Methods for Selection of Valid Instrumental Variables – Theory and Applications

Dissertation zur Erlangung des Grades eines Doktors der Wirtschaftswissenschaft

eingereicht an der Fakultät für Wirtschaftswissenschaften der Universität Regensburg

vorgelegt von:
Nicolas Apfel

Berichterstatter:
Prof. Dr. Andreas Roider, Prof. Frank Windmeijer, Ph.D.

Tag der Disputation: 6. Dezember 2021

DISSERTATION

METHODS FOR SELECTION OF VALID INSTRUMENTAL VARIABLES – THEORY AND APPLICATIONS

NICOLAS APFEL

Nicolas Apfel: Methods for Selection of Valid Instrumental Variables – Theory and Applications, Dissertation, September 2021

ACKNOWLEDGMENTS

The four years of work on this dissertation have been an intense experience and have required countless solitary working sessions. However, what has made this voyage enjoyable were the long discussions, the brainstorming about new research ideas and the conference dinners. The people who have accompanied me along the way have been great teachers, colleagues and friends. It is you who made this voyage worthwhile.

This thesis would not have been possible without the support and guidance of my supervisors. I am indebted to Professor Roider, who had faith in this undertaking and gave me the freedom to explore my academic interests. Thank you, Andreas for the successful cooperation, for encouraging me to aim high, to present my work early on, for building a research network from which I have benefited immensely and for supporting me at each step of my dissertation. I also want to thank Professor Frank Windmeijer for supervising this thesis. I have been extremely lucky that you offered to supervise my thesis. I feel privileged to be your student and could not imagine a better supervisor for this dissertation. This thesis would not have been possible without your patience, insight and guidance.

I have immensely benefited from the repeated interactions with researchers from other universities. I want to thank each one of them. I will always remember the casual "I think research is for you", told to me by Angelika Schmid, my supervisor as an undergraduate. Little did we know that we would spend two years on a project on austerity policies. Even though this particular project did not flourish, your encouragement has meant a lot to me. Benjamin Elsner has shown remarkable patience when discussing my work with me and when explaining how to improve my academic writing. Jan Stuhler has been incredibly supportive with respect to my work. I hope we will meet again at a conference or for a coffee, in Munich. Stephan Huber has been a terrific mentor since my master's studies. Niklas Wallmeier not only has been a great roommate during our departments' joint seminar, he and I have also taken it upon us to devour industrial quantities of fish buns at said seminar, but more importantly he had the right advice at the right time. It has been a great pleasure to get to know Xiaoran Liang and to engage in endless phone calls on all sorts of questions about instrumental variables. Going through the last two years of the PhD together has made things considerably more pleasant and fun. Likewise, it has been uplifting to hear constructive critique along with warm words from colleagues

at conferences and visitors at Regensburg University. I would like to mention Carsten Feuerbaum, Nicolas Salamanca, Alexandra de Gendre, Martin Hackmann, Stephan Heblich and the numerous others I have forgotten. Thank you!

These years would not have been the same without my colleagues and friends at the University of Regensburg. You have brought color in the otherwise cement-gray tower in which our department is located. Thanks to Alexander Lauf for tolerating me in our joint office, and to everyone who has made our lunches and table football matches fun: David Russ, Michael Heyna, Tobias Hartl, Christoph Rust and all those that I have forgotten here. A special thank you also goes to Petra Gilg and Martina Kraus-Pietsch for making me feel at home and for the steady exchange of cookies.

I also want to thank my dear friends. Thanks, Madhi for the good laughs and our dinners, thanks to Valentin for the speedminton matches and breakfast feasts. Thanks to Johannes - rest in peace! Thanks to David for volley ball games and for some of the most liberating economics-related rants. Thanks to Marta, for 30 years of friendship - we will turn psycholinguistics upside down with instrumental variables, trust me!

If I finished this thesis, it is also because of the support of a few very important people in my life: my family, to whom I dedicate this thesis. Ingrid, Raffaele, Emily, Weiwei, thanks for always being there! Without your trust, love and unceasing nodding, when listening to my helpless translations of shift-share instruments to plain language I would not have had the strength to follow through on this quest and I would not be the person I am now. One special thank to Weiwei: thank you for your unconditional love, for the support during the tough last weeks. This would not have been possible without you.

For my family

CONTENTS

LIST OF FIGURES

LIST OF TABLES

I

PREFACE

Most studies in modern economics revolve around the question of how an exposure causes an outcome. This is the central topic of causal inference. Researchers are often confronted with the problem of estimating causal relationships in an accurate way. In practice, not only could the exposure cause the outcome, but the outcome might also affect the exposure, or there might be a third, unobserved factor which confounds the relationship of interest. This complicates the task of inferring a causal relationship.

One common statistical technique to address these problems is to use an instrumental variable (IV), which is a variable related with the outcome only through the exposure. This kind of approach is used very frequently in economic research, but presents its own problems. What if the IV is itself correlated with an unobservable variable or with the outcome directly? In this case the IV is said to be invalid or to fail the exclusion restriction, and the estimation strategy fails. Then, apparently, the researcher is back to square one. A new statistical literature shows a way out of this dilemma in the following way: if there are more IVs than exposures, a suitable subset of IVs can be selected. The main research questions of this thesis are:

- How can these IV selection methods be applied to economic settings?
- How can these methods be improved and made more readily applicable?

This dissertation provides applied and theoretical work which will help researchers in economics estimate causal effects more reliably. In this way, I show that even when IV estimation does not seem suited, with the help of these new methods, researchers can still retrieve informative estimates from the data. I connect the IV selection literature to applied econometric research and illustrate how researchers can easily apply it to the questions under study. I apply the methods to economic examples and improve on the existing methods by proposing an alternative selection method. The methods are also extended to settings which are likely to occur in reality.

The first derivation of the IV estimator dates back to Phillip G. Wright, who discussed how the relationship between price and quantity could be determined with the help of an IV (Wright, 1928). This application gave rise to one of the most vibrant branches of modern econometrics. The method has been applied in labor, development, health, financial, applied macroeconomics and many other fields of research.

The IV needs to fulfill two key assumptions: relevance and validity. *Relevance* means that the instrument should be sufficiently related with the exposure. *Validity* denotes

that there should be no direct effect of the instrument on the outcome and no correlation with unobservables which are then correlated with the outcome. The consequences of a violation of the first assumption, when an IV is weak, is the object of an entire branch of the literature. The relevance assumption can be tested and there is also considerable work on how to improve the strength of the instrument. For example, Belloni, Chen, Chernozhukov, and Hansen (2012) and Lin, Feng, and Li (2015) show that a subset of strong instruments can be selected from a large set of instruments to predict the endogenous regressor. This literature assumes that all instruments are valid. There are few studies about separating valid and invalid instruments when both are present. Therefore, in this thesis I will focus on how to mend violations of validity.

So far, validity has been defended mostly on two grounds. First, a feasible intuition and theoretical underpinning about why the instrument should fulfill the assumption are central. Thorough empirical work spends a great deal of effort to justify this assumption. The theoretical arguments made, of course might not apply to the real data. Second, when there are multiple potential instruments, joint validity of all instruments can be tested with the help of so-called overidentification tests. This entails auxiliary assumptions, such as constant treatment effects and the assumption that at least one IV is valid. In applied econometric work, a combination of intuition, theory and over-identification tests has dominated, in the last decades. One key problem with the current econometric toolkit is that if the overidentification tests reject the Null hypothesis that all instruments are valid, then it is unclear in which direction a researcher should proceed.

Some studies have treated this problem. Kolesár, Chetty, Friedman, Glaeser, and Imbens (2015) and Bowden, Davey Smith, and Burgess (2015) provide inferential methods when there are many invalid instruments. The key assumption is that the direct effects of the instrument on the outcome and the effect of the instrument on the endogenous regressor are uncorrelated. Conley, Hansen, and Rossi (2012) take a Bayesian perspective on the problem and use prior knowledge about the extent of the violation of validity. The disadvantage of these methods is that knowledge on properties of the direct effects is needed.

When in a pool of many IVs some are valid and others are invalid, something can be done to retrieve useful estimates, even though it is unknown which IVs are valid and invalid. D. W. Andrews (1999) proposes moment selection criteria that are based on overidentification test statistics. In cases with many instruments, this approach requires to evaluate all possible overidentified models, making the approach infeasible. In the last five years, a new literature that originated in genetic statistics proposes ways to reduce this high-dimensional IV selection problem. Kang, Zhang, Cai, and Small (2016), Windmeijer, Farbmacher, Davies, and Smith (2019), Guo, Kang, Cai, and Small (2018) and Windmeijer, Liang, Hartwig, and Bowden (2021) are the main contributors to this emerging literature. The key idea of these methods is to find a subset of IVs that is valid, under the assumption that a large enough number of IVs is indeed valid. These methods select a model from a potentially high-dimensional set of models. This is why machine-learning methods are used to help reduce the dimensionality of the problem.

The key contributions of this thesis to the literature are five-fold:

1. I show how the new methods can be applied in an economic context. The first chapter can be read as a practitioner's guide for economists to help researchers obtain more reliable estimates in the context of a widely-used class of instruments: shift-share instruments.

2. I develop a new IV selection procedure, which combines the agglomerative hierarchical clustering algorithm with a downward testing procedure. This method can deal effectively with the weak instrument problem and can be applied to a setting with multiple endogenous regressors and with heterogeneous treatment effects.

3. It is shown how the results for the selection methods apply to settings with small violations of validity.

4. The selection literature is connected and compared to another novel econometric study which looks at the minimal deviations from the exclusion restriction such that the model is non-falsified. I reconcile these two approaches and discuss how these two methods can be used together.

5. I provide code in `R` and `STATA` to allow researchers to apply the new methods.

In the first chapter of my thesis, I show that the IV selection literature proves very helpful when applying it to a class of IVs, so-called shift-share IVs. The research question of the first chapter is the following: can shift-share IV estimation be consistent when many shares violate the exclusion restriction? Many economic studies use shift-share instruments to estimate causal effects. In these research designs, often all shares need to fulfill an exclusion restriction, making the identifying assumption very strict.

This chapter proposes the use of methods that relax validity by consistently selecting invalid shares and extends one method to the multiple endogenous regressor case. I apply the methods to simulated data and two empirical examples: the effect of immigration on wages, and the effect of Chinese import exposure on employment. In the simulations I find that settings with weak instruments and strong violations of the exclusion restriction still lead to good selection results and performance of the estimators, when the sample is large. In both empirical applications, the coefficient estimates can change considerably but these changes are reconcilable with the findings in the literature. This chapter uses existing IV selection methods. A few important, open questions in the literature are how to extend the methods to a setting with multiple endogenous regressors, how to deal with weak instruments, and how to treat local average treatment effects (LATE). In Chapter 2, I therefore propose a selection procedure which tackles these problems.

In the second chapter, I propose an IV selection procedure which combines the agglomerative hierarchical clustering method and the Hansen-Sargan overidentification test for selecting valid instruments for IV estimation from a large set of candidate instruments. I show that under the plurality rule, i.e. when the largest group of instruments is valid, the method can achieve oracle selection and estimation results. Compared to the previous IV selection methods, the method has the advantages that it can deal with the weak

instruments problem effectively, and can be easily extended to settings where there are multiple endogenous regressors and heterogenous treatment effects. I conduct Monte Carlo simulations to examine the performance of my method, and compare it with two existing methods, the hard thresholding method and the confidence interval Method. The simulation results show that the new method achieves oracle selection and estimation results in both single and multiple endogenous regressors settings in large samples when all the instruments are strong. Also, the method works well when some of the candidate instruments are weak, outperforming the hard thresholding method and the confidence interval method. Finally, I apply the method to the estimation of the effect of immigration on wages in the US. The new method still relies on the assumption that there is a group of IVs for which the validity assumption holds exactly. In the next chapter, I therefore relax this assumption and discuss what the implications are for the confidence interval method and the agglomerative hierarchical clustering.

In the third chapter, I investigate the performance of IV selection methods when some IVs are invalid but there is also a component of the invalidity which disappears asymptotically, so that the exclusion restriction does not hold exactly. In realistic settings, there might be small correlations between instruments and the error term so that the validity assumption only holds approximately. When the violations are major, the methods break down. When the violations are minor, the selection methods still produce reliable results. In a special case, the findings imply that the plurality assumption can be relaxed. Monte Carlo simulations confirm the theoretical results. Selecting valid instruments is not the only way in which the problem of invalid instruments can be tackled. Therefore, it is interesting to ask how the methods relate to other approaches that have been proposed recently.

In the fourth chapter, I review recent work by Masten and Poirier (2021) who ask what researchers should do when an overidentified IV model is falsified. In their work, they suggests to report the minimal relaxations that do not lead to a falsification, the falsification frontier (FF), and the associated identified set, the falsification adaptive set (FAS). I compare the two approaches, discuss advantages and disadvantages and restate the confidence interval method in the framework in Masten and Poirier (2021). I introduce issues discussed in the IV selection literature which can be of interest for researchers wanting to apply the FF and FAS. Finally, I show that allowing for violations of exogeneity instead of violations of the exclusion restriction leads to ambiguous results with respect to the FF and FAS.

The IV selection literature originates in genetical statistics, where genes are used as IVs for certain exposures. One general contribution of this thesis is to apply methods originally developed for genetics in an economic context, and to develop them further. Historically, IV estimation has been developed by econometricians and biostatisticians.[1] In this way, this thesis also connects to the aspects which stood at the very beginning of IV estimation.

[1] Interestingly, authorship of the relevant Appendix B in Wright (1928), which contains the derivation of the IV estimator is posthumously contended between Philipp G. Wright, an economist, and his eldest son, Sewall Wright, a genetic statistician (Stock and Trebbi, 2003).

This thesis also connects to the literature on machine learning. Machine learning or statistical learning can be defined as the analysis of high-dimensional data with the help of computationally intensive algorithms. The model-building process is informed by the researcher and is assisted by the machine learning algorithm. In recent years, an emerging literature has tried to connect machine learning to economic research. One part of of this literature puts its emphasis on prediction tasks (Mullainathan and Spiess, 2017). Prediction of an outcome variable is a natural task for machine learning algorithms, because causal mechanisms play only a subordinate role. Another strand of this literature connects machine learning methods with the estimation of causal treatment effects. Some of these methods are discussed in Athey and Imbens (2019). In this thesis, causality is central and there are usually many instruments to select from. This thesis illustrates how statistical learning can propel the analysis of economic questions forward when problems are high-dimensional, and it illustrates this added value that machine learning brings to causal inference.

The remainder of this thesis is structured as follows: in Chapter II, I illustrate how IV selection methods can be applied to shift-share IVs estimation. In Chapter III, I present the agglomerative hierarchical clustering method, a development of the existing instrument selection methods. In Chapter IV, I apply the selection methods to settings where the validity assumption is not exactly fulfilled. In Chapter V, I propose a synthesis between the IV selection literature and a new econometric approach to quantify the extent of invalidity. In Chapter VI, I conclude with a summary, possible shortcomings and directions for future research.

II

RELAXING THE EXCLUSION RESTRICTION IN SHIFT-SHARE INSTRUMENTAL VARIABLE ESTIMATION

1 INTRODUCTION

The shift-share instrument is often used in applied economics to obtain estimates of causal effects. The numerous applications have spanned three decades, beginning with Bartik, 1991 seminal paper, and included several fields, such as migration, labor, international economics and many other. A shift-share strategy exploits shares at an earlier point in time and current aggregate-level changes to create an IV. For example, the share of migrants from a certain origin country is interacted with the inflow from that country. The methods proposed in this chapter are not restricted to the examples mentioned here, but can be applied to a wide range of studies in which shift-share instruments are used.

The key assumptions and properties of shift-share instruments have been investigated only in very recent work. Centrally, in an important setting the exclusion restriction must hold for all initial shares, for the estimator to be consistent. In practice, a potentially large number of shares cannot have a direct effect on the outcome. This assumption is very strict, because it requires the researcher to have perfect structural knowledge about all shares, which is typically unavailable. The natural question to ask, therefore is: Is consistent estimation still possible when the exclusion restriction is violated for some but not all shares?

In this chapter, my main contribution is to show how consistent shift-share estimation is possible, when not all shares fulfill the exclusion restriction. This chapter is a practitioner's guide on how to select invalid shares in the shift-share setting, using two methods developed in bio-statistics. I also extend one of the methods that I present to allow for multiple endogenous regressors. This was not possible so far and is a further contribution of the chapter.

The proposed methods go beyond existing econometric diagnostics typically applied in this setting. Rotemberg weights, proposed by Goldsmith-Pinkham, Sorkin, and Swift (2020), report the sensitivity to misspecification of the different shares. These weights often fail to provide clear-cut guidance as to which shares should finally be included in the construction of the instrument, because they tell the researcher how large the relative bias of the entire shift-share estimator stemming from the bias of a single industry is. Instead, with the methods proposed in this chapter the researcher obtains an estimate for the identity of valid and invalid instruments and a consistent estimate, adjusted from the absolute bias. This bears the advantage that valid instruments do not need to be discarded just because their potential invalidity could lead to bias.

Applying the new methods to simulated data and revisiting the effects of immigration and of Chinese import exposure on the US labor market illustrates that the shares selected as invalid in these applications are consistent with those discussed as problematic in the literature. So far, no way to locate invalid shares has been proposed. This chapter fills this gap and provides a principled approach to share selection. To make the methods more widely accessible to practitioners, I also provide simple-to-use Stata-programs.

I begin by presenting the shift-share IV and its key identifying assumption, the exclusion restriction. In the shift-share approach, there are multiple class-specific shares and shifts

which are interacted to produce the final instrument. Goldsmith-Pinkham, Sorkin, and Swift (2020, GPSS) show that if the exclusion restriction holds for each class-specific share, the IV estimator is consistent. That is, shares should not be directly correlated with the outcome variable through unobservable shocks or longterm effects. The exclusion restriction from the shares perspective is a sufficient condition for the consistency of the shift-share IV estimator. Instruments which fulfill the exclusion restriction are called *valid*, while those that do not fulfill it are called *invalid*. This definition of validity assumes that all instruments are related to the treatment. This exclusion restriction is very restrictive because it must hold for *all* classes. While the general idea behind the shift-share IV is credible, typically structural relationships between instruments and outcome variables are difficult to exclude for each single class.

In Section 3, I show how to obtain consistent estimators, when many shares are invalid. To achieve consistency, invalid shares are selected using two methods: the adaptive Least absolute shrinkage and selection operator (ALasso) by Windmeijer, Farbmacher, Davies, and Smith (2019) and the confidence interval method (CIM) by Windmeijer, Liang, Hartwig, and Bowden (2021). The two methods have been developed primarily for the use in Mendelian randomization, which is the application of IV estimation in genetic epidemiology. In these applications, genetic markers are used as IVs when estimating the effect of an exposure on a health outcome.

The key advantage of the leveraged methods is that they consistently select shares, which violate the exclusion restriction. When the *majority* of instruments is valid, both metods have so-called *oracle properties* and when the *largest group* of shares fulfills the exclusion restriction the CIM has oracle properties. This means that asymptotically the post-selection estimators perform as well as if the researcher knew the identity of invalid IVs. Intuitively, these estimators both exploit the fact that just-identified estimates, which use only one valid IV at a time, converge against the same value. None of the existing methods allow for multiple endogenous regressors. I therefore propose a simple extension of the ALasso. Here, the exclusion restriction becomes stricter with increasing number of endogenous regressors.

To show the implications and generality of the presented methods in practice, in Sections 4 to 6, I apply them to simulated data and two empirical examples. In a first set of simulations, the results confirm that in applications with shares as IVs with increasing sample size the performance approaches that of the estimator which uses only valid shares. This holds already for relatively small sample sizes. Second, allowing for weak instruments and stronger direct effects of the instruments on the outcome does not change the fact that the estimators converge to oracle performance quickly. In a third set of simulations, I test the performance of my extension to multiple endogenous regressors. The simulations confirm that with increasing number of regressors, the allowed number of invalid IVs becomes lower.

Further, I apply the methods to the estimation of the effect of immigration on wages in the US as in Altonji and Card (1991) and Card (2001) and of Chinese import exposure on manufacturing employment, following Autor, Dorn, and Hanson (2013). These two

empirical examples are representative for a long series of applications in international and migration economics, which rely heavily on shift-share IVs.

The use of the new methods suggests a lower effect of immigration on wages in the US. Using data from the US, the coefficients for the estimates that do not account for endogenous shares are positive. Using the estimators which adjust for shares selected as invalid, the estimates become smaller, often even change sign and retain statistical significance, when they were significant in the standard estimations. For example the effects on high-skilled wages change from 0.52 to -0.53 ($p < 0.01$).

Among the selected there are many countries which were suspected of invalidity in the literature, such as the Philippines (Card, 2009). When using a model with lagged immigration, the standard shift-share analysis produces estimates which point in different directions than expected. When using the proposed extension, the coefficients get switched in the expected direction again. Overall, the proposed methods seem to induce a large qualitative difference and should be used as a robustness check in migration settings.

When estimating the effect of import competition on employment, a large number of instruments is selected as invalid in some settings. It is noteworthy that many of the classes which have been discussed as problematic in Autor, Dorn, and Hanson (2013) and are likely to affect estimates (GPSS) are selected as invalid.

These two applications illustrate the value of the proposed methods. The identity of chosen shares is consistent with economic intuition, because many of the origin countries and industries chosen as invalid are also discussed as being potentially problematic in the literature. This speaks for the plausibility of the outcomes of the new methods. Also, many shares which are similar to the discussed groups but were not specifically pointed out as being problematic in the literature, have been chosen by the new methods. The results can change qualitatively, when using the adjusted estimators. This shows that the methods are complementary to economic intuition and can help in designing an appropriate shift-share instrument.

This paper relates to two strands of the literature. Recent work has brought forward two ways of motivating shift-share research designs: one which is justified by quasi-random shocks and one which stresses the exclusion restriction for shares only. Borusyak, Hull, and Jaravel (2021) show consistency of the shift-share estimator when shocks are quasi-randomly assigned, conditionally on the shares. Also in this setting, Adão, Kolesár, and Morales (2019) discuss issues in inference. Unlike these two papers, GPSS put forward an interpretation of shift-share designs, according to which the shares express differential exposure to common shocks. The identification of the causal effect relies on the exclusion restriction for shares. Which setting is appropriate depends on the economic question at hand.

This chapter mainly relates to the exogenous share setting. Still, in situations in which it comes natural to think of the instrument from the perspective of random shifts, when many class-specific shifts are available these can be used to create multiple shift-share IVs or can be used as IVs directly. The selection methods then select among them instead of among the shares. For example, when imports are available for several high-income

countries, in the Autor, Dorn, and Hanson (2013) example, the shifts can be used separately in a regression where the observations are industry-specific.

The new approach that I propose is helpful when both of these approaches fail. When the shifts are not numerous or are not random, consistency as derived by Borusyak, Hull, and Jaravel (2021) does not follow. When the identifying assumption is motivated through the shares, for example through the time lags in the shares, the share exogeneity perspective is relevant. If some of these class-specific shares are subject to criticism, but the general motivation of exclusion still stands, the methods proposed here can offer interesting insights.

This chapter also relates to the literature that proposes the use of machine learning methods for causal inference (e.g. Athey and Imbens, 2019). Shift-share IV estimation does not resemble a high-dimensional problem *at first sight*, because there is only one instrument. Arguably however, all of the shares can be used as separate instruments and the need to select invalid shares substantially increases the complexity of the problem, making the use of machine learning methods appropriate. Mullainathan and Spiess (2017) have pointed out that in the context of economic research, machine learning methods lend themselves mostly to predictive tasks and less to causal inference. This paper provides a remedy for a commonly seen endogeneity problem, which threatens the reliability of causal inference in a wide range of economic studies.

2 SHIFT-SHARE INSTRUMENTAL VARIABLES AND THE EXCLUSION RESTRICTION

In this section, I present the shift-share setup and the exclusion restriction in terms of the shares. I show under which conditions the exclusion restriction is fulfilled and show a setting in which it is plausible that some shares are valid and some invalid. A discussion of indications that a some valid - some invalid setting applies concludes this section.

2.1 *Endogeneity problem and instrument*

First, consider a linear model with a constant treatment effect β:

$$y_{lt} = \mu + d_{lt}\beta + \vartheta_{lt} + \xi_{lt}, \tag{1}$$

where l indicates the location and t the time period. A discussion of the constant treatment effect assumption can be found later in the text. The outcome variable is denoted by y_{lt}, d_{lt} is the treatment, ξ_{lt} is an idiosyncratic error term with $Cov(\xi_{lt}, d_{lt}) = 0$ and ϑ_{lt} denotes unobservable shocks which might be correlated with the treatment, i.e. $Cov(\vartheta_{lt}, d_{lt}) \neq 0$. For example, the outcome variable is employment growth in a certain region and year, the independent variable is growth of the immigrant share and the unobserved shocks ϑ_{lt} are labor demand shocks which might be correlated with the growth of the immigrant share. The intercept is denoted by μ. I abstract from covariates for ease of exposition.

