David A, Joakim Ruist, and Jan Stuhler (2020). *Shift-Share Instruments and the Impact of Immigration*. Tech. rep. National Bureau of Economic Research.

Kang, Hyunseung (2018). *TSHT.R*. https://github.com/hyunseungkang/invalidIV.

Kang, Hyunseung et al. (2016). "Instrumental Variables Estimation With Some Invalid Instruments and its Application to Mendelian Randomization". In: *Journal of the American Statistical Association* 111.513, pp. 132–144. DOI: 10.1080/01621459.2014.994705.

Karp, Richard M. (1972). "Reducibility among Combinatorial Problems". In: *Complexity of Computer Computations*. Ed. by Raymond E. Miller, James W. Thatcher, and Jean D. Bohlinger. Springer US, pp. 85–103. DOI: 10.1007/978-1-4684-2001-2_9.

Kolesár, Michal et al. (2015). "Identification and Inference With Many Invalid Instruments". In: *Journal of Business & Economic Statistics* 33.4, pp. 474–484. DOI: 10.1080/07350015.2014.978175.

Lawlor, Debbie A. et al. (2008). "Mendelian Randomization: Using Genes as Instruments for Making Causal Inferences in Epidemiology". In: *Statistics in Medicine* 27.8, pp. 1133–1163. DOI: 10.1002/sim.3034.

Locke, Adam E., Bratati Kahali, Sonja I. Berndt, et al. (2015). "Genetic Studies of Body Mass Index Yield New Insights for Obesity Biology". In: *Nature* 518.7538, pp. 197–206. DOI: 10.1038/nature14177.

Newey, Whitney K. and Kenneth D. West (1987). "Hypothesis Testing with Efficient Method of Moments Estimation". In: *International Economic Review* 28.3, pp. 777–787. DOI: 10.2307/2526578.

Okbay, Aysu et al. (2016). "Genome-wide association study identifies 74 loci associated with educational attainment". In: *Nature* 533.7604, pp. 539–542.

Pearl, Judea (2009). *Causality*. Cambridge University Press. DOI: 10.1017/cbo9780511803161.

Pötscher, B. M. (1983). "Order Estimation in ARMA-Models by Lagrangian Multiplier Tests". In: *The Annals of Statistics* 11.3, pp. 872–885. DOI: 10.1214/aos/1176346253.

Ruggles, Steven et al. (2015). "Integrated Public Use Microdata Series: Version 6.0 [Dataset]". In: *Minneapolis: University of Minnesota* 23, p. 56.

Sanderson, Eleanor and Frank Windmeijer (2016). "A weak instrument F-test in linear IV models with multiple endogenous variables". In: *Journal of econometrics* 190.2, pp. 212–221.

Sanderson, Eleanor et al. (2019). "An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings". In: *International Journal of Epidemiology* 48.3, pp. 713–727. DOI: 10.1093/ije/dyy262.

Sargan, J. D. (1958). "The Estimation of Economic Relationships using Instrumental Variables". In: *Econometrica* 26.3, pp. 393–415. DOI: 10.2307/1907619.

Sniekers, Suzanne et al. (2017). "Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence". In: *Nature genetics* 49.7, pp. 1107–1112.

Staiger, Douglas and James H. Stock (1997). "Instrumental Variables Regression with Weak Instruments". In: *Econometrica* 65, pp. 557–586. DOI: 10.2307/2171753.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288.

Von Hinke, Stephanie et al. (2016). "Genetic markers as instrumental variables". In: *Journal of Health Economics* 45, pp. 131–148.

Ward, Joe H Jr (1963). "Hierarchical Grouping to Optimize an Objective Function". In: *Journal of the American Statistical Association* 58.301, pp. 236–244.

Wensley, Frances et al. (2011). "Association Between C Reactive Protein and Coronary Heart Disease: Mendelian Randomisation Analysis Based on Individual Participant Data". In: *BMJ* 342.feb15, p. d548.

Windmeijer, Frank (2019). "Two-Stage Least Squares as Minimum Distance". In: *The Econometrics Journal* 22.1, pp. 1–9.

Windmeijer, Frank et al. (2019). "On the Use of the Lasso for Instrumental Variables Estimation with Some Invalid Instruments". In: *Journal of the American Statistical Association* 114.527, pp. 1339–1350. DOI: 10.1080/01621459.2018.1498346.

Windmeijer, Frank et al. (2021). "The confidence interval method for selecting valid instrumental variables". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 83.4, pp. 752–776.

Wooldridge, Jeffrey M (2010). *Econometric analysis of cross section and panel data*. MIT press.

Wright, Philip G (1928). *Tariff on animal and vegetable oils*. Macmillan Company, New York.

Zhao, Qinguan et al. (2019). *Statistical Inference in Two-Sample Summary-Data Mendelian Randomization Using Robust Adjusted Profile Score*. arXiv:1801-09652.

Zhao, Qingyuan et al. (2020). "Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score". In: *The Annals of Statistics* 48.3, pp. 1742–1769.

Zou, Hui (2006). "The Adaptive Lasso and Its Oracle Properties". In: *Journal of the American Statistical Association* 101.476, pp. 1418–1429.