In model 1, the treatment variable has the structure

$$d_{lt} \equiv \sum_{j=1}^{J} z_{jlt} \cdot g_{jlt},$$

where j indicates a *class* (e.g. the industry or the origin country of migrants), z_{jlt} is the class-specific *share* in a certain region and g_{jlt} is the region-specific growth-rate (or *shift*) of that class at time t. For example, $z_{Mexico,CA,2020}$ is the share of Mexicans in California in 2020 and $g_{Mexico,CA,2020}$ is the inflow of migrants from Mexico to California in 2020. These shifts and shares are available for J classes, i.e. origin countries, in the migration example.

In many settings, d_{lt} can be subject to endogeneity problems such as correlation with unobserved shocks and reverse causality. In this model, the regressor is endogenous when $Cov(d_{lt}, \vartheta_{lt}) \neq 0$. In the migration context, Mexican migrants may have chosen to settle down in California precisely because of the high wages at destination. Part of the correlation that is measured with ordinary least-squares regressions would thus be due to migrant selection into regions.

To circumvent this problem, a shift-share approach replaces components of the treatment variable by shares and shifts which are presumably unrelated with changes of the outcome variable. For example the share of Mexicans in California relative to Mexicans in the US is replaced with the same share, at a certain base period t^0 earlier in time (say 1990), while the growth rate of Mexican immigrants in California is replaced by its equivalent at the national level. The resulting shift-share IV is

$$s_{lt} = \sum_{j=1}^{J} z_{jlt^0} \cdot g_{jt}, \tag{2}$$

where g_{jt} is the national growth rate of industry j (i.e. the shift) at time t and s_{lt} is then used to instrument for d_{lt}.

2.2 *Exclusion restriction*

The exclusion restriction is the key identifying assumption for any IV approach. In this setting, the exclusion restriction is stated in terms of shares. This is the setting proposed by GPSS. To show fulfilledness, violation and partial violation of the exclusion restriction, I set up a simple model. The structural equation is augmented by the shares z_{jlt^0}, with coefficients ϕ_j which model the direct effects on the outcome. This is the definition of validity found in Kang, Zhang, Cai, and Small (2016).

The model becomes

$$y_{lt} = d_{lt}\beta + \sum_{j=1}^{J} z_{jlt^0}\phi_j + \vartheta_{lt} + \xi_{lt}, \tag{3}$$

$$d_{lt} = s_{lt}\gamma_s + v_{lt}, \tag{4}$$

$$d_{lt} = \sum_{j=1}^{J} z_{jlt^0}\gamma_j + \varepsilon_{lt}. \tag{5}$$

Equation 4 denotes the first stage. Relevance is given when $\gamma_s \neq 0$. When all shares are to be used as instruments, separately, relevance is given when $\gamma_j \neq 0$ for all j in equation 5. This chapter focuses on the exclusion restriction and takes relevance as given. Why should this IV be relevant? The underlying idea of this instrument is that immigrants settle in regions where they find communities of earlier migrants from their same country of origin, for example because they rejoin family members or there is a network of their country of origin which eases their arrival. This is why the shift-share instrument has also been called "network", "enclave" or "past settlement instrument". The higher probability to settle in regions in which communities of their same origin country can be found creates a correlation between past and present settlement, and the instrument is relevant.

Shares might fail validity because they have a direct effect on the outcome, as measured by α_j but they might also need to be discarded because they are related to the outcome through unobservable shocks, ϑ_{lt}. To show this, I allow for a non-zero correlation between current and past unobservable shocks:

$$\vartheta_{lt} = \rho\vartheta_{lt^0} + \nu_{lt}. \tag{6}$$

Now, assume that the past unobservable shocks can be written as

$$\vartheta_{lt^0} = \sum_{j=1}^{J} z_{jlt^0}\pi_j + \epsilon_{lt}. \tag{7}$$

Then, the structural equation becomes

$$y_{lt} = d_{lt}\beta + \sum_{j=1}^{J} z_{jlt^0}\phi_j + \rho(\sum_{j=1}^{J} z_{jlt^0}\pi_j + \epsilon_{lt}) + \nu_{lt} + \xi_{lt} \tag{8}$$

$$= d_{lt}\beta + \sum_{j=1}^{J} z_{jlt^0}(\phi_j + \rho\pi_j) + u_{lt} \tag{9}$$

where $u_{lt} = \rho\epsilon_{lt} + \nu_{lt} + \xi_{lt}$ and $Cov(d_{lt}, \nu_{lt}) = Cov(d_{lt}, \epsilon_{lt}) = 0$. In order for the initial shares to be valid instruments, they should not be directly related with the outcome ($\phi_j = 0$) and they should not be related to the initial shocks ($\pi_j = 0$) or there should be no serial correlation between initial and current shocks ($\rho = 0$).

Next, I summarize the above and introduce the definition of share validity in the context of this model to state the exclusion restriction more easily. Let $\alpha_j = \phi_j + \rho\pi_j$.

Definition 1. ***Validity***
For $j = 1, ..., J$, share z_{jlt^0} is valid if $\alpha_j = 0$. If $\alpha_j \neq 0$, then z_{jlt^0} is an invalid share.

In the migration example, the first part of validity means that there is no adjustment through other factors of production. The second part means that unobserved shocks are not related to initial shares and/or these shocks are not correlated across time. The strict exclusion restriction can now be stated as

Assumption 1. ***Strict exclusion restriction:*** *All shares z_{jlt^0} are valid.*

Under the strict exclusion restriction and relevance, the shift-share IV estimator is consistent (Proposition 2 in GPSS). Note that when Assumption 1 is fulfilled, the shifts do not play a role for the validity of the instrument.

Another way to achieve consistency of the estimator is relying on random shifts. This is the setting in Borusyak, Hull, and Jaravel (2021) and Adão, Kolesár, and Morales (2019). Which of the settings should be considered is dependent on the application. Still, the methods proposed here are also applicable to the random shocks setting of Borusyak, Hull, and Jaravel (2021), when there are multiple shifts. The following sections discuss how this can be achieved.

Applications in labor and migration economics often are related to share exogeneity, because they stress that past shares are not directly related with the outcome of interest and are hence valid. Twenty-one examples for this are listed in Table 7. This list is not exhaustive. In the mentioned papers, the reader can find explicit statements that share exogeneity motivates the validity of the shift-share IV strategy.

2.3 *Violations of the exclusion restriction*

The definition of validity in the preceding section makes it clear that violations of the exclusion restriction can come from two different sources. First, a non-zero direct effect ϕ_j invalidates the shares. In the migration setting, Jaeger, Ruist, and Stuhler (2020) warn that there might be direct effects through general equilibrium adjustments. The concern is that the economy reacts dynamically to migrant inflows. If this is the case, there is a direct correlation between instrument and outcomes, through native labor, capital and other general equilibrium adjustment channels, invalidating the instrument. One way that this might apply is illustrated by Borjas (2003): if migrants choose to move to regions with persistently high wages and native workers choose to migrate in response to the immigration of foreign workers, then the effect of immigration is positively biased.

Second, when unobserved shocks today and at the initial period t^0 are correlated ($\rho \neq 0$) *and* the initial shocks are related with initial shares ($\pi_j \neq 0$), this induces a non-zero correlation between instrument and error term. A violation is plausible, because serial correlation of unobservables is typically discussed in the literature (see Table 7 and Jaeger, Ruist, and Stuhler, 2020) and initial migrants might well have been attracted by economic conditions. In principle, the bias could go in both directions because migrants might

endogenously select into regions with higher wages, or into regions with lower growth potential.

The exclusion restriction is strict in the sense that it must hold for all J shares. What looks like a single exclusion restriction in a just-identified model is in fact a set of J exclusion restrictions. Therefore, the researcher needs to feel comfortable defending the exclusion restriction for Mexicans, Cubans, Canadians, Indians and all origin countries used when constructing the IV.

In practice, it is very difficult, if not impossible to credibly uphold the strict exclusion restriction. While building an intuition about which shares are valid might be feasible, arguing that none of them had a long-term effect or was correlated with initial shocks is very restrictive. Thinking about which factors determined migrant settlement at an initial point in time makes it clear how difficult and hypothetical such an argument is destined to be. Institutional knowledge about which origin country group was mostly drawn into cities which were experiencing a boom at the time of settlement is typically unavailable. This holds true especially in settings in which a large number of countries of origin is used. Such detailed knowledge about the structural mechanisms at work is only available for very few countries, if any.

Until now, there have not been attempts to make shift-share designs robust to violations of the exclusion restriction in Assumption 1. GPSS propose computing sensitivity-to-misspecification (Rotemberg) weights, which indicate by what percentage the bias of the shift-share IV estimator changes if the bias from a certain share increases by one percent. The authors point out that one should argue prudently for the validity of shares associated with large weights. While these weights indicate the *relative* importance with which an individual invalid share contributes to the bias of the estimator, the latter can still be considerable in absolute terms, even if only shares associated with low weights violate the exclusion restriction. Therefore, it does not suffice to argue for the validity of the shares associated with the largest weights to make a case for a low bias in *absolute* terms.

2.4 *Some valid and some invalid share setting*

These reasons for violations of the strict exclusion restriction indicate that in many settings it can at best be hoped that *some but not all* shares are valid. The general share validity setup as stated by GPSS might be credible, but not for all shares.

The migration example applies to this setting with partial violation of the exclusion restriction. First, when $\rho \neq 0$, some migrant groups might be related with labor demand shocks at the base year ($\pi_j \neq 0$), while others are not. The absence of correlation with unobservable shocks is credible for some shares, because only some origin country groups might have migrated mainly because of economic reasons. This is in line with Jaeger (2007), who finds that migrants with employment visa where most responsive to economic conditions in their location choice. If the visa composition varies by origin groups, then some shares might have been driven mostly by factors orthogonal to economic conditions. Jaeger (2007) also finds that in the beginning of the 1970s the share of employment-based

visa was low. The increase of employment visa over the decades implies that origin country groups in a later base period are more likely to be invalid.

Second, there are multiple sets of shares, which vary by base year. Some base years are correlated with the current shocks. Then for some years, $\rho_0 \neq 0$, while for others $\rho_{-1} = 0$, when the correlation breaks after a few decades. Third, some origin country groups might have had long-term effects on wages, while the effects of others have worn off quickly. This might be the case when origin country shares which consisted mostly of people with family visas did not affect other factors of production in the long-term.

In applications, the discussion of single shares as potentially problematic indicates that the researchers think of a setting in which some shares are valid, while others are invalid. Another telltale sign of such a setting is when researchers report Rotemberg weights and exclude the shares with the highest weights as a robustness exercise. The questions in the application of this diagnostic are: "By how much does the bias of the estimator change, if a certain share is invalid? What happens if we assume that the most influential shares are invalid and exclude them from the estimation?" These questions imply that it is feasible that some shares are valid, while others are invalid.

In the literature, there is also evidence that such a some valid - some invalid setting is indeed the case. Tabellini (2020) raises the concern that specific origin country shares violate the exclusion restriction because Italian or Irish migrants could have chosen their city of location endogenously, based on the possibility to influence the local economy and politics. Hunt (2017) and Wozniak and Murray (2012) use adapted versions of the shift-share instrument where certain origin countries are excluded from the construction.

Arguably, the random shocks setting can be used when the strict exclusion restriction fails, but when the shocks are not numerous and random, which is often the case, this alternative approach is of little help. This offers a further setting where the some valid - some invalid IV setting applies. Several shift-share IVs can be constructed by using various push factors of emigration as shifts, such as economic, conflict- or civil liberties related variables. One might argue that economic variables are most likely to be related across countries, while political variables at origin are more likely to be unrelated with the local economic outcomes at destination. This setting is not based on the validity of shares and illustrates that the methods are in fact more widely applicable, also to the Borusyak, Hull, and Jaravel (2021) context.

3 SELECTION OF VALID IVS IN SHIFT-SHARE ESTIMATION

In this section, I introduce how to obtain modified estimators which are robust to invalid shares. I present the general procedure, the leveraged methods and extensions of these methods.

3.1 *Two-step procedure*

The idea of the procedure is to preselect valid shares beforehand with methods that will be presented in the following. I first introduce some notation. Let $\mathbf{Z}_{\mathcal{V}}$ be the matrix of valid IVs with $\mathcal{V} = \{j : \alpha_j = 0\}$ the set of valid IVs and $\hat{\mathcal{V}}$ the set of IVs *selected* as valid. Let $\mathbf{Z}_{\mathcal{I}}$ be the matrix of invalid IVs with $\mathcal{I} = \{j : \alpha_j \neq 0\}$ the set of invalid IVs and $\hat{\mathcal{I}}$ the set of IVs selected as invalid. Further, $|\mathcal{V}|$ is the number of valid and $|\mathcal{I}|$ is the number of invalid IVs.

In short, the procedure works as follows:

1. Use ALasso or CIM with $\mathbf{y}$ as outcome, $\mathbf{d}$ as exposure and instrument with the share matrix $\mathbf{Z}$. The share matrix consists of all shares that the researcher believes to be valid.[1]
2. Use shares chosen as valid (associated with $\alpha_j = 0$) for constructing the corrected IV

$$\sum_{j \in \hat{\mathcal{V}}} z_{ljt_0} \cdot g_{jt} \tag{10}$$

   and estimate the model with

   (a) 2SLS or limited-information maximum likelihood with the selected shares

   (b) the adjusted shift-share IV.

It is important that the shares selected as invalid are controlled for. The invalid shares can only be omitted from the regression, if they are uncorrelated with the valid shares. However, this is unlikely to be the case in practice. Consistency of the proposed method follows directly if the selection methods used in the first step consistently select the invalid shares.

When validity is plausible only with random shifts and there are multiple shifts, one can also apply an industry-level regression with multiple shifts and select shifts instead of shares, analogously to above. Disregarding which source of validity is put emphasis on, the preselection of variables starts with theoretical arguments. The set of shares (or shifts) selected is hence the intersection of the shares considered to be valid by the researcher and the algorithm.

3.2 *Relaxed exclusion restriction*

In this section, I start with the critical assumptions needed for identification in the adaptive Lasso (*ALasso*) by Windmeijer, Farbmacher, Davies, and Smith (2019) and the Confidence Interval Method (*CIM*) by Windmeijer, Liang, Hartwig, and Bowden (2021), that I will present in the remainder of this section.

The properties of the methods that will be leveraged to improve shift-share estimation are the so-called "oracle properties". Oracle properties mean consistent selection of invalid

[1] The outcome and exposure are denoted by vectors $\mathbf{y}$ and $\mathbf{d}$, while the matrix $\mathbf{Z}$ collects the share instruments. Generally, scalars are in lower-case, vectors are in lower-case bold and matrices in upper-case bold.

IVs and convergence in distribution to the ideal (*oracle*) estimator that uses the model under perfect knowledge about the identity of invalid IVs.
The oracle model is

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

where $\mathcal{I}$ denotes the true set of invalid IVs. The oracle estimator is

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

The oracle properties are defined as

Definition 2. ***Oracle properties***
Let $\mathbf{Z}_{\hat{\mathcal{I}}} = \mathbf{Z} \setminus \mathbf{Z}_{\hat{\mathcal{V}}}$ *with* $\mathbf{Z}_{\hat{\mathcal{I}}}$, $\mathbf{Z}_{\hat{\mathcal{V}}}$ *being the selected invalid and valid instruments respectively. Let* $\hat{\beta}_{\hat{\mathcal{V}}}$ *be the 2SLS estimator given by*

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

1. *Consistent selection of invalid IVs:* $\lim_{n\to\infty} P(\hat{\mathcal{I}} = \mathcal{I}) = 1$
2. *Convergence in distribution:* $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2_{or})$.

Under Assumptions 6 to 9 (below) and as $n \to \infty$, σ^2_{or} *is the asymptotic variance for the oracle 2SLS estimator given by*

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

In other words, if an estimator has oracle properties, it works as well as if one knew the true identity of invalid IVs.

The ALasso has oracle properties when the majority of IVs is valid. All of the IVs also need to be relevant, as noted in equation 5.

Assumption 2. ***Majority condition:*** $|\mathcal{V}| > \frac{J}{2}$.

The CIM has oracle properties when the largest group of IVs is valid. The plurality condition in Windmeijer, Liang, Hartwig, and Bowden (2021) states that the group of valid IVs is larger than any other group. A group is defined as a set of IVs associated with an estimate which asymptotically deviates from the true β by the same constant $q = \frac{\alpha_j}{\gamma_j}$. For the valid group, c is zero. Formally, the plurality exclusion restriction is

Assumption 3. ***Plurality exclusion restriction:*** $|\mathcal{V}| > max_{q\neq 0}|\{j : \frac{\alpha_j}{\gamma_j} = q\}|$.

To compare these two assumptions, consider the following example: there are five IVs. The true effect is $\beta = 1$. For three of these IVs: $\alpha_j = 0$ and hence the three IV-specific estimands are $\beta_j = \beta$, while the remaining estimands are $\beta + \frac{\alpha_j}{\gamma_j}$ with $\alpha_j \neq 0$. In this example: $\beta_1 = \beta_2 = \beta_3 = 1$ while e.g. $\beta_4 = 4$ and $\beta_5 = 5$. Clearly, the majority assumption is fulfilled. The plurality assumption is also fulfilled, because the largest group of IVs is valid. When only two IVs are valid and the third now has an estimand which is $\beta_3 = 3$, the majority is violated, because only 2/5 IVs are valid, but the plurality is still fulfilled,

because there is one valid group of two IVs and three singleton groups. In this sense, the plurality assumption can still hold even when the majority is violated.

Next, I discuss the choice of methods and introduce how the two methods work.

3.3 *Choice of methods*

Before presenting the applied methods, I discuss why I propose to use these and not other methods. The procedure builds on two methods from an emerging literature that investigates IV estimation in presence of invalid IVs. The proposed methods are the only ones which combine the following four benefits.

First, they are computationally feasible. D. W. Andrews (1999) requires to search over all possible models, which is computationally infeasible when the number of IVs is moderately large. Second, they do not require a priori knowledge about an initial set of valid IVs. Caner, X. Han, and Lee (2018) also allow for invalid IVs when a set of valid IVs is known a priori. Third, the methods do not need assumptions on the correlation of first-stage and structural parameters. Kolesár, Chetty, Friedman, Glaeser, and Imbens (2015) assume that first stage and direct effects are uncorrelated, but in applications, this assumption is rather strict. Finally, the direct effect of invalid IVs on the outcome need not be close to zero. This needs to be the case in Conley, Hansen, and Rossi (2012), where additionally prior knowledge on possible values of α is needed. The methods used in this chapter allow for arbitrarily strong direct effects. In fact, their performance even improves when the direct effects are large. In the following, I present the methods, starting with ALasso.

3.4 *Adaptive Lasso*

3.4.1 *Method*

I first present the ALasso for IV selection developed by Windmeijer, Farbmacher, Davies, and Smith (2019, WFDS). The method consists of two parts. In short, the method chooses invalid instruments to then apply 2SLS with the instruments which have been selected as invalid.

The method consists of three parts. In the first part, an initial consistent estimator β_m is obtained through the median of IV estimates of exactly identified models (C. Han, 2008). From this estimate $\hat{\beta}_m$, a plug-in estimate $\hat{\alpha}_m$ can be directly obtained. This estimator is consistent when Assumption 2 holds. The intuition for why the median estimate is consistent is the following: More than one half of IV-estimators which use only one IV at a time are consistent if a majority of IVs is valid. More than half of the points will hence converge to the same value. The median then will pick one of the consistent estimates. Why shouldn't the analysis stop here? Windmeijer, Farbmacher, Davies, and Smith (2019) show that even though it is consistent, the estimator has an asymptotic bias. Also, the limiting distribution is that of the order statistic of a normal distribution and this distribution is unknown, making inference on the parameter difficult. Moreover, the median estimate does

not use the information contained in the additional valid IVs, missing out on efficiency gains.

In the second part, the ALasso uses the initial consistent estimates $\hat{\alpha}_{m,j}$ as weights. The ALasso minimization problem is

$$\hat{\boldsymbol{\alpha}}_{ad}^{\lambda} = argmin \frac{1}{2}||\mathbf{y} - \tilde{\mathbf{Z}}\boldsymbol{\alpha}||_2^2 + \lambda_n \sum_{j=1}^{J} \frac{|\alpha_j|}{|\hat{\alpha}_{m,j}|}, \tag{11}$$

where $\tilde{\mathbf{Z}} = \mathbf{M}_{\hat{d}}\mathbf{Z}$, $\hat{\mathbf{d}}$ is the linear projection of $\mathbf{d}$ on $\mathbf{Z}$ and $\hat{\alpha}_{m,j}$ is the initial consistent estimate of α_j, directly obtained from $\hat{\beta}_m$. For a given value of the penalty parameter λ, some entries of $\boldsymbol{\alpha}$ will be shrunk to zero. In the third part, IVs associated with an $\hat{\alpha}_{ad,j}$ of zero are used as valid in 2SLS estimation and those associated with non-zero coefficients are used as controls.

In summary, the estimation procedure works as follows:

1. Compile the vector of exactly identified estimates
2. Take the median $\hat{\beta}_m$
3. Calculate $\hat{\alpha}_m$
4. Estimate α_{ad}^{λ} by AL
5. 2SLS with IVs chosen as invalid included as controls and those chosen as valid used as IVs.

The key requirement for the ALasso to have oracle properties is that it uses an initial consistent estimate. The key assumption for the ALasso to have oracle properties hence also is that the majority exclusion restriction holds.[2] As compared to the strict exclusion restriction, this assumption is already a considerable relaxation. Moreover, the ALasso having oracle properties does not depend on the different strength or the correlation of instruments.

For any given sample, the ALasso is dependent on the value of the tuning parameter λ. WFDS show that selection under the majority condition is consistent for any sequence of penalty parameters $\lambda_n \to \infty$ and $\lambda_n = o(n^{1/2})$, where n is the number of observations. They propose to use the Hansen-Sargan statistic in a stopping rule, testing at each ALasso step, following D. W. Andrews, 1999 downward testing procedure. For each selection of IVs on the ALasso-path, the J-statistic is evaluated once. The authors first specify a significance level. They propose to use $0.1/ln(n)$, following Belloni, Chen, Chernozhukov, and Hansen (2012). Successively smaller sets of valid IVs are tested. When a prespecified significance level is exceeded, the testing procedure stops.

In applications, one might be interested in including additional endogenous regressors. However, the methods proposed here do not allow for this. Therefore, I propose a simple extension of the ALasso, by using an extension of the median estimator. In the multiple endogenous regressor case, the just-identified estimates use P IVs and the estimates are

[2] According to Theorem 1 and Proposition 3 in WFDS, the ALasso has oracle properties when the majority exclusion restriction is fulfilled.

stacked into matrices. In this extension I take the median along each dimension. This gives the following vector of marginal medians:

$$\hat{\boldsymbol{\beta}}_m = \Big(med(\tilde{\boldsymbol{\beta}}_1),\ ...,med(\tilde{\boldsymbol{\beta}}_P)\Big). \tag{12}$$

The key assumption for this estimator to be consistent is that the fraction of exactly identified models, which uses valid instruments exceeds 0.5. I call this the "qualified majority condition", because the condition on the number of valid instruments, $|\mathcal{V}|$, becomes stricter than the simple majority assumption.

Assumption 4. ***Qualified majority condition*** $\frac{\binom{|\mathcal{V}|}{P}}{\binom{J}{P}} \overset{!}{>} 0.5$.

For the simulations and applications, it is important to know how many valid IVs the new condition requires in the following settings. If we fix $J = 20$, for $P = 2$ the minimum $|\mathcal{V}|$ needed to achieve an initial consistent estimate is 15 and for $P = 3$ it is 17. For $P = 2$ and $J = 38$, it is 28. These assumptions are now much more strict than the simple majority condition in WFDS. A more detailed presentation of the method and discussion of the multiple regressor setting can be found in Appendix A.2.

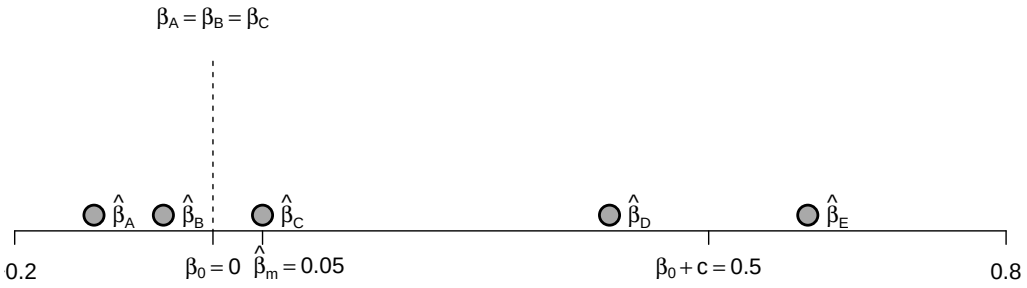
3.4.2 *Illustration*

In order to illustrate the proposed methods, consider the following toy example. Assume that one is interested in the effect of change in immigration ($\mathbf{d}$) on change in wages ($\mathbf{y}$). The parameter of interest is β in equation 8. There are cross-sectional observations from 722 commuting zones in the US. The matrix of instruments $\mathbf{Z}$ is composed of employment shares of immigrants in 1970. There are five origin countries A, B, C, D and E and hence $\mathbf{Z}$ is a 722×5 matrix. Assume that the effect of immigrants on wages is $\beta_0 = 0$. For countries A, B and C, the exclusion restriction is fulfilled and there is no direct effect of immigration on wages. However, the shares of countries D and E are invalid, because the base-period settlement of migrants from these countries was driven by economic factors and hence was not as-good-as-random. The selection of these five countries is the result of researcher's scrutiny, who ignores non-random settlement for countries D and E With ALasso, the first step is to use each share separately to obtain a vector of just-identified estimates. The IV estimates are illustrated in figure 1. The dotted, vertical line shows the true effect β_0. This effect is identified with the valid country shares A, B, and C. Let's assume that the inconsistency of the country share IV estimators D and E is $q = 0.5$. The just-identified estimates for country shares A to E are illustrated by the grey circles. In this example, the median of these estimates is $\hat{\beta}_m = \hat{\beta}_C = 0.05$. Note that if the majority exclusion is fulfilled, an IV estimator which uses a valid share is always used. From $\hat{\beta}_m$, a consistent estimate of $\boldsymbol{\alpha}$ can be obtained by

$$\hat{\boldsymbol{\alpha}}_m = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{y} - \mathbf{M}_{\hat{d}}\hat{\beta}_m).$$

This vector is $\hat{\boldsymbol{\alpha}}_m = (0.1, 0.2, -0.3, 1, 0.9)$. These estimates do not clearly indicate which of the shares is valid and which invalid, but they can be plugged into the ALasso minimization

Figure 1: Illustration



Note: This figure illustrates a setting with five origin country shares. The values on the x-axis correspond to the values of $\hat{\beta}$. The dotted line shows the true effect $\beta_0 = 0$, to which the three IV estimators which use country shares A, B and C separately, converge. Estimators which use country shares D and E converge to 0.5. The grey dots illustrate the finite sample point estimates.

problem in equation 11. For a specific λ, this gives us a new vector of $\boldsymbol{\alpha}$-estimates, $\hat{\boldsymbol{\alpha}}_{ad}$, where some entries are equal to zero, for example $\hat{\boldsymbol{\alpha}}_{ad} = (0, 0, 0, 1.2, 0.95)$. The vector which indicates which shares have been selected as invalid is $\hat{\mathcal{V}} = (0, 0, 0, 1, 1)$, where zero-entries denote valid shares. The first to third share vectors in $\mathbf{Z}$, are associated with a zero in the $\hat{\boldsymbol{\alpha}}_{ad}$-vector and are hence selected as valid, while the shares with non-zero values of $\hat{\boldsymbol{\alpha}}$ are chosen as invalid. These shares are finally used as instruments directly in the 2SLS, LIML or SSIV estimator.

Beginning with a very large value of λ, more importance is given to the second part of the ALasso minimization problem and no country is chosen as invalid, because all elements of $\boldsymbol{\alpha}$ are assigned zero. ALasso is estimated via the LARS-algorithm by Efron, Hastie, Johnstone, and Tibshirani (2004), which produces a path of models, illustrated in table 1. The lower λ gets, the more shares are chosen as invalid. The Hansen-J over-identification test is performed for each model along this path and the corresponding p-value is compared with the pre-specified significance level of $0.1/ln(722) = 0.0152$ at each step. In this illustrative example, the Hansen-test would correctly suggest to select the oracle model in column 3, with countries A, B and C selected as valid and D and E as invalid, because for this model, the p-value of the test is larger than 0.0152.

3.5 *Confidence interval method*

3.5.1 *Method*

Next, I present the CIM (Windmeijer, Liang, Hartwig, and Bowden, 2021, WLHB), which relies on an exclusion restriction which is relaxed even further. The CIM builds on Guo, Kang, Cai, and Small's (2018) hard thresholding method. Therefore, first, I introduce the hard thresholding method.

As inputs of the method, simple OLS first-stage and reduced form estimates are needed. In the first hard thresholding stage, it is tested whether the first-stage parameters are sufficiently distant (over a certain threshold) from zero. If that is the case, the individual instrument is deemed to be relevant. In the second hard thresholding stage, each individually

Table 1: Example of selection path

country \ $\lambda =$	5	2	1.6	1	0.3	0
A	0	0	0	0	0	1
B	0	0	0	0	1	1
C	0	0	0	1	1	1
D	0	1	1	1	1	1
E	0	0	1	1	1	1
Hansen (p)	0.0005	0.001	0.07	0.21	-	-

Note: This table illustrates an ALasso selection path, showing selection vector $\mathcal{V}$ for different values of penalty parameter λ.

strong instrument is excluded and used to estimate β. For each IV estimation, a plug-in estimate of the coefficient vector can be derived. Again, using a threshold, it is decided whether the instrument shall be included or whether it qualifies as a valid instrument. Now, the direct effect estimate shall lie inside a certain threshold. In this way, the estimates on the validity of a given instrument are obtained. These estimates are called "votes". If an instrument is voted by a majority of just-identified models, then it is included in the set of selected variables. The plurality rule instead includes the instruments that have received a plurality of votes. For example, with $J = 10$, the first three "experts" select five IVs as valid, while the others select less than five as valid, meaning that the direct effect estimate is below the cut-off. If that is the case, the instruments selected by the first three experts are the ones selected as valid by the plurality rule. Instruments chosen by plurality voting are such that they have received the maximal number of votes given, even if they have tied with other IVs.

The hard thresholding methods requires the choice of a threshold. Among other improvements, the CIM allows to choose this threshold more systematically. The idea behind the CIM is that IV estimators which use one valid instrument at a time converge to the same value. The method works as follows:

1. Set a critical value ψ and calculate a confidence interval (CI) for each just-identified estimate.
2. Confidence intervals are ordered by their lower endpoints.
3. Lower endpoints of CIs are compared to the upper endpoints of each CI preceding it in order. If the upper endpoint of the k-th interval is larger than the lower endpoint of the j-th interval, the estimates are said to belong to the same group. The points cil_j and ciu_j denote lower and upper endpoints of the CI when using the j-th instrument. The number of overlapping intervals when comparing from instrument j's CI downwards is $no_{[j]} = \sum_{k=1}^{j-1} 1(ciu_k > cil_j)$.
4. The largest group corresponds to the set of estimates with the most overlapping confidence intervals.

Again, the result is dependent on the value of a tuning parameter. This time ψ plays the role of the tuning parameter. For large values of ψ, all CIs will overlap and hence all variables will be chosen as valid. Gradually decreasing the value of ψ narrows the

confidence intervals down, and decreases the number of IVs chosen as valid. Analogously to ALasso, the HS test is used in a testing procedure to choose an optimal level of ψ. WLHB formally prove consistency of the Hansen-Sargan testing procedure.

The exclusion restriction needed for ALasso is stricter than the exclusion restriction needed for CIM. Why should a researcher then rely on ALasso? One advantage of ALasso over CIM is that the path of ALasso is more stable than that of CIM. A cautious researcher should use both methods and compare their results. If many IVs are chosen as invalid, suggesting a violation of the majority assumption, one should concentrate on the results of CIM.

3.5.2 *Illustration*

The CIM computes confidence intervals for each just-identified estimate and orders them by the lower endpoint of the CIs. Then it counts how often a given CI overlaps with the preceding CIs. The largest overlapping group is chosen as valid. Figure 2 illustrates the method in the toy example presented above. The second comparison, from the confidence interval of country C downwards, already selects the largest group, which includes countries A, B, and C. The other groups include only one or two IVs. In practice, the algorithm starts with a large critical value for the CIs, so that all confidence intervals overlap. Decreasing this critical value ψ produces a selection path analogous to the ALasso selection path illustrated in table 1 and the algorithm stops when a prespecified level of the Hansen-J test is exceeded. The two-stage least squares estimators adjusted by ALasso or CIM now estimate the following model:

$$\mathbf{y} = \mathbf{d}\beta + \mathbf{Z}_{DE}\boldsymbol{\alpha}_{DE} + \mathbf{u}$$

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

where the shares of people from countries D and E are additionally used as controls and the rest of the country shares are used as instruments.

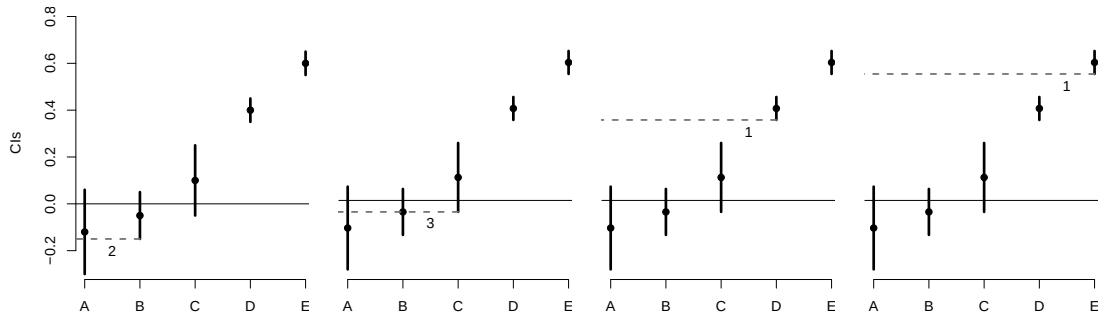
The illustration of the methods in figures 1 and 2 can also be used to understand the heterogeneous effect case. Imagine that countries A, B and C still constitute the largest group, but country shares D and E are now valid IVs which estimate effects $\beta_D = \beta_E = 0.5$ consistently. The CIM treats countries D and E as a different causal mechanism, reporting the effect of countries A, B and C.

3.6 *Discussion: Weak instruments and heterogeneous effects*

There are two main concerns with the proposed methods: weak instruments and heterogenous effects. In applications, weak IV bias is a concern, because each share is used individually to predict the endogenous variable and because the bias increases with the number of IVs.

There are three answers to this problem: First, the limited information maximum likelihood (LIML) which has better finite-sample properties than the two-stage least squares

Figure 2: Illustration of confidence interval method



Note: This figure illustrates the CIM. Black lines and dots show point estimates with intervals. Which country share is used as IV is denoted on the x-axis. Starting from the second confidence interval, the lower endpoints are compared with the upper endpoints of the CIs preceding in order. The dashed line illustrates this comparison. If the CIs overlap, they are said to belong to the same group. The number under the dashed line denotes how many IVs are in a certain group, according to the corresponding comparison. In this graph, the second panel from the left shows the largest group, which includes countries A, B, and C.

can be used after selection. Second, I also use the Anderson-Rubin test statistic in the downward testing procedure to use a test which can detect violations of the exclusion restriction in presence of weak instruments. Third, in simulations I show that the algorithms also perform well, when IVs are weak. Still, these are just practical answers to weak instruments. How to address weak instruments in valid IV selection is the object of future research.

The methods presented in this chapter rely on the constant treatment effect assumption. GPSS allow for location-specific coefficients β_l. When all the first-stage coefficients have the same sign (monotonicity), each class-specific IV estimates a weighted average of location-specific effects,

$$\underset{L\to\infty}{plim}\hat{\beta}_j = E(\omega_{lj}\beta_l), \tag{13}$$

according to Proposition 4.1 in GPSS. The exact definition of ω_{lj} can be found there. In a LATE-framework the shift-share estimator is therefore a weighted combination of class-specific weighted averages. These class-specific estimates $\hat{\beta}_j$ therefore might differ.

Can the selection methods proposed before deal with heterogeneous effects? One practical solution is to perform the analogous analyses for clusters of industries, inside of which the constant treatment effect assumption is believed to hold. If the constant effect assumption is more generally violated, another approach is needed. Combining heterogeneous treatment effects with valid IV selection is also the object of current research and an exhaustive answer is beyond the scope of this chapter. Still, a first take to this problem can look as follows.

First, note that ALasso does not allow for heterogeneous treatment effects, since the constant effect assumption is upheld throughout the paper, by WFDS. However, WLHB mention the possibility of heterogeneity in the plurality setting. The Hansen-Sargan test might also reject the Null hypothesis of valid moments when some IVs estimate local effects. In this case, the methods would treat valid IVs with heterogeneous effects as invalid and discard the information contained in them.

Following this comment, I restate the plurality assumption in the setting with heterogeneity. In brief, if the largest group of IVs is valid and associated with the same class-specific effect β_j (which implicitly is a weighted average), the remaining IVs - even when valid - are discarded. Assume a setting with class-specific heterogeneous treatment effects with the index $h \in \{1, ..., H\}$, denoted by β_h. An estimator which uses invalid IVs is inconsistent with the constant $q \in \mathbb{R}$, as before. The J exactly identified estimators converge against $R + 1$ values μ_r with

$$\hat{\beta}_j \xrightarrow{P} \mu_r = \beta_h + q, \quad r \in \{0, ..., R\}$$

Each group of IVs $\mathcal{G}_q$ now is a set whose elements produces estimators which converge against a certain μ_q:

$$\mathcal{G}_q = \{j : \hat{\beta}_j \xrightarrow{P} \mu_q = \beta_h + q\}$$
$$\mathcal{V} \equiv \mathcal{G}_0 = \{j : \hat{\beta}_j \xrightarrow{P} \beta_0\}.$$

$\mathcal{V}$ is a group of valid IVs ($q = 0$) which is associated with a specific treatment effect β_0. Other IVs may be invalid or valid with effects that differ from those in the plurality group. The plurality assumption then is

Assumption 5. ***LATE Plurality:***

$$|\mathcal{V}| > \underset{q \neq 0}{max}(|\mathcal{G}_q|).$$

In the following, I use the proposed methods in a Monte Carlo exercise and apply the methods to two real-world examples. I first reproduce the original estimates by using the standard shift-share IV which uses all shares, irrespectively of their validity. I then compare this regression with the result of the adjusted estimators, using ALasso and the CIM.

4 EXAMPLE 1: MONTE CARLO SIMULATIONS

The following simulations illustrate that ALasso and CIM select the invalid shares as invalid, when the sample is reasonably large. The adjusted estimators perform better in terms of bias as compared to the standard 2SLS estimator, which uses all the shares, irrespective of their validity. Moreover, I show that even with weak instruments good selection results are possible and that strong violations of the exclusion restriction even improve their performance. Finally, I show that with multiple regressors a higher fraction of IVs needs to be valid for the ALasso to have oracle properties.

4.1 *Single regressor*

4.1.1 *Setup*

The data is created based on the model

$$y_l = d_l\beta + \mathbf{z}_l'\boldsymbol{\alpha} + u_l \tag{14}$$

and the first stage

$$d_l = \mathbf{z}_l'\boldsymbol{\gamma} + \varepsilon_l. \tag{15}$$

The coefficient of interest β is set to 0 and the elements of the first-stage coefficient vector $\boldsymbol{\gamma}$ to 0.6. The latter is also a typical value of first-stage coefficients found in the empirical examples. To create ten shares, I draw $J = 10$ columns from a uniform distribution between 0 and 0.1. The vector of direct effects of the IVs, $\boldsymbol{\alpha}$, is set to $\boldsymbol{\alpha} = (0.2, 0.2, 0.2, 0, 0, 0, 0, 0, 0, 0)$, so that a majority of shift-share products is still valid. In a second simulation, the vector is set to $\boldsymbol{\alpha} = (0.1, 0.1, 0.2, 0.2, 0.3, 0.3, 0, 0, 0, 0)$, such that only the largest group of IVs is valid. The error terms u_l and ε_l are distributed as

$$\begin{pmatrix} u_l \\ \varepsilon_l \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right). \tag{16}$$

I vary the sample size from 400 to 6000 observations, gradually increasing by 400. The number of repetitions is 100 for each parameter combination.

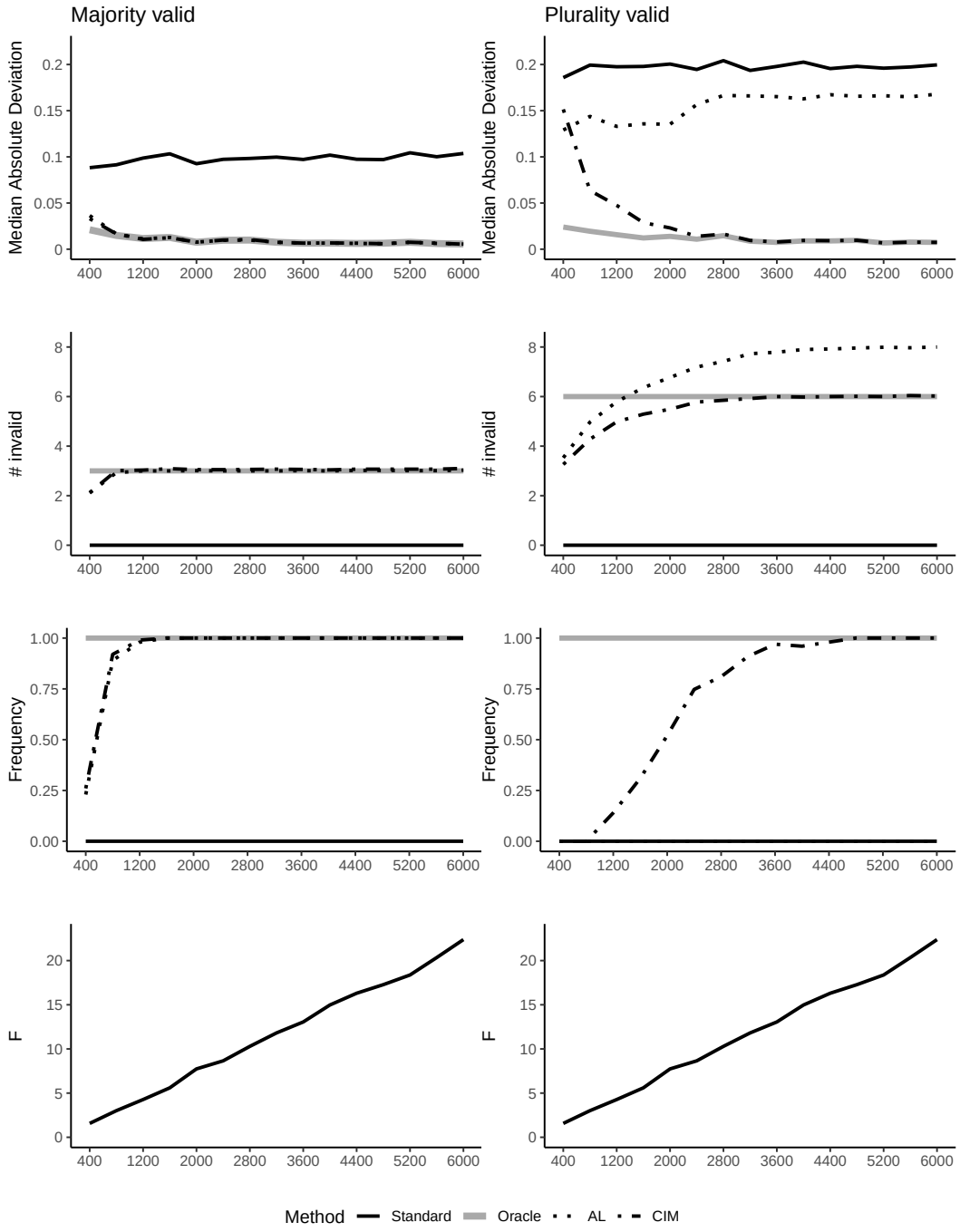
4.1.2 *Results*

The two baseline estimators are the standard and the oracle 2SLS. The standard 2SLS is the estimator for which all shares are assumed to be valid, and the oracle shift-share estimator is the ideal estimator for which only valid shares are used as IVs and invalid ones are used as controls. I compare these two estimators with the 2SLS estimator adjusted by ALasso and CIM. I report the median absolute deviation (MAD), the mean number of IVs selected as invalid, the frequency with which all invalid IVs have been selected as invalid and the F-statistic of the oracle model.

The main result is that the adjusted 2SLS estimators outperform the standard 2SLS estimator in the majority setting for all sample sizes and approach the performance of the ideal estimator with growing sample size. In the plurality setting only the CIM approaches the performance of the oracle estimator. This is in line with the predictions: When the majority rule holds, both methods should work well and ALasso is expected to fail when only the plurality rule holds but majority does not.

The graphs on the left of Figure 3 depict the setting in which a majority (seven out of ten) of shares is valid. The MAD of the standard estimator is at about 0.1 and does not decrease as the sample size gets larger. The oracle 2SLS estimator's median absolute deviation is below 0.05 and gets closer to zero as n increases. Notably, the MAD of the 2SLS adjusted by ALasso visualized by the dotted line and the 2SLS estimator adjusted by CIM, visualized by the dashed-dotted line, have a MAD equal to that of the oracle estimator (the grey, solid line) already for a moderate sample size of $N \geq 1000$. In the second and third rows, it becomes clear why that is the case. From a sample size of 1000 upwards, only three shares are chosen as invalid on average and in 100% of the cases the chosen IVs are the invalid ones, as can be seen in the third row. Whenever one appeals to

Figure 3: Simulation results



Note: Performance of 2SLS in Monte Carlo simulations as described in section 4 using shares as IVs, adjusted with ALasso and CIM. 100 replications have been used for each sample size. IVs chosen as invalid are included as controls. Horizontal axis: Number of observations. First row: median absolute deviation, second row: number of IVs chosen as invalid, third row: relative frequency with which all invalid IVs have been chosen as invalid. Grey line: oracle 2SLS, black line: standard 2SLS, dotted line: ALasso, dashed-dotted line: CIM.

the asymptotic properties of an estimator, the question of how large the sample size needs to be is legitimate. In the setting of this simulation, the adjusted estimators already attain oracle properties from a relatively low sample size on.

The graphs on the right of Figure 3 show the results from the setting in which only a plurality of shares is valid (four out of ten). Here, the estimator adjusted by ALasso fares only slightly better than the standard estimator, with a MAD at about 0.15, which does not monotonously decrease with growing sample size. On average, about eight shares are selected as invalid, as the sample gets large, but in no MC replication all invalid IVs are correctly selected as invalid. This can be seen from the right graph in row three: the dotted line and the solid black line coincide. Selection via the CIM achieves a performance which is equal to the oracle 2SLS from a sample size of about 2000 on. The average number of IVs selected as invalid reaches six when $n = 2000$ and the frequency with which all invalid are selected as invalid is close to one at $n = 3000$.

4.2 *Weak IVs and strong violations*

Next, I ask how weak instruments and stronger direct effects change the performance of the estimators. Notably, in the lowest graphs in figure 3 the F-statistic grows steadily with sample size but it is lower than 10 for many sample sizes. Still, even when F is lower than 5, the performance of ALasso and CIM is close to that of the oracle estimator. This suggests that there are settings in which weak IVs do not severely affect the selection performance.

To investigate weak instruments as well as stronger direct effects, I ran additional simulations, which can be found in the supplementary material, in Figures 6 and 7. In a nutshell, the results are that lowering the maximal F statistic in the simulations changes the results by little and increasing the entries of $\boldsymbol{\alpha}$ to make the direct effect stronger even improves performance in small-sample settings.

4.3 *Multiple regressors*

In a further simulation, I also compare the performance of the ALasso with the vector of marginal medians as an initial estimator, the standard and the oracle 2SLS estimators, but now with multiple endogenous regressors. I use 20 IVs in total and gradually increase the number of invalid instruments from zero to 18 or 17 (when $P = 3$). I expect that the performance of the ALasso breaks down at the predicted cutoffs, when there are more than ten ($P = 1$), five ($P = 2$) and three ($P = 3$) invalid IVs.

The results can be found in figure 5 in the supplementary material together with a description of the setup. As expected, the performance of the ALasso with the vector of marginal medians breaks down as soon as the majority rule is violated, i.e. when more than ten out of twenty IVs are invalid. When there are two endogenous regressors the performance of the ALasso diverges from that of the oracle estimator when there are seven or more invalid IVs. The consistency of the vector of marginal medians is guaranteed only as long as five or less IVs are invalid. In this particular setting, however, the ALasso continues performing well when six IVs are invalid. This can be the case when some just-identified estimates that use invalid IVs end up above and some below the consistent estimates, making the qualified majority assumption stricter than needed. When there are

three endogenous regressors, again the ALasso continues performing as well as the oracle as long as five or less instruments are invalid.

Overall, the results suggest that the extension of the ALasso also has oracle properties as long as sufficiently many instruments are valid, with the qualified majority condition becoming stricter with growing number of invalid instruments.

5 EXAMPLE 2: THE EFFECT OF IMMIGRATION ON WAGES

5.1 *Setting*

The second empirical application is the estimation of the effect of immigration on wages in the United States. Basso and Peri (2015) estimate the linear model[3]

$$\Delta y_{lt} = \beta \Delta immi_{lt} + a_t + u_{lt}, \tag{17}$$

with three time periods t (1990, 2000, 2010) and 722 commuting zones l. On the left hand side, Δy_{lt} stands for the three dependent variables used in separate regressions: the change in log weekly wages, and the change in log weekly wages of high- and low-skilled workers. On the right-hand-side, $\Delta immi_{lt}$ is the change in share of immigrants in total employment and β is the coefficient of interest. Decade fixed-effects are denoted by a_t and u_{lt} is the error term. Commuting zone fixed effects are accounted for by first-differencing. The authors use data from the Census Integrated Public Use Micro Samples and the American Community Survey (Ruggles et al., 2015).

As discussed before, estimating by OLS does not account for migrant sorting into regions: migrants might select into more prosperous or declining regions, creating a correlation between migrant location and outcome, which cannot be accounted to the impact of immigration. To tackle this problem, a shift-share IV, which uses origin-specific migrant shares in 1970 and changes in migrant populations, is used. The shift-share IV is

$$s_{lt} = \sum_j \left(z_{lj,1970} \cdot g_{jt} \right), \tag{18}$$

where $z_{lj,1970}$ is the share of immigrants from country j in a location l at base period 1970, and 19 origin country groups are used. The change of immigrants from country j (the shift) is denoted by g_{jt}. All of the origin countries are assumed to plausibly fulfill the exclusion restriction a priori.

The SSIV estimates are in column 1 of Table 2. The coefficient estimate of the change in log weekly wages of the natives is 0.09, but it is insignificant. The 2SLS estimate is 0.479 ($p < 0.05$) and the LIML estimate is 0.568 ($p > 0.1$). For the change in log weekly wages of the high-skilled, the coefficients are 0.35 for SSIV, 0.519 for 2SLS and 0.672 for

[3] I choose to use this paper as a reference even though it is unpublished, for the following reasons: the number of locations is large, which is helpful, because the methods I use make asymptotic arguments. Many of the published papers for which data is available have observation numbers which are often very low. For example, Card (2009), which GPSS use as an illustration, uses only 124 city-observations.

Table 2: Impact of immigration on wages

	(1) Standard	(2) AL (HS)	(3) CIM (HS)	(4) AL (HS)	(5) CIM (HS)	(6) AL (AR)	(7) CIM (AR)
Panel A: Change in average log weekly wages							
SSIV	0.0921	0.0921	0.0921	0.0921	0.0921	0.625^{+}	0.0176
	(0.438)	(0.438)	(0.438)	(0.438)	(0.438)	(0.336)	(0.438)
F	21.80	21.80	21.80	21.80	21.80	11.30	16.08
2SLS	0.479*	0.479*	0.479*	0.479*	0.479*	0.189*	-0.287
	(0.195)	(0.195)	(0.195)	(0.195)	(0.195)	(0.0798)	(0.206)
F	171.3	171.3	171.3	171.3	171.3	148.3	12.08
LIML	0.568	0.568	0.568	0.568	0.568	0.183*	-0.367
	(0.445)	(0.445)	(0.445)	(0.445)	(0.445)	(0.0809)	(0.237)
F	171.3	171.3	171.3	171.3	171.3	148.3	12.08
# inv	0	0	0	0	0	9	10
Sign.	-	0.05	0.05	0.1	0.1	0.01302	0.01302
Panel B: Change in avg. log weekly wages of high-skilled							
SSIV	0.347	0.347	0.347	0.347	0.347	1.062*	0.268
	(0.299)	(0.299)	(0.299)	(0.299)	(0.299)	(0.447)	(0.294)
F	21.80	21.80	21.80	21.80	21.80	26.17	15.15
2SLS	0.519**	0.519**	0.519**	0.519**	0.519**	0.180^{+}	-0.529**
	(0.173)	(0.173)	(0.173)	(0.173)	(0.173)	(0.0973)	(0.171)
F	171.3	171.3	171.3	171.3	171.3	6.557	22.28
LIML	0.672	0.672	0.672	0.672	0.672	0.164	-0.604**
	(0.503)	(0.503)	(0.503)	(0.503)	(0.503)	(0.103)	(0.188)
F	171.3	171.3	171.3	171.3	171.3	6.557	22.28
# inv	0	0	0	0	0	11	9
Sign.	-	0.05	0.05	0.1	0.1	0.01302	0.01302
Panel C: Change in avg. log weekly wages of low-skilled							
SSIV	-0.655^{+}	-0.655^{+}	-0.641^{+}	-0.655^{+}	-0.657^{+}	-0.721^{+}	-0.151
	(0.345)	(0.345)	(0.344)	(0.345)	(0.368)	(0.378)	(0.205)
F	21.80	21.80	19.92	21.80	16.89	13.90	9.101
2SLS	0.129	0.129	0.185	0.129	0.162	-0.0684	-0.501
	(0.160)	(0.160)	(0.189)	(0.160)	(0.188)	(0.0793)	(0.699)
F	171.3	171.3	156.9	171.3	138.6	88.74	0.650
LIML	0.0745	0.0745	0.135	0.0745	0.119	-0.0825	-4.118
	(0.226)	(0.226)	(0.324)	(0.226)	(0.274)	(0.0821)	(18.82)
F	171.3	171.3	156.9	171.3	138.6	88.74	0.650
# inv	0	0	1	0	4	9	15
Sign.	-	0.05	0.05	0.1	0.1	0.01302	0.01302

Note: This table reports estimates of β in equation 17. $N = 2166$ (722 CZ $\times$ 3). Standard errors (in parentheses) are clustered by commuting zone. First-stage F-statistics are reported. Observations are weighted by beginning-of-period population. Outcome variables are listed in the panel heads. In the first row of each panel, shift-share IV results are reported. In the second and third rows, the full vector of shares is used for 2SLS and LIML. In the last rows, the number of countries chosen as invalid and the thresholds used in the HS procedure are reported. In column (1), all shares are assumed to be valid. Column heads of columns 2 to 5 denote which method has been used for selection.

LIML. The coefficient is significant for 2SLS. For low-skilled workers, the effects on wages are negative at -0.66 and statistically significant for the shift-share analysis, and they are statistically insignificant and close to 0.1 for 2SLS and LIML. Overall, the standard estimates suggest positive effects on high-skilled and null or negative effects on low-skilled wages. The first-stage F-statistics are at 171.3 for the overidentified models and at 21.8 for the SSIV.

The possible violations in the migration context have been discussed in section 2.3. However, it is unclear whether these concerns really applied to some origin country groups and if yes to which. The remainder of this section indicates which of the shares are invalid and how large the effect is without bias.

5.2 *Results*

The results of applying ALasso and CIM on the immigration example are in Table 2. Each panel presents the results for one of the outcomes. A list of selected countries can be found in table 3. Overall, with the new methods, the coefficients of immigration decrease and often switch sign. The decrease in coefficients tends to be stronger when CIM is used in the selection step.

When choosing the significance level of $0.1/ln(N)$ (0.01375) in the downward testing procedure as proposed in WLHB, no country is chosen as invalid. Thus, the adjusted estimators are identical to the original ones (col. 1). Bowsher (2002) shows that the use of many IVs leads to low power when using the HS test. Larger significance levels of the HS test are more conservative, which is the inverse logic as with conventional tests of coefficient significance (Roodman, 2009). Hence, a more conservative strategy would be to set the threshold to a more conventional level, for example to 0.05 or 0.1. Increasing this threshold in the testing procedure leads to the selection of a few countries for low-skilled wages (Panel C), but does not change the results qualitatively.

One might be concerned that the IVs are weak and hence the HS test is unreliable. To address this concern, I also use the Anderson-Rubin test in the downward selection procedure. As a threshold I use $0.1/ln(N)$, as originally proposed in WFDS. Now, both methods select many IVs. Often, more than half of shares are selected as invalid. If a majority is invalid, ALasso in fact does not have oracle properties. I therefore rely on the CIM. The preferred analyses are hence those in the last column of table 2 for 2SLS and LIML. All estimates decrease strongly, and mostly become negative.

For overall weekly wages, the coefficients become negative and are statistically insignificant. For wages of the high-skilled, the estimates from overidentified models become negative and statistically significant when selecting via CIM, which is in stark contrast with the original results. Interestingly, the absolute size of coefficients is very similar to the original ones, with the difference that they changed direction. For wages of the low-skilled, now 15 countries are chosen as invalid by CIM. The coefficients of 2SLS and LIML are negative but none of them are significant. This might be because the F-statistic becomes low. Still, for most other analyses the F-statistics are still reasonably high.

It is reassuring to see that the differences between 2SLS and LIML estimators are smaller with the corrected as compared to the standard estimators. The remaining differences in the adjusted shift-share (SSIV), 2SLS and LIML estimates may stem from different reasons. First, LIML approximately eliminates the finite sample bias that is due to weak instruments when using 2SLS. Second, the weighting scheme of each just-identified IV estimate differs across SSIV and 2SLS. The weights shown in GPSS are dependent on the shift variable. The weighting implicit in the 2SLS estimator which uses only shares, does not take shifts into account. Therefore, different results may also arise because different methods estimate different weighted combinations of just-identified estimates.

5.3 *Results for dynamic effects*

Taking into account the critique of Jaeger, Ruist, and Stuhler (2020), who argue that using a single regressor compresses the long- and short-term effects, means to include lagged migration. The equation now becomes

$$\Delta y_{lt} = \beta_c \Delta immi_{lt} + \beta_l \Delta immi_{lt-1} + a_t + u_{lt}, \tag{19}$$

where β_c is the coefficient of interest for the contemporaneous impact and β_l denotes the coefficient of interest for the lagged impact. I include an additional shift-share IV, now using 1980 as a base period. When using 2SLS or LIML, the number of shares increases to 38 (19 per base year).

The results are shown in Table 4. The standard estimates always suggest positive effects in the short and negative effects in the long run, across estimators and outcomes. This is exactly the opposite of what Jaeger, Ruist, and Stuhler (2020) expect: partial equilibrium effects should be negative and general equilibrium adjustments are expected to offset these negative effects. These unexpected coefficient estimates might be due to the same endogeneity problems as before: both base years of the number of foreign borns might be directly correlated with the endogenous treatment.

Using the extension of the ALasso presented in Appendix A, with the Hansen-Sargan downward testing procedure, only for wages of the high-skilled the UK and Ireland are selected as invalid once and the adjusted estimates still have the same signs.

When using the Anderson-Rubin test instead, nine to eleven countries are selected for each outcome. The countries selected have a large overlap. Most variables selected come from the year 1980. I focus on 2SLS and LIML results, because the Cragg-Donald statistic of the SSIV is very low. The estimates now have the expected sign: the coefficient of contemporaneous immigration is negative and that of lagged immigration is positive.

5.4 *Overidentification from multiple shifts*

One might fundamentally question the share exogeneity interpretation of the shift-share design in the migration setting. In principle all shares could directly affect wages. If this is

Table 3: Countries selected as invalid

DV	Method	SL	CanAL	Rest of Americas	Mexico	Scandinavia and Northern Europe	UK and Ireland	China	Japan	Korea	Philippines	Vietnam	India	Western Europe	Africa	Oceania	Other	Mediterranean Countries	Central and Eastern Europe	Baltic States	Rest of Asia
Single regressor																					
dlweekly	AL	0.01302	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
dlweekly	CIM	0.01302	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
dlweekly	AL	0.05	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
dlweekly	CIM	0.05	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
dlweekly	AL	0.1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
dlweekly	CIM	0.1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
dlweekly	AL AR	0.01302	x	-	x	-	x	x	-	x	x	-	-	-	-	-	-	x	x	-	x
dlweekly	CIM AR	0.01302	-	-	-	-	-	-	x	-	x	-	x	x	x	x	x	x	x	-	x
dlweekly_hskill	AL	0.01302	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
dlweekly_hskill	CIM	0.01302	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
dlweekly_hskill	AL	0.05	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
dlweekly_hskill	CIM	0.05	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
dlweekly_hskill	AL	0.1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
dlweekly_hskill	CIM	0.1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
dlweekly_hskill	AL AR	0.01302	-	x	x	-	x	x	-	x	x	-	-	-	x	x	-	x	x	-	x
dlweekly_hskill	CIM AR	0.01302	-	-	-	-	-	-	x	-	x	-	x	x	x	x	x	-	x	-	x
dlweekly_lskill	AL	0.01302	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
dlweekly_lskill	CIM	0.01302	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
dlweekly_lskill	AL	0.05	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
dlweekly_lskill	CIM	0.05	-	-	-	-	-	-	-	-	-	-	x	-	-	-	-	-	-	-	-
dlweekly_lskill	AL	0.1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
dlweekly_lskill	CIM	0.1	-	-	-	-	-	-	-	-	x	-	-	-	x	-	-	-	x	-	x
dlweekly_lskill	AL AR	0.01302	x	-	-	-	x	x	x	x	x	-	-	-	x	-	-	-	x	-	x
dlweekly_lskill	CIM AR	0.01302	x	-	x	x	-	x	x	x	x	-	x	x	x	-	x	x	x	x	x
Multiple regressors																					
dlweekly	AL HS	0.1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
dlweekly	AL AR	1970	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	x	-	x
		1980	x	x	-	x	x	-	-	-	-	x	-	-	x	-	x	-	-	-	x
dlweekly_hskill	AL HS	1970	-	-	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		1980	-	-	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-
dlweekly_hskill	AL AR	1970	-	-	-	-	x	-	-	-	-	-	-	-	-	-	-	-	x	-	x
		1980	x	x	x	x	x	-	-	-	-	-	x	-	x	-	x	-	-	-	-
dlweekly_lskill	AL HS	0.1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
dlweekly_lskill	AL AR	1970	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	x	-	x
		1980	x	x	-	x	x	x	-	-	-	-	-	-	-	-	-	x	-	-	-

Note: This table reports the countries chosen as invalid in tables 2 and 4 for the reanalysis of Basso and Peri (2015). The left columns display the method and the outcome variable used. x denotes a country selected as invalid.

the case, the shift-share IV can be motivated via random shifts as in Borusyak, Hull, and Jaravel (2021). This offers an alternative starting point for the selection methods. This shows that the new methods are not restricted to the exogenous share world.

If validity was still a concern for all shares, one additional way to check for robustness of the results is to motivate the exclusion restriction through quasi-random shifts and to use country-of-origin specific push factors related to war, civil liberties or natural disasters. One example for such an approach is Llull (2017). The selection methods can then be used with the different shifts in an overidentified model. There would be reason to believe that some instruments are valid while others are invalid. Some shifts are related to war, others to politics, again others to other country-of-origin factors. Some might be correlated with unobservable shocks which drive wages at destination, for others it is difficult to think of a reason why that would be the case.

If multiple shifts are available, multiple shift-share instruments can be generated and used in an over-identified model. The SSIV constructed with shocks that fulfill the conditions in Borusyak, Hull, and Jaravel (2021) can then be selected. Alternatively, one could also directly use the class-level regression, and use the shocks as IVs directly. The latter approach will be used in the international trade application. Borusyak, Hull, and Jaravel (2021) show consistency of the IV-estimator, taking into account that the data is non-iid. This does not pose a challenge for the selection methods, because the key assumption is that a large-enough group of IV-specific estimators is consistent, regardless of how that consistency is established.

Moreover, selection of a particular group does not necessarily mean that the other groups are invalid instruments. Different shocks might produce heterogeneous effects. The migration inflow due to war might be different from that due to a decrease in civil liberties, which is more likely to induce migration of the elite.

To illustrate this approach, I used eleven shifts and produced eleven shift-share instruments, still using the 1970 country shares.[4] I directly estimate the dynamic model suggested by Jaeger, Ruist, and Stuhler (2020), including lagged immigration. My findings can be found in Table 8. In brief, the main results stay the same: with the HS-testing procedure, only few IVs are selected as invalid, while with the Anderson-Rubin test, more IVs are selected. With the AR testing procedure, all estimates turn negative but insignificant. This could be due to a loss in relevance, as the first-stage F-statistic becomes low. The shifts selected as invalid can be found in table 9. The variables that are selected most often are the IVs constructed with battle-related deaths, with the political freedom indicator (every analysis) and with the Press Freedom Score (five times). Since the last two express similar phenomena, it makes sense that they constitute a group.

[4] The shifts and their sources are as follows: migration (as in the preceding subsections), battle-related deaths, onesided violence and nonstate violence (Uppsala Conflict Data program, www.ucdp.uu.se), population (World Development Indicators), Civil Liberties, Political Rights, Freedom House Status (Freedom House, 2020), Polity Score (Polity V project), Press Freedom Status and Press Freedom Score (Freedom House, 2017).

5.5 *Discussion*

There are five key takeaways from the application of ALasso and CIM to the estimation of the effect of immigration on wages. First, the results from adjusted estimators suggest a strong positive bias of standard estimates. This is in line with most of the literature, that expects an upward bias. This doesn't seem to be due to weaker instruments after selection, because the first-stage statistics are still reasonably high and the use of the LIML estimator, which has better finite-sample properties in presence of weak IVs suggests the same direction of the bias.

Second, the selection of shares is consistent with economic intuition. The selection of Central and Eastern Europe (including Russia) in almost all analyses can be explained by the emigration from the Soviet Union in the 1970s and the Post-Soviet countries in the 1990s. The emigrants predominantly chose coastal cities which had large country-of-origin communities, but also cities which had experienced lasting prosperity. The conditions which have made these places attractive might be correlated over time. This makes a violation of the exclusion restriction likely. The share of migrants from the UK and Ireland, which has been picked by Tabellini (2020) as an example for possibly invalid shares is chosen nine times. When shares from multiple base years are used, mostly IVs with base year 1980 are selected. This is consistent with more job-related visa in 1980 as compared to more family-related visa 1970 and is in line with the common practice in the literature of choosing longer lags to break the possible correlation between shares and current unobservable shocks.

Third, the application shows the added value of the methods to existing econometric tools. The Rotemberg weights help understand which share's invalidity is most likely to bias results, but it does not tell the researcher whether this bias is large in absolute terms and it does not lend guidance on which country should be excluded effectively. A few of the countries flagged as potentially problematic by high weights have been selected. The Philippines have received the highest sensitivity-to-misspecification weight in GPSS. Indeed, they have been selected seven times by ALasso and CIM, and adjusting for them results in large qualitative changes of the coefficients. However, if the Rotemberg weights for some origin countries are low, their invalidity could still contribute to a large part of the inconsistency of estimators. Notably, many country groups which are not worrisome according to the top-5 Rotemberg weights, such as Central and Eastern Europe have been chosen as invalid, while some that have high weights have not been selected. This shows how the new methods can guide the selection of shares beyond the discretion of researchers.

Fourth, the fraction of shares selected as invalid can be high. A maximum of 15 out of 19, are selected as invalid suggesting that the majority assumption is likely to be violated. This suggests that the ALasso can not consistently select valid IVs in the migration setting with one regressor. Also, there is a large overlap of selected countries as invalid, by variables and methods used. This is reassuring in that it confirms that the share selection is not erratic.

Table 4: Impact of immigration

	(1) Standard	(2) AL (HS)	(3) AL (AR)	(4) Standard	(5) AL (HS)	(6) AL (AR)	(7) Standard	(8) AL (HS)	(9) AL (AR)
DV:		Wages			High-skilled			Low-skilled	
SSIV $\Delta immi_t$	5.194	5.194	3.329***	4.703	4.864	6.153	1.046	1.046	-3.011
	(6.727)	(6.727)	(0.934)	(6.469)	(6.941)	(4.689)	(0.823)	(0.823)	(10.24)
$\Delta immi_{t-10}$	-2.083	-2.083	-	-1.778	-1.838	-2.620	-0.695*	-0.695*	0.709
			1.297***						
	(2.583)	(2.583)	(0.340)	(2.485)	(2.658)	(2.708)	(0.324)	(0.324)	(3.689)
J	0.278	0.278	1.356	0.278	0.263	0.248	0.278	0.278	0.170
TSLS $\Delta immi_t$	1.126**	1.126**	-0.484	0.877*	1.249***	-0.494+	0.864**	0.864**	-0.278
	(0.371)	(0.371)	(0.399)	(0.343)	(0.367)	(0.297)	(0.273)	(0.273)	(0.183)
$\Delta immi_{t-10}$	-0.439*	-0.439*	0.510+	-0.249	-0.709**	0.411+	-	-	0.222+
							0.476***	0.476***	
	(0.180)	(0.180)	(0.288)	(0.173)	(0.250)	(0.216)	(0.124)	(0.124)	(0.126)
J	38.47	38.47	26.92	38.47	41.72	30.55	38.47	38.47	28.85
LIML $\Delta immi_t$	5.984	5.984	-3.049	10.04	5.988	-2.845	3.855	3.855	-1.197+
	(18.59)	(18.59)	(2.210)	(109.6)	(10.40)	(2.130)	(10.04)	(10.04)	(0.670)
$\Delta immi_{t-10}$	-3.253	-3.253	2.310	-5.571	-4.137	1.966	-2.202	-2.202	0.899+
	(11.01)	(11.01)	(1.582)	(64.22)	(7.492)	(1.450)	(5.946)	(5.946)	(0.498)
J	38.47	38.47	26.92	38.47	41.72	30.55	38.47	38.47	28.85
# inv	0	0	10	0	2	11	0	0	9
Sign.	-	0.1	0.01302	-	0.1	0.01302	-	0.1	0.01302

Note: This table reports estimates of β in equation 19. The number of observations is $N = 2166$. Standard errors (in parentheses) are clustered by commuting zone. J denotes the Cragg-Donald test statistic. Observations are weighted by beginning-of-period population. Outcome variables are listed in the "DV" line. In the first row of each block separated by horizontal lines, results for contemporaneous immigration are reported and in the second line, results for lagged immigration are reported. In the last two rows, the number of countries chosen as invalid and the thresholds used in the testing procedure are reported. In columns 1, 4 and 7 all shares are assumed to be valid. Column heads of columns 2, 3, 5, 6, 8 and 9 denote which method has been used for selection.

Fifth, when including lagged immigration, the coefficient estimates have the expected sign, only with the proposed ALasso adjustment. The origin-country variables selected are mostly those from the year 1980. This is consistent with Jaeger, Ruist, and Stuhler (2020), who worry that dynamic adjustments might take around ten years. In my analysis, I also use data from 1990. Hence, my analysis confirms Jaeger, Ruist, and Stuhler's (2020) result that using contemporaneous and lagged immigration can help uncover the effects of immigration. It also confirms the common practice of taking longer lags of the country-of-origin distribution to plausibly fulfill the exclusion restriction.

6 EXAMPLE 3: THE CHINA SHOCK

In this subsection, I apply the proposed methods to the estimation of the effect of import exposure on manufacturing employment in the US. I first present the original approach used by the authors, discuss instrument endogeneity issues and then present my own results.

6.1 *Setting*

Autor, Dorn, and Hanson (2013, ADH) study the impact of Chinese imports on employment in manufacturing in the US. The regression equation is

$$\Delta L_{lt}^m = \Delta IPW_{ult}\beta_1 + X_{lt}'\beta + a_t + u_{lt}, \tag{20}$$

where the left-hand side is decadal change in manufacturing employment in commuting zone l, β_1 is the coefficient of interest and ΔIPW_{ult} is import exposure, defined as $\Sigma_j z_{jlt} g_{jt}$. Here, z_{jlt} are the shares of workers in commuting zone l employed in industry j at time t and g_{jt} measures the growth of imports from China in industry j. This regression is estimated in first-differences to exclude commuting-zone fixed effects and augmented by a time dummy a_t and a set of commuting-zone-level controls, X_{lt}. The time period used ranges from 1990 to 2007 and there are 397 industry shares, indexed by four-digit SIC codes.

The endogeneity issue that affects this analysis is that both employment and imports might be correlated with unobserved shocks to US demand. To address this problem, a shift-share instrument is used, which replaces the share of workers with the same share ten years earlier and uses import exposure of other high-income countries rather than the US. ADH find a coefficient of -0.596. I report the same coefficient for the original estimate in row 1, column 1 (1,1) of table 5. When using all shares separately in a 2SLS estimation, a lower coefficient of -0.183 is found (2,1). The same model is also estimated by LIML (3,1). These are the baseline coefficients to which the adjusted estimation results will be compared.

6.2 *Results*

One might understand the analysis from the viewpoint of GPSS in the framework of a pooled exposure research design, in which employment shares capture local exposure to common import shocks. My results show which industry shares one should worry about if one chooses to rely on share validity.

ADH discuss the possible invalidity of three specific industries: the computer industry, construction materials as well as apparel, footwear and textiles. GPSS show that electronic computers display the highest sensitivity-to-misspecification weight, making the validity of this specific share especially important.

The results of the ALasso-adjusted IV estimators are presented in table 5. Using ALasso, the coefficients change by little. With the default threshold of the over-identification test at 0.01375 ($0.1/ln(N)$) as in WFDS, the test does not reject the Null hypothesis, all shares can be used for the construction of the shift-share IV and all coefficients are identical to the original estimates (column 2 of table 5). To account for the problem of too many instruments in the HS-test, I set the threshold to 0.05. Now, only one industry is selected. When excluding this industry from the construction of the instrument in column 3, the estimate is virtually unaltered.
When applying CIM, the industry chosen by ALasso and seven additional industries are selected as invalid. The estimates becomes larger in absolute terms but the CI still includes the original estimate. Hence, the application is also robust to omitting shares chosen as invalid.

When estimating the post-selection model by LIML, the estimates are very different from 2SLS. This might indicate that many IVs are weak and 2SLS is therefore biased. In

Table 5: Impact of Chinese import exposure

	(1) Standard	(2) AL (HS)	(3) AL (HS)	(4) CIM (HS)	(5) CIM (HS)	(6) AL (AR)	(7) CIM (AR)
SSIV	-0.596	-0.596	-0.603	-0.596	-0.692	-0.720	-0.924
	(0.0988)	(0.0988)	(0.0990)	(0.0988)	(0.125)	(0.0781)	(0.174)
F	47.64	47.64	47.52	47.64	37.64	71.83	28.68
2SLS	-0.183	-0.183	-0.201	-0.183	-0.173	-0.440	0.117
	(0.0419)	(0.0419)	(0.0420)	(0.0419)	(0.0485)	(0.0527)	(0.0813)
F	69.86	69.86	69.86	69.86	37.75	75.71	44.82
LIML	-2.742	-2.742	-1.032	-2.742	-7.503	-0.925	2.017
	(75.48)	(75.48)	(4.883)	(75.48)	(1119.3)	(0.355)	(3.852)
F	69.86	69.86	69.86	69.86	37.75	75.71	44.82
# inv	0	0	1	0	8	63	128
Sign.		0.01375	0.05	0.01375	0.05	0.01375	0.01375

Note: This figure reports the estimates of β_1 in equation 20. $N = 1444$ (722 CZ $\times$ 2). Column 1: Results when all shares assumed as valid, Columns 2 - 4: Endogenous shares selected by ALasso, Column 5: ALasso selection by SIC2-class, Columns 6 - 8: IV selection with CIM. First row of results: just-identified model with shift-share IV, Second row: 2SLS using all shares separately, Third row: LIML, using shares separately. Standard errors (in parentheses) are clustered by state and observations weighted by start of period CZ share of national population.

order to adjust for this, I use the Anderson-Rubin test in the downward testing procedure instead of the HS-test. When using adaptive Lasso with the AR-test, the method now selects 63 shares as invalid. When selecting with CIM and the AR-test (column 7 of table 5), even 128 shares are chosen as invalid. The estimate for SSIV now moves to -0.92, but the 95% significance interval still includes the original estimate. The 2SLS estimate becomes positive, with a coefficient of 0.12, while the coefficient of LIML is positive and large.

Borusyak, Hull, and Jaravel (2021) discuss that the shift-share IV estimator might implicitly compare manufacturing and non-manufacturing industries if the sum of manufacturing shares is not included as control. ADH include the share of non-manufacturing industries to tackle this. To also follow the former approach, I re-estimate all selected models from row 1, table 5, in table 11, in the Supplemental Material, including the sum of shares as control. In the first column, I replicate the coefficient of -0.489 from Borusyak, Hull, and Jaravel (2021, Table 1, Column 2). The selected models in this subsection with the sum of shares included as controls are reported in the following columns. Again, with ALasso and CIM, the coefficient only changes slightly.

The industries chosen as invalid are listed in table 10 of the supplementary material. These industries concord with those discussed in ADH. The first industry labeled as problematic was the computer industry. Industries belonging to electronic and computer equipment (SIC35 and 36) are among those chosen most often, constituting up to 29 percent of the shares selected as invalid. The second industry class that is discussed in ADH is related to construction. Up to 16 percent of selected shares come from industries that are associated with construction (32, 33, 34). The third industry discussed in ADH is apparel, footwear and textiles. Also, up to 16 percent of shares selected comes from these industries (22, 23, 31).

The analysis offers additional information beyond the sensitivity to misspecification illustrated by Rotemberg weights. Games, Toys and Children Vehicles as well as Household

Audio and Video Equipment have obtained the second- and third-largest Rotemberg weights in GPSS, and they have been selected by ALasso and CIM. The industry with fourth-largest Rotemberg has been selected by CIM and the one with the fifth-largest weight has been selected by both methods. However, the SIC-4 industry with the largest weight has not been selected, while numerous industries from the SIC-2 industry related to it and SIC-codes from the food sector have been selected. This shows how the proposed methods can guide share selection when constructing the instrument in expected ways but it can also inspire to think about the possible endogeneity of some of the industries.

Overall, if one believes that some shares are valid and some invalid the selection procedures single out industries which are also in harmony with the ones discussed by ADH. The results are relatively robust to the use of the new methods.

6.3 *Overidentification from multiple shifts*

If share exogeneity is not credible, one can also understand exogeneity of the shift-share instrument from a random shift perspective, as in Borusyak, Hull, and Jaravel (2021), who run an equivalent industry-level regression which uses the shift-variable as instrument. In table C4 of their paper, they use an overidentified model with all eight shifts from other high-income countries instead of the aggregated shift.[5] The common concern is usually that imports of high-income countries are correlated with unobservable shocks. The estimates for these estimations lie at roughly -0.24 for both 2SLS and LIML. It is reassuring to see that the coefficients of the two methods coincide.

I reproduce the results from this table in unreported estimations. The HS- and AR-tests do not reject at any conventional significance level, and therefore the selection algorithms select all eight shifts as valid. This robustness to the use of the new methods illustrates how in this example identification should be thought of in terms of shifts. This is in line with Borusyak, Hull, and Jaravel, 2021 and Adão, Kolesár, and Morales, 2019 interpretation of the exclusion restriction as shock exogeneity.

Researchers can also leverage a larger set of import shifts from even more countries in which they are confident that most of the shifts are valid and select shifts via ALasso and CIM. This example hence shows that even in settings where the exogenous share interpretation is controversial the two methods can be helpful.

7 CONCLUSION

This chapter proposes adjusted shift-share IV estimators which require that only a majority or plurality of shares is valid. New statistical methods are used to select invalid shares. The STATA-programs offer a simple way to apply the proposed methods.

In the migration setting, many shares are chosen as invalid and the adjusted estimates are much lower than the original ones, suggesting negative effects of immigration on wages.

[5] In this analysis, the authors add lagged sum of shares for each period as control variables. I follow this modeling choice to keep results comparable.

When including lagged migration, with the adjustments the coefficients have the expected signs. In this setting the proposed methods can be helpful for retrieving a causal estimate. In the China shock example the results are mostly robust to the use of the new methods. The results are also robust to the use of the new methods when the exclusion restriction is motivated through random shifts. In simulations I show that even in settings with weak instruments the estimators can continue to perform well. Severe violations of the exclusion restriction even improve the performance of the estimators in small-sample settings.

The methods are complementary to the recent literature on shift-share IVs. Before using them, it is important to think carefully about which source of validity is most feasible. The methods can be most helpful when researchers think about the shift-share instrument from the perspective of valid shares and whenever some shares are suspected to be directly correlated with the outcome variable. If there are many class-specific shocks, for example for multiple high-income countries, there is also scope for applying the methods in the quasi-random shocks setting. When doing so in this chapter, my conclusions do not change, qualitatively.

I conclude with two shortcomings of the methods. First, the original methods only allow for one endogenous regressor. I develop an extension of ALasso to multiple endogenous regressors which calls for stricter qualified majority assumptions. These new assumptions are confirmed in simulations. Further improvements would be to develop methods which can be readily extended to the multiple endogenous regressor case without making the exclusion restriction stricter.

Second, the validity of all shares might be a concern. Given that validity of shares relies on similar arguments, it is possible that they are all inconsistent in similar ways. In this case, consistent selection can not be guaranteed. In fact, even though the majority and plurality assumptions are considerable relaxations of the strict exclusion restriction, they are still strict. Importantly, researchers should find a set of variables whose validity can be credibly defended from a theoretical point of view. The methods proposed here complement thorough theoretical considerations and do not replace them; a convincing justification of the exclusion restriction is still imperative for the new estimators.

Appendices

A METHODOLOGICAL APPENDIX

A.1 *Additional notation*

The $n \times n$ projection matrix is $\mathbf{P}_X = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, the annihilator matrix is $\mathbf{M}_X = \mathbf{I} - \mathbf{P}_X$, $\hat{\mathbf{d}} = \mathbf{P}_Z\mathbf{d}$ are the fitted values from running a regression of the endogenous regressors on the instruments and $\tilde{\mathbf{Z}} = \mathbf{M}_{\hat{d}}\mathbf{Z}$. Define $g := |\mathcal{V}|$ as the number of valid instruments, i.e. the instruments for which $\alpha_j = 0$.

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

where $\sigma^2_{or} = \sigma^2_u \left(plim(\frac{1}{n}\hat{\mathbf{d}}'\mathbf{M}_{\mathcal{I}}\hat{\mathbf{d}})^{-1}\right)$.

A.2 *Details on adaptive Lasso with multiple endogenous regressors*

In this Appendix, I provide additional details on the ALasso.

A.2.1 *Model and assumptions*

With multiple endogenous regressors, the first stages are

$$\mathbf{d}_p = \mathbf{Z}\boldsymbol{\gamma}_p + \boldsymbol{\varepsilon}_p, \text{ for } p \in \{1, ..., P\}. \tag{21}$$

There are now P endogenous regressors $\mathbf{d}_1$, ..., $\mathbf{d}_P$, which can be subsumed in a matrix $\mathbf{D}$, J instrument vectors $\mathbf{z}_1$, ..., $\mathbf{z}_J$, which can be subsumed to a matrix $\mathbf{Z}$ and error terms $\mathbf{u}$ and $\boldsymbol{\varepsilon}_p$ for $p \in \{1, ..., P\}$ which can be correlated with $\text{cov}(\mathbf{u}, \boldsymbol{\varepsilon}_p) \neq 0$. The latter covariance measures the endogeneity of regressor $\mathbf{d}_p$. The coefficient vector of interest is $\boldsymbol{\beta}$ $(P \times 1)$. The $(J \times 1)$-vector of first-stage coefficients for regressor p is $\boldsymbol{\gamma}_p$. If we estimate all possible combinations of exactly identified models, taking P out of J instruments at a time, we get $\binom{J}{P}$ estimate vectors. Let $[j]$ be such a just-identifying IV combination. Let $\mathbf{Z}_{[j]}$ with $j \in \{1, ..., \binom{J}{P}\}$ be the $n \times P$ matrix of one combination of instruments.

Considering the model from the main text (equations 4 and 8) and following WFDS, I assume the following:

Assumption 6. ***Relevance***
For all possible combinations $[j]$, let $\boldsymbol{\gamma}_{[j]}$ be the combinations of the $[j]$-th rows of $\boldsymbol{\gamma}$. Then assume that for all $[j]$

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

Assumption 7. ***Rank assumption***

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

Assumption 8. ***Error structure***

Let $\mathbf{w}_i = (u_i \quad \boldsymbol{\varepsilon}_i')'$. *Then* $E(\mathbf{w}_i) = 0. Var(u_i) = \sigma^2_u, \quad Var(\varepsilon_{ip}) = \sigma^2_{\varepsilon_p} \quad Cov(u_i, \varepsilon_{ip}) = \sigma_{u,\varepsilon_p}$

$$e.g.\ for\ P=2 \quad E[\mathbf{w}_i\mathbf{w}_i'] = \begin{pmatrix} \sigma_u^2 & \sigma_{u,\varepsilon_1} & \sigma_{u,\varepsilon_2} \\ \sigma_{u,\varepsilon_1} & \sigma_{\varepsilon_1}^2 & \sigma_{\varepsilon_1,\varepsilon_2} \\ \sigma_{u,\varepsilon_2} & \sigma_{\varepsilon_1,\varepsilon_2} & \sigma_{\varepsilon_2}^2 \end{pmatrix} = \mathbf{\Sigma}$$

Assumption 9.

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

Assumption 10.

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

Assumptions 4, 6 and 7 imply Assumption 3 in C. Han (2008) (with the difference that in our example we abstracted from covariates). Assumptions 7, 8, 9 and 10 ensure the existence of the Hansen-Sargan test statistic. Assumption 6, is the standard relevance assumption for each just-identified model. This assumption implies, that the first stage coefficients are all non-zero and is crucial for identification.

As will be shown below, the vector of marginal medians of the matrix of estimates from exactly identified models is a consistent estimator of $\boldsymbol{\beta}$. This proof follows closely the one implied by C. Han (2008).

A.2.2 *Additional examples of the qualified majority*

For the two-regressor case, $P = 2$, it follows:

Corollary 1. *When $P = 2$ and the number of candidate instruments grows to infinity, the fraction of valid IVs has to exceed 70,7107%.*

Proof:

$$\frac{\frac{g!}{2!(g-2)!}}{\frac{J!}{2!(J-2)!}} = \frac{(g-1)g}{(J-1)J} \overset{!}{>} 0.5 \quad \Rightarrow \quad g \overset{!}{>} \frac{1+\sqrt{1+4\cdot 0.5\cdot(J-1)\cdot J}}{2}. \tag{22}$$

Multiply both sides in 22 with $1/J$ and take $\lim_{J\to\infty}$. □

To get a better feeling about how the majority condition is altered with the number of endogenous regressors in the model, we fix J at 100. The minimal g needed for oracle properties is found by plugging into 4 and sequentially decreasing g until the condition is fulfilled. For $P = 3$ the fraction of valid instruments needs to be at least 80, for $P = 4$, 85; for $P = 5$, 88; for $P = 6$, 90; for $P = 15$, 96 and for $P = 30$, 99. The relationship between number of invalid IVs and fraction of models using only valid IVs is visualized in 4.

With growing number of endogenous regressors, the number of exogenous instruments needed also grows. In the limiting case, when the number of endogenous regressors is maximal, it is equal to the number of candidate instruments. In this case, we have only

one exactly identified model. The method does not provide any benefit then, because it cannot discard any instrument, as the model would then be underidentified. Assumption 4 now becomes a consensus rule: all of the instruments need to be valid.

A.2.3 *Consistency of the vector of marginal medians*

I rewrite estimators as in the proof of Proposition A1 in WLHB. First, partition the matrix $\mathbf{Z} = (\mathbf{Z}_1 \quad \mathbf{Z}_2)$, where $\mathbf{Z}_1$ is a $n \times P$ and $\mathbf{Z}_2$ is a $n \times (J-p)$ matrix. $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1 \quad \boldsymbol{\gamma}_2)'$ 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

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

Each of the $\binom{J}{P}$ estimators can be written as

$$(\hat{\boldsymbol{\beta}}_{2SLS} \quad \hat{\boldsymbol{\gamma}}_{2SLS})' = \hat{\mathbf{H}}^{-1}\hat{\boldsymbol{\Gamma}} = \hat{\mathbf{H}}^{-1}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{D}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{u}) = \hat{\mathbf{H}}^{-1}(\hat{\boldsymbol{\gamma}}\boldsymbol{\beta} + \boldsymbol{\alpha})$$

Note that the part in parentheses is equal to

$$\begin{pmatrix} \boldsymbol{\gamma}_1\boldsymbol{\beta} + \boldsymbol{\alpha}_1 \\ \boldsymbol{\gamma}_2\boldsymbol{\beta} + \boldsymbol{\alpha}_2 \end{pmatrix}.$$

The resulting $J \times 1$ vector will be

$$plim(\hat{\boldsymbol{\beta}}_{2SLS} \quad \hat{\boldsymbol{\gamma}}_{2SLS})' = \begin{pmatrix} \boldsymbol{\beta} + \boldsymbol{\gamma}_1^{-1}\boldsymbol{\alpha}_1 \\ -\boldsymbol{\gamma}_2\boldsymbol{\gamma}_1^{-1}\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 \end{pmatrix}$$

Hence, the inconsistency of $\hat{\boldsymbol{\beta}}_{2SLS}$ is $\mathbf{q} = \boldsymbol{\gamma}_1^{-1}\boldsymbol{\alpha}_1$. $\tilde{\boldsymbol{\beta}}$ is the $\binom{J}{P} \times P$ matrix, which consists of the stacked $\hat{\boldsymbol{\beta}}_{2SLS}$ row vectors. The easiest way to compute a median for this multidimensional set of points is to take the median along each of the P dimensions separately. I have considered alternatives to the vector of marginal medians, such as the Tukey median or Oja's simplex median, but have not implemented them because they are not continuous functions. The resulting P-dimensional vector is called the vector of marginal medians:

$$\hat{\boldsymbol{\beta}}_m = \Big(med(\tilde{\boldsymbol{\beta}}_1), ...,med(\tilde{\boldsymbol{\beta}}_P)\Big). \tag{23}$$

Proposition 1. *Under assumptions 4, 6 and 7, $\hat{\boldsymbol{\beta}}_m \xrightarrow{P} \boldsymbol{\beta}_0$, where $\boldsymbol{\beta}_0$ is the true $P \times 1$ $\boldsymbol{\beta}$-vector.*

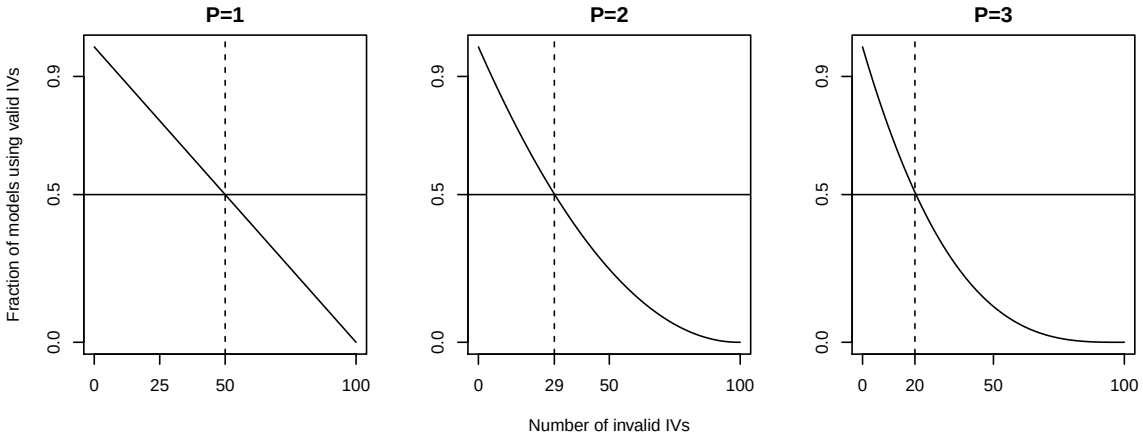
Proof: Let $\tilde{\beta}_p^{[j]}$ be the 2SLS-estimator for the p-th coefficient β_p when a certain combination $[j]$ of IVs is used. Then

$$plim(\tilde{\beta}_p^{[j]}) = \beta_0 + q_p^{[j]} \quad \forall \quad j$$

hence

$$plim\{(\tilde{\boldsymbol{\beta}}_1, \tilde{\boldsymbol{\beta}}_2, ..., \tilde{\boldsymbol{\beta}}_P)\} = (\boldsymbol{\beta}_0 + \mathbf{q}_1, \boldsymbol{\beta}_0 + \mathbf{q}_2, ..., \boldsymbol{\beta}_0 + \mathbf{q}_P)$$

Figure 4: Fraction invalid



Note: Fraction of models using valid instruments against number of valid instruments. In the left panel $P = 1$, in the middle panel $P = 2$ and in the right panel $P = 3$. The number of invalid IVs marked with the dotted line shows the maximal number of invalid IVs allowed to be present in order for the qualified majority rule to hold.

where $q_p^{[j]}$ is the p-th entry of $\mathbf{q}^{[j]} = \boldsymbol{\gamma}_1^{-1}\boldsymbol{\alpha}_1$, i.e. the inconsistency from using at least one invalid instrument in the $[j]$-set. There are $\binom{J}{P}$ entries in $\mathbf{q}_p$, which is the vector collecting inconsistency terms for a certain regressor, when using all possible combinations of IVs. The median function is continuous. Hence, by the continuous mapping theorem (CMT):

$$plim\{med(\tilde{\boldsymbol{\beta}}_1), med(\tilde{\boldsymbol{\beta}}_2), ..., med(\tilde{\boldsymbol{\beta}}_P)\} = \{med(\boldsymbol{\beta}_{0,1} + \mathbf{q}_1), ..., med(\boldsymbol{\beta}_{0,P} + \mathbf{q}_P)\}$$

Under assumption 4, $q_p^{[j]} = 0$ holds for a majority of entries inside each column. Then

$$plim\{\hat{\boldsymbol{\beta}}_m\} = \boldsymbol{\beta}_0. \quad \square$$

A.2.4 *Adaptive Lasso*

Given the consistent estimator $\boldsymbol{\beta}_m$, the following procedure is analogous to WFDS. For a detailed overview over the original ALasso method, refer to WFDS and Zou (2006). A consistent estimator for $\boldsymbol{\alpha}$ can be achieved by rewriting the moment conditions $E(\mathbf{Z}'\mathbf{u}) = 0$:

$$\hat{\boldsymbol{\alpha}}_m = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{y} - \mathbf{D}\hat{\boldsymbol{\beta}}_m). \tag{24}$$

The ALasso estimator as proposed in Zou (2006) can be used, where the penalty term is weighed by the initial consistent estimate:

$$\hat{\boldsymbol{\alpha}}_{ad}^{\lambda} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \frac{1}{2}||\mathbf{y} - \tilde{\mathbf{Z}}\boldsymbol{\alpha}||_2^2 + \lambda_n \sum_{j=1}^{J} \frac{|\alpha_j|}{|\hat{\alpha}_{m,j}|}. \tag{25}$$

The ALasso estimator $\boldsymbol{\beta}_{ad}$ is then retrieved from the conditions $E(\hat{\mathbf{D}}'\mathbf{u}) = E\Big(\hat{\mathbf{D}}'(\mathbf{y} - \hat{\mathbf{Z}}\boldsymbol{\alpha} - \hat{\mathbf{D}}\boldsymbol{\beta})\Big) = 0$

$$\boldsymbol{\beta}_{ad}^{\lambda} = \Big(\hat{\mathbf{D}}'\hat{\mathbf{D}}\Big)^{-1}\hat{\mathbf{D}}'(\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\alpha}}_{ad}^{\lambda}). \tag{26}$$

A.3 *Additional simulations*

A.3.1 *Weak instruments and strong violations*

For the additional simulations in figures 6 and 7, I use the same setup as in the main text in section 4 but I change the first-stage parameter vector $\boldsymbol{\gamma}$ and the parameter of direct effect coefficients α. Table 6 explains the new simulation settings. Lines 1 and 5 replicate the original setup, already reported in table 3.

Table 6: Additional simulations

Setting:	$\boldsymbol{\gamma}$	$\boldsymbol{\alpha}$
Majority		
Strong IV, weak violation	0.6	0.2
Weak IV, weak violation	0.3	0.2
Strong IV, strong violation	0.6	0.4
Weak IV, strong violation	0.3	0.4
Plurality		
Strong IV, weak violation	0.6	$0.1, 0.1, 0.2, 0.2, 0.3, 0.3$
Weak IV, weak violation	0.3	$0.1, 0.1, 0.2, 0.2, 0.3, 0.3$
Strong IV, strong violation	0.6	$0.2, 0.2, 0.4, 0.4, 0.6, 0.6$
Weak IV, strong violation	0.3	$0.2, 0.2, 0.4, 0.4, 0.6, 0.6$

A.3.2 *Multiple regressors*

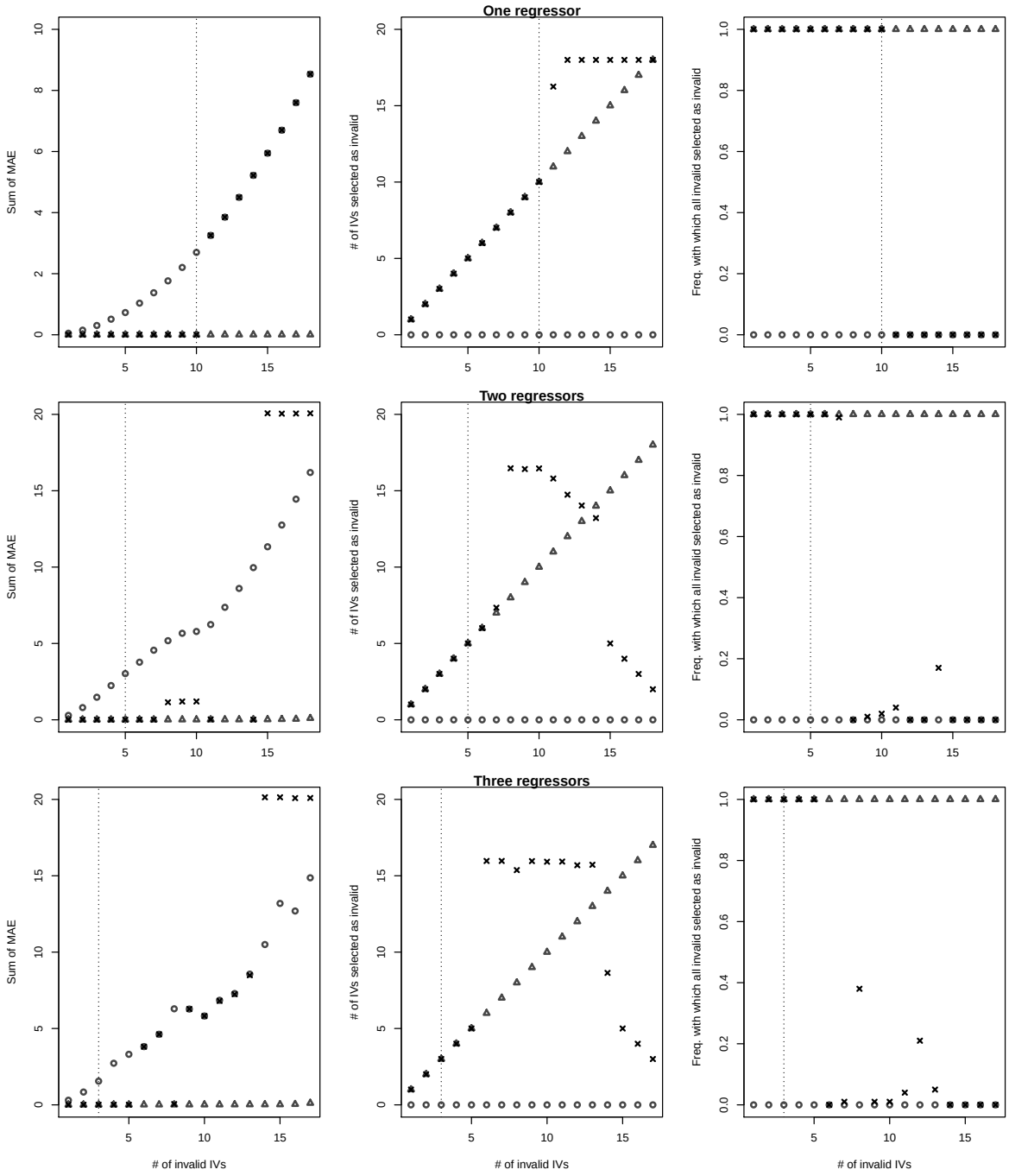
In the simulations involving multiple endogenous regressors, the following is equal in all settings: There are 100 iterations per number of invalid IVs, the sample size is $n = 10,000$, there are $J = 20$ IVs. The coefficient of interest β is set to 0. $\mathbf{Z}$ is a $N \times 20$ matrix drawn from a uniform distribution between 0 and 1. The vector indicating invalidity, $\boldsymbol{\alpha}$ in equation 8, is set to $\boldsymbol{\alpha} = (1, 0, ..., 0, 0)$, when one IV is invalid, to $\boldsymbol{\alpha} = (1, 2, 0, ..., 0, 0)$ when two IVs are invalid, etc.

The error term from the structural equation is normally distributed with $u_i \sim N(0, 0.25)$. The first-stage error terms $\boldsymbol{\varepsilon}_p$ are constructed as $\boldsymbol{\varepsilon}_p = N(0, 1) + 0.5\mathbf{u}$, obtaining the variances $Var(\boldsymbol{\varepsilon}_p) = 1.0625$, and the covariances $Cov(\mathbf{u}, \boldsymbol{\varepsilon}_p) = 0.125$.

For the case with one regressor, the first-stage coefficient vector $\boldsymbol{\gamma}$ is a matrix of ones only. When there are two endogenous regressors, the matrix of coefficients is made of two vectors. The first-stage coefficient vectors for the first and second regressors are $\boldsymbol{\gamma}_1 = (0.05, 0.1, 0.15, ..., 0.95, 1)$ and $\boldsymbol{\gamma}_2 = (1, 0.95, 0.85, ..., 0.1, 0.05)$. For $P = 3$, the matrix is composed of the two just-mentioned vectors and of an additional one: $\boldsymbol{\gamma}_3 = (0.05, 0.1, 0.15, 0.2, ..., 0.05, 0.1, 0.15, 0.2)$.

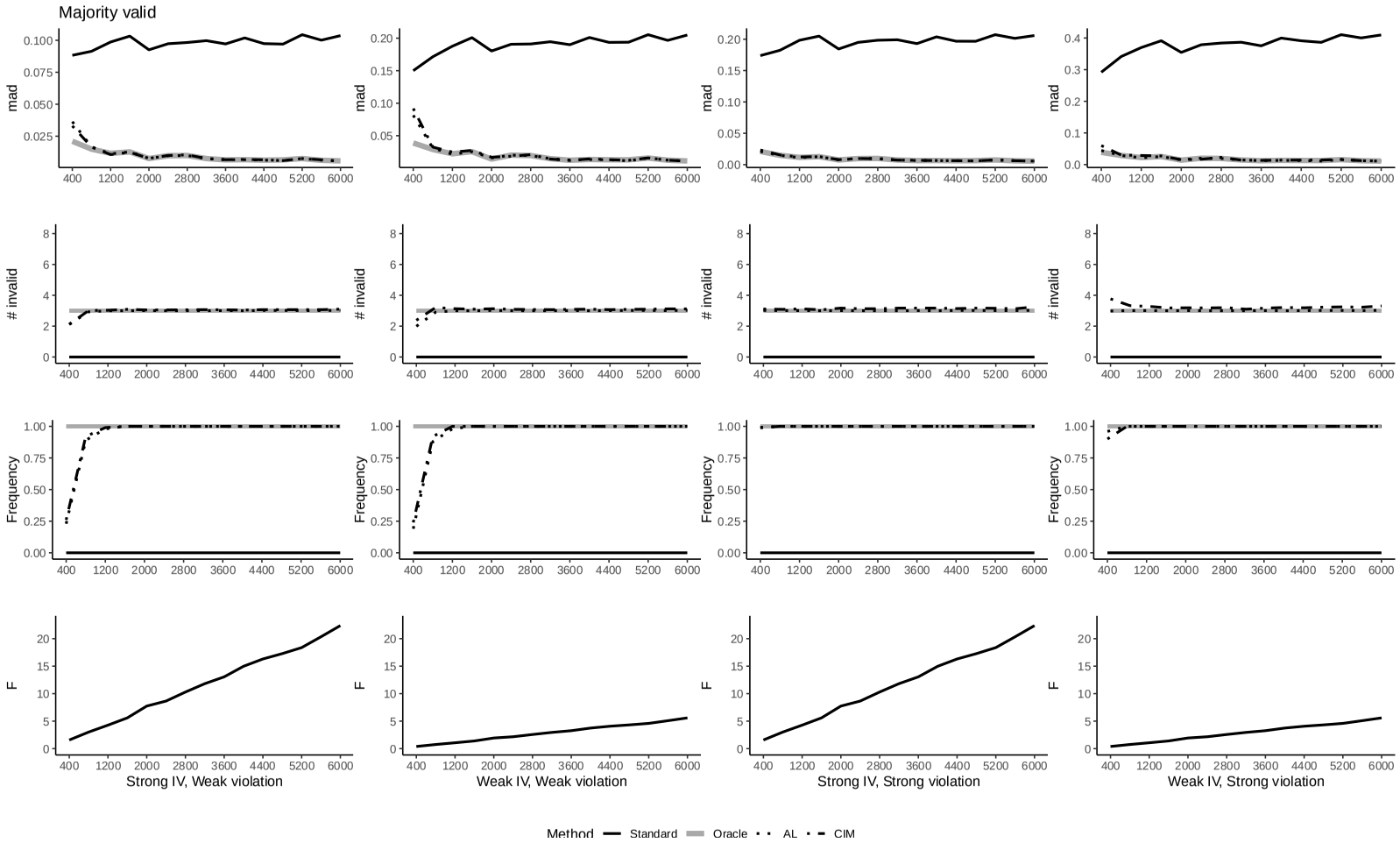
B FIGURES

Figure 5: Simulation results multiple



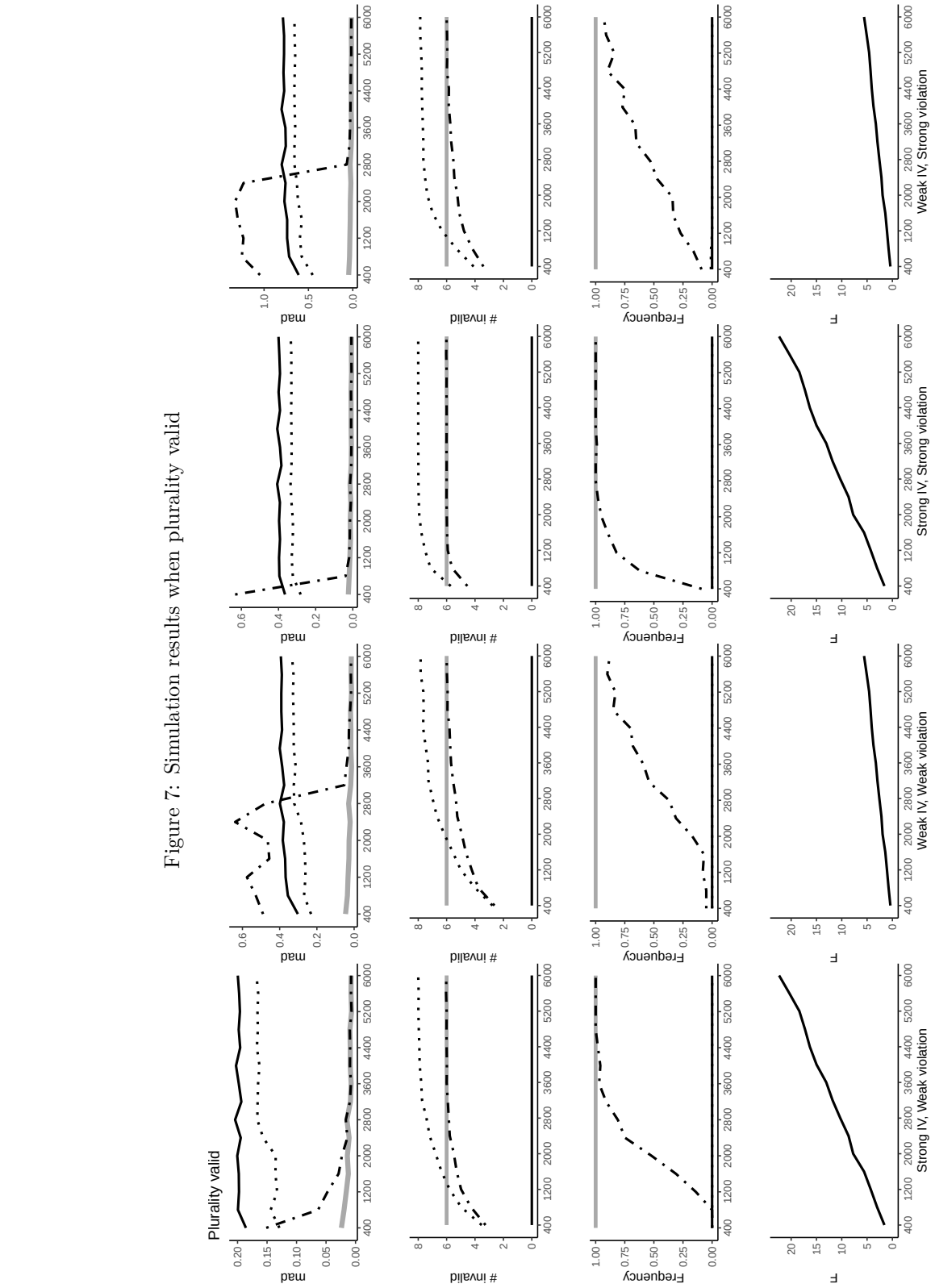
Note: Performance of standard, oracle and post-ALasso 2SLS in Monte Carlo simulations as described in section 4 using shares as IVs, with one, two and three endogenous regressors. 100 replications have been used for each number of invalid IVs. IVs chosen as invalid are included as controls. Horizontal axis: Number of invalid IVs. First column of graphs: median absolute deviation, second column: number of IVs chosen as invalid, third column: relative frequency with which all invalid IVs have been chosen as invalid. Circles: oracle 2SLS, triangles: standard 2SLS, crosses: post-ALasso 2SLS.

Figure 6: Simulation results when majority valid



Note: Performance of 2SLS in Monte Carlo simulations as described in section 4 using shares as IVs, adjusted with ALasso and CIM. 100 MC replications for each sample size. IVs chosen as invalid are included as controls. Horizontal axis: Number of observations. First row: median absolute deviation, second row: number of IVs chosen as invalid, third row: relative frequency with which all invalid IVs have been chosen as invalid. Fourth row: F-statistic of oracle model. Grey line: oracle 2SLS, black line: standard 2SLS, dotted line: post-AL 2SLS, dashed-dotted line: post-CIM 2SLS.

Figure 7: Simulation results when plurality valid



Note: Same as in figure 6, but with plurality valid.

C TABLES

Table 7: Justification of exclusion restriction in the literature - Part 1

Author	Journal	Citation
Amior (2020)	Working Paper (WP)	"The enclave instrument's validity depends on the exogeneity of the initial (origin-specific) migrant population shares (Goldsmith-Pinkham, Sorkin and Swift, 2018)."
Edo, Giesing, Öztunc, and Poutvaara (2019)	Europ. Econ. Rev. (EER)	"The identifying assumption is that the distribution of immigrants in 1968 is not correlated with voting [...]. This exclusion restriction means that, for instance, local economic shocks in 1968 are not correlated with voting more than 20 years later [...]. The assumption would be invalid if the initial distribution of immigrants is correlated with persistent local factors that influence future votes."
Bratti and Conti (2018)	Reg. Stud.	"The main identifying assumption is that [...] the between-province variation within the same NUTS-2 region in the distribution of immigrants by different nationalities in 1995 was approximately random with respect to provinces' future innovation prospects."
Aydemir and Kirdar (2017)	EER	" The validity of our instrument requires that the ratio of earlier repatriates to non-repatriates across locations be unrelated to the change in unemployment rate from 1985 to 1990 in any way other than through its effect on the number of 1989 repatriates. [...] The key concern as to the validity of our instrument is that if earlier repatriates chose their locations based on economic circumstances, we could expect their location of residence in 1985 to be related to the change in the economic conditions from 1985 to 1990 in that location."
Hunt (2017)	J. of Human Res. (JHR)	"The instrument will be invalid if nonimmigration shocks to high school completion are correlated with 1940 immigrant densities".
Foged and Peri (2016)	AEJ: Applied	" The plausibility of the exclusion restriction is predicated on the independence of the dispersal policy from labor demand conditions. " Note: the dispersal policy determines the shares.
Moreno-Galbis and Tritah (2016)	EER	"For our exclusion restriction to be valid, we require the natives' distribution within each educational group across occupations [...] to be independent from immigrants' labor supply shock."
Basso and Peri (2015)	WP	"This instrument is based on the idea that the distribution of foreign born of nationality c in CZ i in t_0 is uncorrelated with subsequent demand shifts and productivity changes in that CZ."
Bosetti, Cattaneo, and Verdolini (2015)	J. of Int. Econ.	"The underlying exclusion restriction for this instrument is that the 1991 settlement of migrants by origin is not correlated with the economic situation after 1996. "
Cattaneo, Fiorio, and Peri (2015)	JHR	" The assumption behind this instrument is that the distribution of immigrants of specific nationality across countries or occupations in 1991 is the result of historical settlements and past historical events. " " It should, therefore, be correlated with the share of foreign-born, but not with the region-sector specific demand shocks. "
Dustmann and Glitz (2015)	J. of Labor Econ. (JoLE)	"The idea is that immigrants tend to settle in areas in which other immigrants of the same country of origin have already settled earlier [...] but that these historical settlement patterns are not related to current demand-induced changes in local labor supply." "Under the plausible assumption that current regional demand-induced labor market shocks are uncorrelated with past immigrant settlement patterns, this instrument leads to estimates that have a causal interpretation."
S. P. Kerr, W. R. Kerr, and Lincoln (2015)	JoLE	The authors mention a concern about the shift-share IV: "whether the initial distribution of country groups for skilled immigrants used in the interaction is correlated with something else that affects the measured outcomes."

(continued on next page)

Justification of exclusion restriction in the literature - Part 2

Author	Journal	Citation
Orrenius and Zavodny (2015)	JoLE	"To be valid, the instrument requires assuming that the distribution of immigrants by country or region of origin across states 10 years ago is not correlated with shocks that affect the probability that natives in a state major in a STEM field 10 years later."
Peri, Shih, and Sparber (2015)	JoLE	The IV's "validity is based, in large part, on the assumption that the 1980 employment share of foreign STEM workers varied across cities because of factors related to the persistent agglomeration of foreign communities in some localities. These historical differences [...] affected the change in the supply of foreign STEM workers but were unrelated to shocks affecting city-level native wage and employment growth. [...] For example, the initial distribution of foreign STEM may be correlated with persistent city factors that influenced future labor market outcomes, resulting in omitted variables bias."
D'Amuri and Peri (2014)	J. of the Europ. Econ. Assoc. (JEEA)	"The underlying assumption is that while new immigrants tend to settle where existing immigrant communities already exist, in order to exploit ethnic networks and amenities, their historical presence is unrelated to current cell-specific changes in labor demand. [...] Current changes in labor demand have no correlation with the past presence of immigrants, which only affects the supply of labor and skills in that cell."
Dustmann, Frattini, and I. P. Preston (2013)	Rev. of Econ. Stud. (REStud)	"We instrument the change in this ratio using two alternative but closely related instruments: the 1991 ratio of immigrants to natives for each of these regions, from the Census of Population, interacted with year dummies, and four period lags of the ratio of immigrants to natives in each region from the LFS." Note that shares are used as IVs directly.
Smith (2012)	JoLE	"The exclusion restriction for this instrument requires that the composition of the immigrant population in t-1 [...] affects changes in native labor market outcomes only through its effect on changes in immigrant stocks."
Cortes and Tessada (2011)	AEJ: Applied	"The instrument will help in identifying the causal effect of immigration concentration on time use of native women as long as the following conditions hold: (1) The unobserved factors determining that more immigrants decided to locate in city i versus city i' (both cities in the same region) in 1970 are not correlated with changes in the relative economic opportunities for skilled women offered by the two cities during the 1980s and 1990s."
Farré, González, and Ortega (2011)	BE J. of Econ. Anal. & Pol.	"Our exogeneity assumption is that regional shocks to the demand for female skilled labor between 1999 and 2008 are uncorrelated with immigrant location patterns prior to 1991."
Dustmann, Fabbri, and I. Preston (2005)	Economic J.	"Pre-existing immigrant concentrations are unlikely to be correlated with current economic shocks if measured with a sufficient time lag, since existing concentrations are determined not by current economic conditions, but by historic settlement patterns of previous immigrants."
Ottaviano and Peri, 2005	NBER WP	"Since the instrument uses only the initial composition of foreign-born residents in a city and subsequent average immigration rates in the U.S. by nationality, it is not correlated with any city-specific factor that would affect actual immigration in the city during the decade. As a consequence it is by construction orthogonal to any city-specific shock to productivity, amenities and labor market conditions."

Table 8: Impact of immigration

	(1) Standard	(2) AL (HS)	(3) AL (AR)	(4) Standard	(5) AL (HS)	(6) AL (AR)	(7) Standard	(8) AL (HS)	(9) AL (AR)
DV:		Wages			High-skilled			Low-skilled	
2SLS $\Delta immi_t$	0.726***	0.290	-0.463	0.654***	0.583***	-0.341	0.626***	0.226	-0.147
	(0.151)	(0.181)	(0.555)	(0.159)	(0.0976)	(0.254)	(0.117)	(0.279)	(0.508)
$\Delta immi_{t-10}$	-0.291**	-0.378**	-0.0620	-0.168*	-0.438***	-0.139	-0.482***	-0.390*	-0.115
	(0.0889)	(0.117)	(0.324)	(0.0670)	(0.0794)	(0.158)	(0.0735)	(0.193)	(0.266)
J	15.52	7.119	5.834	15.52	14.65	4.865	15.52	7.119	5.746
LIML $\Delta immi_t$	0.954	0.154	-0.488	0.822	0.569***	-0.376	0.762*	0.180	-0.150
	(0.664)	(0.266)	(0.573)	(0.644)	(0.101)	(0.272)	(0.325)	(0.346)	(0.511)
$\Delta immi_{t-10}$	-0.344***	-0.367*	-0.0552	-0.168	-0.500***	-0.135	-0.551***	-0.389^{+}	-0.114
	(0.0942)	(0.186)	(0.333)	(0.106)	(0.108)	(0.169)	(0.0551)	(0.231)	(0.267)
J	15.52	7.119	5.834	15.52	14.65	4.865	15.52	7.119	5.746
# inv	0	3	8	0	2	7	0	3	7
Sign.	-	0.1	0.01302	-	0.1	0.01302	-	0.1	0.01302

Note: This table reports estimates of β in equation 19. The number of observations is $N = 2166$. Standard errors (in parentheses) are clustered by commuting zone. J denotes the Cragg-Donald test statistic. Observations are weighted by beginning-of-period population. Outcome variables are listed in the "DV" line. In the first row of each block separated by horizontal lines, results for contemporaneous immigration are reported and in the second line, results for lagged immigration are reported. In the last two rows, the number of countries chosen as invalid and the thresholds used in the HS procedure are reported. In columns 1, 4 and 7 all shares are assumed to be valid. Column heads of columns 2, 3, 5, 6, 8 and 9 denote which method has been used for selection.

Table 9: Shocks selected as invalid

DV	Method	SL	Migration	Battle-related deaths	Onesided violence	Nonstate violence	Population	FH Civil Liberties	FH Political	FH Status	Polity	Press Freedom Status	Press Freedom Score
dlweekly	AL (HS)	0.1	-	x	-	-	-	-	x	-	-	-	x
dlweekly	AL (AR)	0.01302	-	x	x	x	-	x	x	x	-	x	x
dlweekly_hskill	AL (HS)	0.1	-	x	-	-	-	-	x	-	-	-	-
dlweekly_hskill	AL (AR)	0.01302	-	x	x	x	-	x	x	x	-	-	x
dlweekly_lskill	AL (HS)	0.1	-	x	-	-	-	-	x	-	-	-	x
dlweekly_lskill	AL (AR)	0.01302	-	x	x	-	x	x	x	-	-	x	x

Note: This table reports the countries chosen as invalid in tables 2 and 4 for the reanalysis of Basso and Peri (2015). The left columns display the method and the outcome variable used. x denotes a shock selected as invalid.

Table 10: Industries chosen as invalid

Analysis	Table, Column	Countries / Excluded SIC codes
		China Shock
aLasso	5, 3	Broadwoven Fabric Mills, Manmade Fiber and Silk (2221)
aLasso	5, 4	(2221)
AL SIC	5,5	Ice Cream and Frozen Desserts (2024), Canned and Cured Fish and Seafoods (2091), Macaroni, Spaghetti, Vermicelli, and Noodles (2098), (2211), Women's, Misses', Children's, and Infants' Underwear and Nightwear (2341), Printing Ink (2893), Aluminum Foundries (3365), Industrial and Commercial Machinery and Equipment, Not Elsewhere Classified (3599), Household Laundry Equipment (3633), Current-Carrying Wiring Devices (3643), Household Audio and Video Equipment (3651), Communications Equipment, Not Elsewhere Classified (3669), Semiconductors and Related Devices (3674), Electronic Coils, Transformers, and Other Inductors (3677), Electronic Components, Not Elsewhere Classified (3679), Photographic Equipment and Supplies (3861)
CIM	5, 7	Pleating, Decorative and Novelty Stitching, and Tucking for the Trade (2395), Gray and Ductile Iron Foundries (3321), Ordnance and Accessories, Not Elsewhere Classified (3489), Industrial Patterns (3543), (3633), (3677), (3679), Laboratory Analytical Instruments (3826), Measuring and Controlling Devices, Not Elsewhere Classified (3829)
CIM	5, 8	Apparel and Accessories, Not Elsewhere Classified (2389), (2395), (3321), (3489), (3543), (3633), (3651), (3677), (3679), (3826), (3829), (3861)
CIM	5, 9	Food: 15, Tobacco:1, Textile mill products: 6, Apparel & Other: 13, Lumber & Wood: 7, Furniture & Textures: 5, Paper: 3, Printing & Publishing: 1, Chemical: 9, Petroleum & Coal: 2, Rubber & misc plastics: 3, Leather: 3, Stone, Clay, Glass & Concrete Products: 6, Primary Metal Industries: 9, Fabricated Metal Prdcts, Except Machinery & Transport Eqpmnt: 9, Industrial and Commercial Machinery and Computer Equipment: 19, Electronic & Other Electric Equipment: 14, Transportation Equipment: 6, Instruments & Related Products: 10, Miscellaneous Manufacturing Industries: 6

Note: This table reports industries chosen as invalid in table 5 for the reanalysis of Autor, Dorn, and Hanson (2013). The left column displays the method and the outcome variable used. The middle column reports table and column number, in which the analysis can be found. The right column displays the industries selected as invalid. For ALasso AR and CIM AR the number of 4-digit industries in each 2-digit industry are reported, for brevity.

Table 11: Impact of chinese import exposure (sum of shares as control)

	(1) Original	(2) AL (HS)	(3) AL (HS)	(4) CIM (HS)	(5) CIM (HS)	(6) AL AR	(7) CIM AR
SSIV	-0.489	-0.489	-0.503	-0.489	-0.580	-0.672	-0.786
	(0.0864)	(0.0864)	(0.0866)	(0.0864)	(0.112)	(0.0769)	(0.152)
F	46.37	46.37	45.89	46.37	37.15	80.77	33.79
# inv	0	0	1	0	8	63	128
Sign.	-	0.01375	0.05	0.01375	0.05	0.01375	0.01375

Note: This figure reports the estimates of β_1 in equation 20. $N = 1444$ (722 CZ $\times$ 2). Column 1: Results when all shares assumed as valid, Columns 2 - 4: Endogenous shares selected by ALasso, Column 5: ALasso selection by SIC2-class, Columns 6 - 8: IV selection with CIM. Sum of shares used as control in all columns. Standard errors (in parentheses) are clustered by state and observations weighted by start of period CZ share of national population.

D DOCUMENTATION FOR ADO-FILES

The following subsection provides the documentation for the `ssada` - and `sscim` - programs in Stata. **Preliminaries:** Save `ssada` and `sscim` to your personal ado-directory.

D.1 *Adaptive Lasso shift-share*

The Stata implementation of ALasso shift-share is called `ssada`. The code is a variation of `sivreg` (Farbmacher, 2017) and shares its syntax. The differences are that in `ssada` analytical weights are allowed, the standard errors to be reported in the post-ALasso regression can be chosen and locals containing valid and invalid IVs are returned. `moremata` is required.

SYNTAX

```
ssada depvar indepvars [if] [in] [aw], ///
endog(varlist) exog(varlist) id(string) [options]
```

OPTIONS

Required:

`endog`	Endogenous variable
`exog`	Exogenous controls as well as potentially endogenous shares used for construction of the shift-share IV. The shares should have the following naming: e.g. **stub**1, **stub**2, **stub**3, ...
`id`	String denoting variables by which observations are identified

Optional:

`aw`	Analytical weights (`aweight`) are allowed
`vce`	Specifies the type of standard error reported. Same as in standard `vce-option`. Default is `robust`
`c`	`real` specifying the significance level as $c/ln(n)$ for the Andrews-Hansen stopping rule. Default is 0.1

STORED RESULTS `ssada` stores the results of the last post-ALasso `ivregress`-command in **e()**. Moreover, the following macros are returned:

e(wv)	A local containing the varnames of variables chosen as valid by the ALasso algorithm
e(wi)	A local containing the varnames of variables chosen as invalid by the ALasso algorithm

D.2 *Confidence interval method shift-share*

The setup of `sscim` is adopted from `ssada` (a large part of lines 1-150).

SYNTAX

```
sscim depvar indepvars [if] [in] [aw], ///
endog(varlist) exog(varlist) ssstub(string) [options]
```

OPTIONS

Required:

`endog`	Endogenous variable
`exog`	Exogenous controls as well as potentially endogenous shares used for construction of the shift-share IV
`ssstub`	Stub of shares. The shares should have the following naming: e.g. **stub**1, **stub**2, **stub**3, ...

Optional:

`aw`	Analytical weights (`aweight`) are allowed
`vce`	Specifies the type of standard error reported. As in `vce()`-option in `ivreg2`, e.g. `vce("cluster(state)")`. Default is `"robust"`
`c`	`real` specifying the significance level as $c/ln(n)$ for the Andrews-Hansen stopping rule. Default is 0.1
`psif`	Specifies initial critical value with which confidence intervals are calculated, according to ψ =`psif`$\times\sqrt{2.01^2 * ln(N)}$. Set this larger than one if in the beginning already more than one IV are chosen as invalid. Default is 1.

STORED RESULTS `sscim` stores the results of the last post-CIM `ivregress`-command in **e()** and the following macros:

e(wv)	A local containing the varnames of variables chosen as valid by the CIM
e(wi)	A local containing the varnames of variables chosen as invalid by the CIM

POST-ESTIMATION

For both `ssada` and `sscim`, the same post-estimation results as for `ivregress` apply.

III

AGGLOMERATIVE HIERARCHICAL CLUSTERING FOR SELECTING VALID INSTRUMENTAL VARIABLES

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 pathway between them is confounded. Consistent IV estimation requires that all instruments are valid. This requires that

(a) Instruments are associated with the endogenous variables (relevance condition)

(b) 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. The key challenge therefore is how to estimate the causal effect in situations where many IVs are invalid.

In this chapter, 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. The estimator that we develop relies on the plurality rule (Guo, Kang, Cai, and Small, 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 set of true 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 set of truly valid instruments.

A prominent example for a setting with many IVs, all of which have to be valid is the estimation of the effect of immigration on wages in labor economics. To identify causal effects, researchers often rely on the lagged origin-country specific immigration pattern, measured by previous shares of immigrants. If none of the previous shares by origin country are directly or indirectly correlated with the outcome variable, the causal effect can be consistently estimated. This assumption is invoked very often 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. Here, researchers use genetic variation to estimate the causal effect of an exposure on a health-related outcome. 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).

[1] See Table 7 in the preceding chapter for a non-exhaustive list of papers in this literature.

In the applied literature, the two solutions used most often are to select valid instruments from the set of potential instruments based on economic intuition, or to 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, it is important to develop data-driven IV selection methods to select invalid instruments, when complete knowledge about the candidate instruments' validity is absent.

Previous work has tackled the IV selection problem in the single endogenous variable case. Kang, Zhang, Cai, and Small (2016) propose a selection method based on the least absolute shrinkage and selection operator (LASSO). WFDS 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, Kang, Cai, and Small (2018) propose the two-stage hard thresholding method with voting (HT) that 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 to the majority rule. Under the same identification condition, WLHB propose the CIM, which has better finite sample performance.
Our research adds to the literature in five ways:

1. We combine AHC with a traditional statistical test, the Sargan over-identification test, to yield a novel downward testing algorithm for IV selection. This new method provides the theoretical guarantee that under the plurality rule it can select the true set of valid instruments consistently, and is computationally feasible.

2. We extend the method to settings with multiple endogenous regressors. Such an extension is not available for the aforementioned methods, but it is straightforward in our setting.

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

4. We also discuss the application of our method to a setting with heterogeneous treatment effects. Importantly, we can retrieve and inspect the entire group structure, a possibility that most of the previous methods do not offer.

5. Our algorithm is computationally less complex than the CIM method. Also, the only pre-specified parameter for our algorithm is the critical value for the Sargan test, which has been well established in the existing literature to guarantee consistent selection.

We conduct Monte Carlo simulations to examine the performance of our method, and compare it with two existing methods: the hard thresholding method (HT) and the CIM. We compare with these two methods, because they also rely on the plurality rule. The simulation results show that our method achieves oracle performance in both single and multiple endogenous regressors settings in large samples when all the instruments are strong. Also, our method works well when some of the candidate instruments are weak, outperforming HT and CIM. We apply our method to the estimation of the short- and

long-run effects of immigration on wages in the US. We also provide an R-package that makes implementation of our method easy in practice.

The remainder of this chapter is structured as follows. In Section 2, we state the model and assumptions and illustrate some of the well-established properties of the 2SLS just-identified estimator. In Section 3, we describe the basic method and the algorithm when there is a single endogenous variable, and investigate its asymptotic properties. In Section 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 5, we provide Monte Carlo simulation results. In Section 6, we apply our method to estimate the effects of immigration on wages. Section 7 concludes.

2 MODEL AND ASSUMPTIONS

2.1 *Model setup*

In the following, we introduce notational conventions used throughout this 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 regressors in $\mathbf{D}$. The coefficient vector of interest is $\boldsymbol{\beta}$ $(P \times 1)$. The $J \times P$ matrix $\boldsymbol{\gamma}$ contains first-stage coefficients. Let s be the number of instruments in set of valid 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.

We start from the model setup with a single endogenous regressor, i.e. throughout Section 2 and Section 3, $P = 1$. The extension of our method to the cases with multiple endogenous regressors can be found in Section 4.1. All proofs in the Appendix are for a general number of endogenous regressors P. Following the literature on invalid IV selection (for example Kang, Zhang, Cai, and Small, 2016), we adopt the following observed data model which takes the potentially invalid instruments into account:

$$\mathbf{y} = \mathbf{d}\beta + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{u}, \tag{27}$$

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

$$\mathbf{d} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{28}$$

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 has a direct effect on the outcome variable and hence is invalid. Following Definition 1 in Guo, Kang, Cai, and Small (2018), we define a valid instrument as:

Definition 3. *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.*

This definition of validity precludes any direct relation from the instrument to the outcome, or any correlation directed towards an unobservable which then again codetermines the outcome. The set of valid IVs is denoted by $\mathcal{V}$, while the set of invalid IVs is $\mathcal{I}$.

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:

$$\mathbf{y} = \mathbf{d}\beta + \mathbf{Z}_{\mathcal{I}}\boldsymbol{\alpha}_{\mathcal{I}} + \mathbf{u} = \mathbf{X}_{\mathcal{I}}\boldsymbol{\theta}_{\mathcal{I}} + \mathbf{u}. \tag{29}$$

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

2.2 *Assumptions*

The assumptions that follow are the same as in WLHB. The first assumption makes sure that the just-identified estimators all exist.

Assumption 11. *Relevance:*

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

Assumption 12. *Rank assumption*

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

The relevance assumption is crucial for identification.

Assumption 13. *Error structure*
Let $\mathbf{w}_i = (u_i \quad \varepsilon_i)'$. *Then,* $E(\mathbf{w}_i) = 0$ *and* $E[\mathbf{w}_i\mathbf{w}_i'] = \begin{pmatrix} \sigma_u^2 & \sigma_{u,\varepsilon} \\ \sigma_{u,\varepsilon} & \sigma_\varepsilon^2 \end{pmatrix} = \boldsymbol{\Sigma}$ *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* $\boldsymbol{\Sigma}$ *are finite.*

Assumption 14.

$$plim(n^{-1}\mathbf{Z}'\mathbf{Z}) = E(\mathbf{z}_i\mathbf{z}_i') = \mathbf{Q}_{ZZ}; \quad plim(\mathbf{Z}'\mathbf{d}) = E(\mathbf{z}_i d_i) = \mathbf{Q}_{Zd}$$
$$plim(n^{-1}\mathbf{z}_i u_i) = E(\mathbf{z}_i u_i) = 0; \quad plim(n^{-1}\mathbf{z}_i\varepsilon_i) = E(\mathbf{z}_i\varepsilon_i) = 0$$
$$plim(n^{-1}\sum_{i=1}^{n}\mathbf{w}_i) = 0 \qquad plim(n^{-1}\mathbf{w}_i\mathbf{w}_i') = \boldsymbol{\Sigma}$$

Assumption 15. $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} vec(\mathbf{z}_i\mathbf{w}_i') \xrightarrow{d} N(0, \mathbf{\Sigma} \otimes \mathbf{Q}_{ZZ})$ *as* $n \to \infty$

This assumption implies conditional homoskedasticity and is made for ease of exposition. The assumptions above will be modified when there is more than one endogenous regressor. From (1) and (2), we have the outcome-instrument reduced form

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

where $\Gamma_j = \gamma_j\beta + \alpha_j$. Γ_j and γ_j are entries of the reduced-form and first-stage estimands:[2]

$$\mathbf{\Gamma} = E(\mathbf{z}_i\mathbf{z}_i')^{-1}E(\mathbf{z}_i y_i) \quad and \quad \boldsymbol{\gamma} = E(\mathbf{z}_i\mathbf{z}_i')^{-1}E(\mathbf{z}_i d_i).$$

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 instrument, and treating the remaining IVs as controls. There are J just-identified IV estimators. We write these estimators as in WLHB.

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

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

Property 1. *Properties of just-identified estimates*
Under Assumptions 11 to 15 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, Kang, Cai, and Small (2018) as:

Definition 4. *Group of instruments:*
A group of IVs is a set of IVs associated with IV-specific just-identified estimands which deviate from the true β *by a constant* $q = \frac{\alpha_j}{\gamma_j}$*:*

$$\mathcal{G}_q = \{j : \frac{\Gamma_j}{\gamma_j} = \beta + \frac{\alpha_j}{\gamma_j} = \beta + q\}.$$

where $\frac{\Gamma_j}{\gamma_j}$ is equivalent to the just-identified estimand, using the j-th IV and controlling for the rest. Then the group consisting of all valid instruments is

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

Let the number of groups be Q.

The next assumption was proposed by Guo, Kang, Cai, and Small (2018) and is the key assumption for identification. It states that among the Q groups formed by $\mathbf{z}_1, ..., \mathbf{z}_J$, the

[2] In dissonance with the remaining notation in this chapter, $\mathbf{\Gamma}$ is a vector. I use this notation, because it is conventionally used, for example in WLHB.

largest group of instruments is valid. A group is defined as above, as a set of instruments whose just-identified estimators converge to the same value $\beta + q$.

Assumption 16. *Plurality rule*

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

where the set of valid IVs is $\mathcal{V} := \mathcal{G}_0$.

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 out the Q IV groups and then select the largest group as the set of valid instruments. In this chapter, we explore clustering methods to discover the IV groups. First, we fit the general clustering framework to the IV selection problem, which is summarized in the minimisation problem in 30. 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 which combines the AHC method (Ward's method) with the Sargan test for overidentifying restrictions to select the valid instruments, which allows us to select the valid instruments without pre-specifying K.

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:

$$\hat{\mathcal{S}}(K) = \underset{\mathcal{S}}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{\hat{\beta}_j \in \mathcal{S}_k} ||\hat{\beta}_j - \bar{\mathcal{S}}_k||^2, \tag{30}$$

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

$$\hat{\mathcal{S}}_m(K) = \{\mathcal{S}(K) : |\hat{\mathcal{S}}(K)| = \max_k |\hat{\mathcal{S}}_k(K)|\}$$

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, $K = Q$, then there is a partition minimizing the sum in Equation 30. 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. $\hat{\mathcal{S}}(K) = \mathcal{S}_0$,

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

where $g(\cdot)$ is the probability limit of the goal function. 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 30, which is such that $\mathcal{S} = \mathcal{S}_0$. A necessary condition for this to hold is that $K = Q$.

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 AHC procedure proposed by Ward (1963), with the Sargan test of overidentifying restrictions. Our selection algorithm has two parts. 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 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 2). Ward's Algorithm works as follows

Algorithm 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. The total number of clusters in the beginning hence is J.*

3. ***Joining:*** *The two clusters which 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) originally 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 4.4.

After generating the clustering path by Algorithm 1, we select the set of valid instruments following Algorithm 2:

Algorithm 2. *Downward testing procedure*

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

2. *Do 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 when for the first time, the model selected by the largest cluster at some K does not get rejected by the Sargan test.*

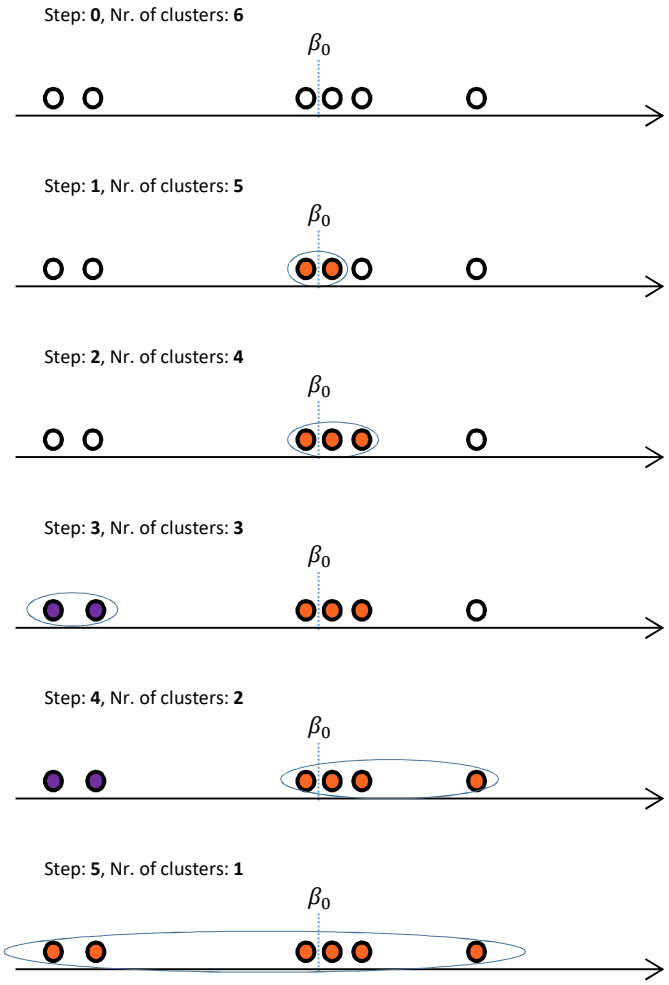
5. *Select the instruments associated with the cluster from Step 4 as valid instruments.*

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. We show later 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

Figure 8: Illustration of the algorithm with one regressor



$\xi_{n,J-|\hat{\mathcal{I}}|-P} = o(n)$. In practice, we choose the significance level $\frac{0.1}{log(n)}$ following WLHB. The procedure is illustrated in figure 8. Here, we have 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 graph the circles above the real line denote the just-identified estimate for the coefficient β_0 estimated by each of the six instruments. From left to right, we number these estimates and their corresponding instruments as No.1 to No.6.

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 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 (2,3 and 4). If this is the case, only the valid IVs are selected as valid.

3.3 *Oracle selection and estimation property*

In this section, we state the theoretical properties of the IV selection results obtained by Algorithm 1 and Algorithm 2 and the post-selection estimators. See Section 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 1. *Consistent selection*
Let ξ_n be the critical value for the Sargan test in Algorithm 2. Let $\hat{\mathcal{V}}^{dts}$ be the set of instruments selected from Algorithm 1 and Algorithm 2. Under Assumptions 11 to 16, for $\xi_n \to \infty$ and $\xi_n = o(n)$,

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

Recall from the preceding chapter, that the oracle estimator is $\hat{\beta}_{or} = \left(\hat{\mathbf{d}}'\mathbf{M}_{\mathcal{I}}\hat{\mathbf{d}}\right)^{-1}\hat{\mathbf{d}}'\mathbf{M}_{\mathcal{I}}\mathbf{y}$. 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 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

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

Under Assumptions 11 to 16, 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

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

The proof of Theorem 2 follows from the proof of Guo, Kang, Cai, and Small (2018, Consistent selection leads to oracle properties, Theorem 2).

3.4 *Computational complexity*

Recent implementations of the hierarchical agglomerative clustering algorithm have a computational cost of $O(J^2)$ (Amorim, Makarenkov, and Mirkin, 2016). In the downward

testing procedure, a maximum of $J-1$ different models needs to be tested. Therefore, the computational cost of the downward testing algorithm is $O(J^2)$. This is an improvement on the CIM which has a time complexity of $O(J^2 log(J))$ and where the maximal number of tests is $J(J-1)/2$.

4 EXTENSIONS

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

4.1 *Multiple endogenous regressors*

One shortcoming of previous methods that try to select invalid instruments is that they only allow for one endogenous regressor. Therefore, in this section we show how our method can be naturally extended to select invalid instruments when $P > 1$. First of all, the input of 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 identities 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 return to assumptions 6 to 10, from the preceding section. These are just modifications of Assumptions 11 to 15, to suit the general multiple regressor case.

The plurality assumption also needs modification for $P > 1$. For $P = 1$, Assumption 16 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. The modification starts with revisiting the asymptotics of the just-identified estimators for $P > 1$. The technical details can be found in Appendix E.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 just-identified estimates:

Property 2. *Properties of just-identified estimates with $P \geq 1$*
Under Assumptions 6 to 10 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. Because when $P > 1$, not each IV is associated with a single scalar q, we introduce the concept of a *family*:

Definition 5. ***Family:***
A family is a set of just-identifying IV combinations that is associated with just-identified estimators which converge to the same value.

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

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 which 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 automatically is a family. Analogously to Assumption 16, we assume that $\mathcal{F}_0$ is the largest family:

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

We show in Appendix E.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 16.a. *Family plurality*

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

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 $\boldsymbol{\gamma}_{[j]}^{-1}\boldsymbol{\alpha}_{[j]}$ 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 $\alpha_j \neq 0$. If the dispersion of $\boldsymbol{\gamma}_{[j]}^{-1}\boldsymbol{\alpha}_{[j]}$ 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 E.1 for an illustration) only that now we need to account for the presence of families. Firstly, 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 potentially the cluster associated with the valid instruments at K. Again, this largest cluster is $\hat{\mathcal{S}}_m(K)$.

$$\hat{\mathcal{S}}_m(K) = \{\hat{\mathcal{S}}(K) : |\hat{\mathcal{S}}(K)| = \max_k |\hat{\mathcal{S}}_k(K)|\}$$

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 $\hat{\mathcal{S}}_m(K)$. 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 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:

$$\hat{\mathcal{V}}^{Sar}(K) = \{\hat{\mathcal{V}}_m(K) : \hat{\mathcal{V}}_m(K) = max|\hat{\mathcal{V}}^M(K)| \ \ \& \ \ min\, Sar(\hat{\mathcal{V}}^M(K))\} \tag{31}$$

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:

$$\hat{\mathcal{V}}^{dts} = \{\hat{\mathcal{V}}^{Sar}(K),\, K = min(1, ..., \binom{J}{P} - 1) : Sar(\hat{\mathcal{V}}^{Sar}(K)) < \xi_{n,J-|\hat{\mathcal{I}}|-P}\} \tag{32}$$

The method has oracle properties as stated in Theorem 1 and Theorem 2. Here, we formally establish the theoretical results for the general case with an arbitrary number of regressors, $P \geq 1$. See Appendix E.4 for proofs of all theorems. Suppose Algorithm 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 that asymptotically, Algorithm 1 merges $\mathcal{S}_j$ and $\mathcal{S}_k$.

Lemma 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 6 to 10 and 16.a in Algorithm 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 1, we start from the number of clusters $K = \binom{J}{P}$. For each step onward, according to Step 3 in Algorithm 1, there would be two clusters joining with each other and forming a new cluster. Based on Lemma 1, along the path of Algorithm 1, members of different families will not be joined 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 2. *Under assumptions 6 to 10, and 16.a in steps 3 and 4 of Algorithm 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 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 1 and Corollary 2, estimators from different families are joined together only when all elements of their own family have been completely joined to their clusters. This implies that in particular 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 1 contains the valid family with probability going to 1 as there must be one step such that $K = Q$.

Corollary 3. *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 1 covers the family which uses only valid IVs, $\mathcal{F}_0$. In Appendix E.4 we show that by Algorithm 2, we can locate this $\mathcal{F}_0$ and select the valid instruments consistently. This consistent selection property is summarized in Theorem 1 which holds for $P \geq 1$ under Assumptions 6 to 10 and 16.a. These assumptions also must hold for Theorem 2 to hold.

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 6. In practice, however, these assumptions might not be valid 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 16 for *strong and valid* instruments as in Guo, Kang, Cai, and Small (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.[3]

Inherently, the AHC method can rule out weak and invalid instruments. This is because for these instruments, under Model 27 and 28, it can be shown that their just-identified estimators tend to infinity.[4] 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, when 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 a first-stage hard thresholding and simply selects all weak valid instruments as invalid, the AHC method is more flexible and instead uses the algorithm to decide which of the weak valid instruments are classified as invalid.

This mechanism has two advantages for valid weak instruments selection. Firstly, compare with the HT method which drops all such instruments. AHC can avoid loss of information as the individually weak instruments can be informative all together. Secondly, 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 biasing effect of including additional weak valid instruments on the 2SLS estimator would be small as their weights of contribution to the 2SLS estimator are small.

In comparison, the CIM can be problematic in presence of weak instruments among the candidates as it tends to select weak invalid instruments as valid, causing severe bias of the post-selection estimator. This is because the CIs of the weak invalid instruments

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

[4] 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 WLHB, 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 $n \to \infty$.

tend to have large ranges. Thus, most of them will be overlapping with all other CIs, 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 WLHB. In Section 5.2, we provide a detailed comparison via Monte Carlo experiments.

To summarize, the AHC 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 limit the bias-inducing effect of including weak instruments in IV estimation.

4.3 *Heterogeneous treatment effects*

The IV estimator also has a LATE interpretation, as estimating the average treatment effect of a sub-population, 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 treatments 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 17. *Independence* $\{y_{0i}, y_{1i}, d_{0i}, d_{1i}\} \perp\!\!\!\perp z_i$

Assumption 18. *First Stage* $P(d_i = 1|z_i = 1) \neq P(d_i = 1|z_i = 0)$

Assumption 19. *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:

$$\frac{E(y_i|z_i = 1) - E(y_i|z_i = 0)}{E(d_i|z_i = 1) - E(d_i|z_i = 0)} = E(y_{1i} - y_{0i}|d_{1i} > d_{0i}) \tag{33}$$

In the following, we show a setting in which the LATEs are dependent on a potentially unobserved variable w. 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

$$d_i = 1(h^z(w_i, z_i) > \eta_i) \tag{34}$$

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

$$y_{0i} = g_0(w_i) + v_{0i}$$
$$y_{1i} = g_1(w_i) + v_{1i}$$

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

Assumption 20. *Conditional effect ignorability:* $E(y_{1i} - y_{0i}|d^z_{1i}, d^z_{0i}, w_i) = E(y_{1i} - y_{0i}|w_i)$

The authors then show that under this assumption the LATE can be written as a function of w:

$$\beta_j = E(y_{1i} - y_{0i}|u, d_{1i} > d_{0i}) = g_1(w_i) - g_0(w_i) \tag{35}$$

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

$$\mathcal{G}_q = \{j : \beta_j = q\} \tag{36}$$

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

Note that Lemma 1 and Corollaries 2 and 3 also hold in the heterogeneous effects setting. In this case, the algorithm can find groups of heterogeneous treatment effects. Now, Algorithms 1 and 2 are altered. Instead of steps 4. and 5., in Algorithm 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 rejects. Finally, all cluster centers are reported.
In the same way as before:

Theorem 3. *Consistent selection of LATE groups*
Let ξ_n be the critical value for the Sargan test in Algorithm 2. Under Assumptions 17 - 20 and Lemma 2, 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.

4.4 *Different proximity measures*

In Algorithm 1 we have made use of the Euclidean distance to assess the proximity of clusters. Especially in settings with multiple regressors, there might be better choices to assess proximity. Aggarwal, Hinneburg, and Keim (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, the authors suggest to rely on the Manhattan distance instead of the Euclidean distance, in high dimensions. 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 consider the use of alternative distances in Algorithm 1. We consider the Manhattan and the Minkowski distance, which is similar to the fractional distance as proposed in Aggarwal, Hinneburg, and Keim (2001), with the difference that the absolute value of the distances is taken.

Furthermore, Algorithm 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 AHC algorithm, and the results are very similar to those of using the Euclidean distance and the Ward-linkage function.

5 MONTE CARLO SIMULATIONS

5.1 *All candidate instruments are strong*

We conduct Monte Carlo simulation experiments to illustrate the performance of our AHC method in IV selection and estimation, and compare with that of the existing CI and HT methods in situations where Assumption 6 is satisfied. We follow closely the setting in WLHB: There are 21 candidate instruments, 12 of which are invalid, while 9 are valid with $\boldsymbol{\alpha} = c_\alpha\,(\boldsymbol{\iota}_6',\ 0.5\boldsymbol{\iota}_6',\ \mathbf{0}_9')'$ 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 12 for sample sizes $N = 500$, $1000, 2000$ for 1000 Monte Carlo replications. We report the median absolute error (MAE) and the standard deviation (SD) of the IV estimators, and the coverage rate of the 95% confidence intervals (*Coverage*). For IV selection results we report three statistics: the number of selected invalid instruments (*# invalid*), the frequency of selecting all invalid instruments as invalid (*p allinv*) and the frequency of selecting the oracle model (*p oracle*).

Table 12: 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.

For $N = 500$, the oracle 2SLS estimator (*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 at 0.929. The naive 2SLS estimator (naive) which treats all candidates instruments as valid irrespective of their validity, however, has a much larger median absolute error of about 1.056 and its 95 % CI never covers the true value. This does not change even when increasing the sample size to 2000, as expected. When using the HT method with 500 observations, the MAE is even larger than that of the naive 2SLS estimator and the method never chooses the oracle model, leading none of the CIs to cover the true value. This is in line with the IV selection results - the frequency of including all invalid instruments as invalid, and that of selecting the oracle model are 0. When using CIM, the MAE is already low when $N = 500$, 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. Results are very similar for our AHC method. 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 $N = 1000$ and $N = 2000$, and the probabilities to select the oracle model are close to one, while for HT it is lower, showing that CIM and AHC have better finite 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

Table 13: 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.

uniform distributions as $\boldsymbol{\gamma}_1 = unif(1,2)$, $\boldsymbol{\gamma}_2 = unif(3,4)$ and $\boldsymbol{\gamma}_3 = unif(5,6)$, when there is a third endogenous regressor. The rest of the parameters are the same as before. With this setting we estimate $\boldsymbol{\beta} = \mathbf{0}$ for $m = 1000$ replications. The results can be found in Table 13. It can be seen that the performance of our method approaches that of the oracle estimator as the sample size grows large. But as the number of endogenous variables increases from 1 to 3, it needs a larger sample size to achieve oracle selection.

5.2 *Some weak instruments among the candidate instruments*

Now we check the performance of the previously mentioned methods when Assumption 6 and Assumption 11 are violated, i.e there are weak instruments among the candidates. For

Table 14: 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.

individually weak instruments, we consider the local to zero setup and set their first stage parameters as $\gamma_j = C/\sqrt{n}$ with $C = 0.1$.

Firstly, consider the same setting as in Section 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 5.1. We focus on the large sample performance in the presence of weak instruments and fix the sample size to $N = 2000$.

Simulation results are calculated based on 1000 Monte Carlo replications. We present the results in Table 14, where MAE, *# invalid* and *p allinv* are defined in the same way as in Section 5.1. Here we report three different IV selection statistics: the frequency of selecting all valid and strong instruments as valid (*strongvalid*), the frequency of selecting all weak invalid instruments as invalid (*weakin*), and the frequency of selecting all weak valid instruments as invalid (*weakva*). In these designs, 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 (either strong or weak) are selected as invalid, because including any invalid instruments in IV estimation can cause severe bias.

In Table 14 we can see that in the presence of weak instruments, the CI method can be very 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), meaning that it 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, HT and AHC).

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, 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 as *p allinv* is only at 0.053. This results in a significantly larger MAE than the oracle model.

The AHC performs well in general, because 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, thus in some cases they are not totally separated by the algorithm. This is not the 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 (MAEs of AHC are very close to those of the oracle models). It is noticeable 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 performance of AHC in the presence of weak IVs with two endogenous variables in large samples (fix sample size $N = 5000$). Simulations are

Table 15: Weak IV simulation designs with two endogenous regressors

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 16: 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.

conducted in four designs with 9 candidate instruments (see Table 15). In Design 1, each instrument is valid but only strong for one endogenous variable, respectively, violating Assumption 6. We are interested to see if the AHC method can include all the instruments as valid. In Design 2, still all the candidate instruments are strong for only one treatment variable, but some of them are invalid. In the last design, we make some of the instruments weak for both variables and a mixture of valid and invalid instruments. Results are presented in Table 16. In all designs, AHC achieves selection results close to the oracle model and hence very similar MAEs as well. 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 just-identified estimates use combinations of IVs that are strong, which can provide sufficient information for selecting valid instruments and hence delivering consistent estimates.

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.[5] Most papers in the literature only estimate the contemporaneous effects of immigration on labor market outcomes. Jaeger, Ruist, and Stuhler (2020) point out that there might be general equilibrium adjustments that affect wages in the long run, for example through the attraction of capital or the responses of native labor. This calls for including lagged immigration into the regression equation.
To illustrate our new method, we estimate the following linear model:

$$\Delta y_{lt} = \beta^{short} \Delta immi_{lt} + \beta^{long} \Delta immi_{lt-10} + a_t + \varepsilon_{lt}, \tag{37}$$

as in Basso and Peri (2015).

Here, 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 $\Delta immi_{lt}$ and $\Delta immi_{lt-10}$, which is the change of the share of immigrants in employment. The coefficients of interest are the short-term effect β^{short} and the long-term effect β^{long}. Decade fixed-effects are captured by a_t and ε_{lt} is the error term. Commuting-zone fixed effects are eliminated through first-differencing. This regression is canonical in migration economics. The authors use data from the Census Integrated Public Use Micro Samples 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. This identification strategy dates back to Altonji and Card (1991). The papers that invoke this exclusion restriction are numerous and are summarized in the preceding chapter.

In a recent paper, Goldsmith-Pinkham, Sorkin, and Swift (2020) discuss a class of IVs, so-called shift-share IVs, which are extensively used in labor economics. They show that a sufficient condition for the validity of the shift-share instrument is that all shares are valid and that an over-identified model with all shares as instruments can be used equivalently to the just-identified model which uses a single IV formed by all the shares. The authors show that a sufficient condition for validity of the shift-share IV is that all shares are valid. They also provide robustness-to-misspecification weights which illustrate how big the impact of the invalidity of each instrument is on the invalidity of the entire estimator.

Therefore, we use all shares of migrants from a certain origin country j, at a base period t_0 in region l. This share is denoted by z_{jlt_0}. We use origin-specific shares from 19 origin country groups and base years 1970 and 1980 as separate IVs and obtain $J = 38$ IVs. However, these previous settlement patterns might be invalid. Jaeger, Ruist, and Stuhler (2020) show that IV estimators that rely on this kind of exclusion restriction might

[5] An overview of the literature can be found in Dustmann, Schönberg, and Stuhler (2016).

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 well 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 hence should also be instrumented. To circumvent these problems, we apply the new estimator, which allows for direct effects of many migrant settlement variables on wages by pre-selecting the valid instruments.

RESULTS The results can be found in Table 17. 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 $0.1/log(n)$ (0.013).

When using AHC with this significance level in the downward testing procedure, two origin country shares are selected as invalid: the share of Mexicans in the US in 1970 and 1980. The coefficient estimate of the short-term effect 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 migrants were attracted to border regions 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, invalidating the initial shares. Moreover, GPSS find that Mexico has the highest sensitivity-to-misspecification weight, that is the overall bias will be sensible to any invalidity stemming from the Mexican share. This application has shown that our new method can make a big difference in practical terms, because it can help researchers identify IVs which violate the exclusion restriction.

7 CONCLUSION

We have proposed a novel method to select valid instruments. This method can be particularly helpful in cases when there are many candidate instruments and 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

Table 17: 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), J = 38. Standard errors in parentheses. Observations weighted by beginning-of-period population. Significance level in testing procedure: 0.013.

candidate instruments. Another suitable example is Mendelian Randomization, the use of IVs in epidemiology.

The advantages of our method are that it extends straightforwardly to the setting with multiple endogenous regressors and 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.

Appendices

E METHODOLOGICAL APPENDIX

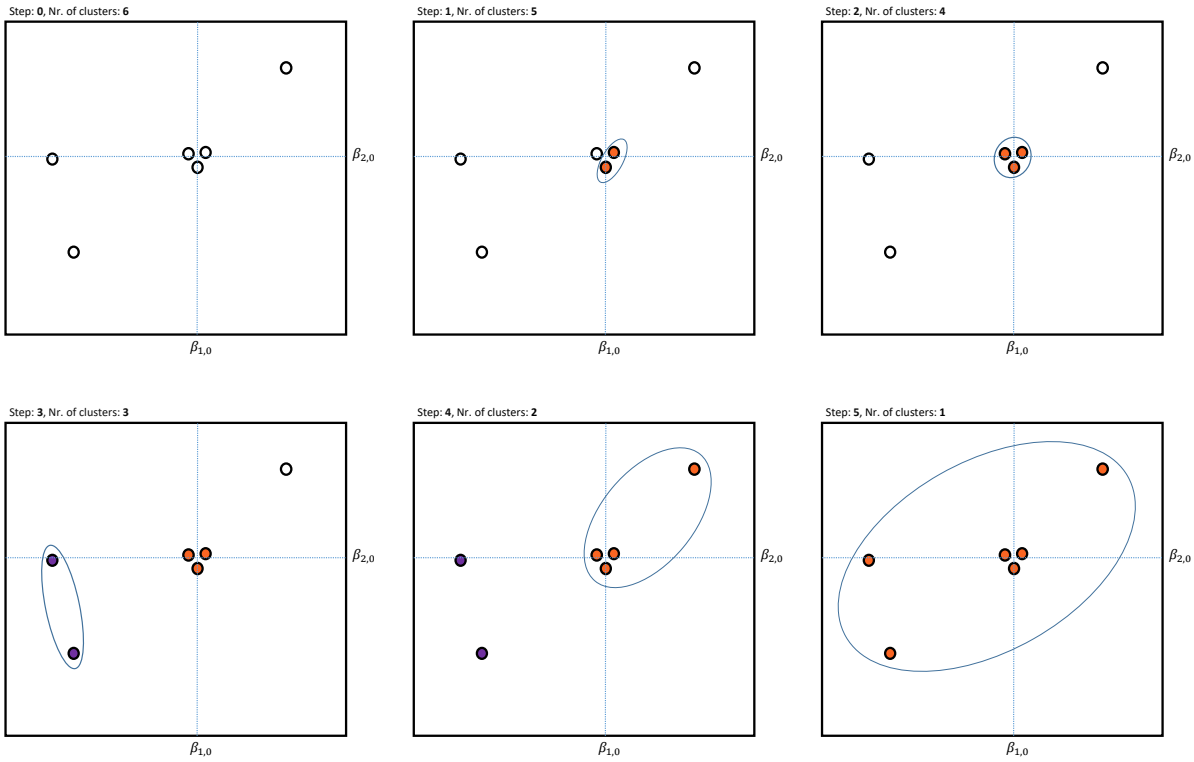
E.1 *Illustration of the IV selection procedure for $P = 2$*

In figure 9, 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 9: Illustration of the algorithm with two regressors



E.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 WLHB 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

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

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

$$(\hat{\boldsymbol{\beta}}_{[j]} \quad \hat{\boldsymbol{\alpha}}_{[j]}')' = \hat{\mathbf{H}}^{-1}\hat{\boldsymbol{\Gamma}} = \hat{\mathbf{H}}^{-1}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{D}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{u}) = \hat{\mathbf{H}}^{-1}(\hat{\boldsymbol{\gamma}}\boldsymbol{\beta} + \boldsymbol{\alpha} + (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{u})$$

Note that $\hat{\boldsymbol{\gamma}}\boldsymbol{\beta} + \boldsymbol{\alpha}$ is equal to

$$\begin{pmatrix} \hat{\boldsymbol{\gamma}}_1\boldsymbol{\beta} + \boldsymbol{\alpha}_1 \\ \hat{\boldsymbol{\gamma}}_2\boldsymbol{\beta} + \boldsymbol{\alpha}_2 \end{pmatrix}.$$

By Assumption 14, we have the following asymptotics

$$plim(\hat{\boldsymbol{\beta}}_{[j]}' \quad \hat{\boldsymbol{\alpha}}_{[j]}')' = plim(\hat{\mathbf{H}}^{-1} \begin{pmatrix} \hat{\boldsymbol{\gamma}}_1\boldsymbol{\beta} + \boldsymbol{\alpha}_1 \\ \hat{\boldsymbol{\gamma}}_2\boldsymbol{\beta} + \boldsymbol{\alpha}_2 \end{pmatrix}) = \begin{pmatrix} \boldsymbol{\beta} + \boldsymbol{\gamma}_1^{-1}\boldsymbol{\alpha}_1 \\ -\boldsymbol{\gamma}_2\boldsymbol{\gamma}_1^{-1}\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 \end{pmatrix}$$

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

E.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 E.2.

Remark 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 6. □

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

E.4 *Oracle Properties*

This section gives proofs for Lemma 1 and Theorems 1 and 3. All proofs apply for the general case that $P \geq 1$.

Proof of Lemma 1

Overall, we want to show that the probability that a cluster $\mathcal{S}_j$ with elements from the true underlying partition $\mathcal{S}_{0q}$ is merged with a cluster with elements from the same partition $\mathcal{S}_{0q}$ goes to 1.
The proof is structured as follows:

1. We note that the means of clusters which are associated with elements from the same family also converge to the same vector as each estimator 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 11 to 14:

$$\begin{aligned} plim(\hat{\boldsymbol{\beta}}_{[j]}) = plim(\hat{\boldsymbol{\beta}}_{[k]}) = \mathbf{q} \\ plim(\hat{\boldsymbol{\beta}}_{[l]}) = \mathbf{r} \end{aligned} \tag{38}$$

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 $\mathcal{S}_l \subset \mathcal{S}_{0r}$.

$$plim\ \bar{\mathcal{S}}_j = \frac{\sum\limits_{\hat{\beta}_{[j]} \in \mathcal{S}_j} \hat{\mathbf{fi}}_{[j]}}{|\mathcal{S}_j|} = \frac{|\mathcal{S}_j|\mathbf{q}}{|\mathcal{S}_j|} \text{ where } \mathcal{S}_j \subset \mathcal{S}_{0q} \tag{39}$$

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

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

$$\Leftrightarrow \quad lim\, P(\frac{|\mathcal{S}_j||\mathcal{S}_k|}{|\mathcal{S}_j| + |\mathcal{S}_k|}||\bar{\mathcal{S}}_j - \bar{\mathcal{S}}_k||^2 < \frac{|\mathcal{S}_j||\mathcal{S}_l|}{|\mathcal{S}_j| + |\mathcal{S}_l|}||\bar{\mathcal{S}}_j - \bar{\mathcal{S}}_l||^2) = 1 \tag{40}$$

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

Part 3: We want to prove equation (40) 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 (39)

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

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

$$\lim\, P(a < \varepsilon) = 1 \tag{41}$$

and

$$\lim\, P(|b - c| < \varepsilon) = 1. \tag{42}$$

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,

$$\lim\, P(b < \frac{1}{2}c) \neq 0 \tag{43}$$

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 (43).

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 1

Proof. The proof for Theorem 1 is structured as follows:

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

2. We show that Algorithm 2 can recover $\mathcal{F}_0$ from the selection path from Algorithm 1.

Part 1 follows from Corollary 3 directly.
Part 2: Firstly, we establish the properties of the Sargan statistic. The following two equations can be also found in WLHB (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 $\boldsymbol{\theta}_{\mathcal{I}} = [\beta \quad \boldsymbol{\alpha}_{\mathcal{I}}']'$, the Sargan test statistic is given by

$$S(\hat{\boldsymbol{\theta}}_{\mathcal{I}}) = \frac{\hat{\mathbf{u}}(\hat{\boldsymbol{\theta}}_{\mathcal{I}})'\mathbf{Z}_{\mathcal{I}}(\mathbf{Z}_{\mathcal{I}}'\mathbf{Z}_{\mathcal{I}})^{-1}\mathbf{Z}_{\mathcal{I}}'\hat{\mathbf{u}}(\hat{\boldsymbol{\theta}}_{\mathcal{I}})}{\hat{\mathbf{u}}(\hat{\theta}_{\mathcal{I}})'\hat{\mathbf{u}}(\hat{\boldsymbol{\theta}}_{\mathcal{I}})/n} \tag{44}$$

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, 2019, the Sargan statistic has the following properties:

Property 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 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 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 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 16.a, $\mathcal{F}_0$ is the largest family among all Q families. Then following Property 3 and $\xi_{n,J-|\hat{\mathcal{I}}|-P} \to \infty$ for the Sargan test we have

   $$\lim_{n\to\infty} P(S(\hat{\boldsymbol{\theta}}_{\hat{\mathcal{I}}}) < \xi_{n,J-|\hat{\mathcal{I}}|-P}) = 1. \tag{45}$$

indicating that $\mathcal{V}$ would be selected as the set of valid instruments asymptotically.

2. $K = Q$. By Corollary 3 we know that the K clusters are associated with the Q families respectively, and by Assumption 16.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 (45) and Algorithm 2 selects $\mathcal{V}$ as the set of valid instruments.

To summarize, asymptotically, at steps $1 \leq K < Q$, Algorithm 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 1. □

Proof of Theorem 3

The proof of Theorem 3 works in the same way as the proof of Theorem 1.

Proof. The proof for Theorem 3 is structured as follows:

1. We note that asymptotically the selection path generated from Algorithm 1 contains all groups $\mathcal{G}_q$.

2. We show that Algorithm 2 can recover all $\mathcal{G}_q$ from the selection path from Algorithm 1.

Part 1 again follows directly from Corollary 3.
Part 2: Firstly, we establish the properties of the Sargan statistic.

Property 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 4, we have

$$\lim_{n\to\infty} P(S(\hat{\boldsymbol{\theta}}_{\mathcal{G}_q}) < \xi_{n,J-|\mathcal{G}_q|-P}) = 0. \tag{46}$$

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 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

$$S(\hat{\boldsymbol{\theta}}_{\hat{\mathcal{G}}_k}) = S(\hat{\boldsymbol{\theta}}_{\mathcal{G}_q}). \tag{47}$$

   By Property 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 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 (46).

To summarize, asymptotically, at steps $1 \leq K < Q$, Algorithm 2 does not stop; then it moves to step $K = Q$ and selects the partition that leads to each family being associated with its own cluster.

Combine *Part 1* and *Part 2*, we prove Theorem 3. □

IV

INSTRUMENTAL VARIABLE SELECTION METHODS UNDER NEAR EXCLUSION

1 INTRODUCTION

In two-stage least squares (2SLS) estimation it is central that IVs fulfill the exclusion restriction. This means that instruments should not be related with the outcome other than via the treatment. In recent years, an emerging literature has proposed econometric methods to select which of the IVs is valid in the sense that it fulfills the exclusion restriction. These methods build on the assumption that there is a large enough group of instruments that *exactly* fulfills the exclusion restriction. This key assumption might often hold approximately, but seldom exactly, because there might be a small correlation between instruments and the error term.[1]

In this chapter, the research question hence is: How do the selection methods behave when the exclusion restriction is close to being fulfilled? The models used in the literature assume a *global* component of invalidity, denoted by α_j, which co-determines the value of the estimand. I model the near-exclusion setting with an additional *local* part denoted by τ_j that goes to zero as n grows large. I find that when the local violations are mild relative to the order of convergence of the tuning parameters, violations of the exclusion restriction do not touch the theoretical guarantees of the selection methods. If local deviations of the invalid IVs are major while those of the valid instruments are minor, the key assumption of the selection methods can even be relaxed. These findings are reassuring for empirical researchers in many applied disciplines, such as economics and epidemiology who use the IV selection methods, because they show that deviations from perfect exclusion restrictions of the instruments are allowed to a certain extent. As these small deviations are likely to be present in real-world data, my results enhance the practical relevance of the selection methods.

This exercise is relevant for researchers, because in order for the selection methods to work, one must assume that IV-specific estimands form exact groups. In reality, arguably these exact groups do not exist, and the estimands are scattered around certain values instead of being concentrated exactly at them. When modeling the groups with the help of the local invalidity, in the finite sample there are no exact groups but only groups that live in an environment of their global value, which is co-determined by α_j. I show how large these local deviations from the exact groups are allowed to be so that the selection results still hold. This setting should not be taken literally but as an approximation that mimics more realistic settings.

The selection methods that are discussed in this chapter originated in D. W. Andrews (1999), who proposed moment selection criteria to consistently select the valid and invalid instruments. These criteria are based on a test of overidentifying restrictions, the Hansen-Sargan test (Sargan, 1958). This procedure becomes infeasible with a moderate number of IVs. Therefore, Kang, Zhang, Cai, and Small (2016) use the Lasso to select valid and invalid instruments. WFDS show that the Lasso selection results depend on the correlation and relative strength of instruments. They propose the adaptive Lasso which consistently selects invalid instruments, if more than half of instruments is valid. To determine the

[1] Instruments that fulfill the exclusion restriction are called *valid*, while those that do not fulfill the exclusion restriction are called *invalid*.

penalty parameter, they use the downward testing procedure proposed by D. W. Andrews (1999), which involves the Hansen-Sargan test. Guo, Kang, Cai, and Small (2018) propose a method which is based on a series of pairwise tests and relies on the largest group of IVs being valid, the plurality assumption. WLHB reduce the number of pairwise tests with their CIM and again combine it with a downward testing procedure which makes use of the Hansen-Sargan test. Apfel and Liang (2021) use cluster analysis to find the largest group of valid IVs and provide extensions for the multiple regressor and weak instrument case, also using the downward testing procedure.

The entire literature is based on the assumption that IV-specific estimands cluster exactly at certain values. In particular, the group of globally valid instruments (with $\alpha_j = 0$) should be exactly valid. There has been a literature on near exogeneity, including Newey (1985), Conley, Hansen, and Rossi (2012), Caner (2014) and I. Andrews, Gentzkow, and Shapiro (2017). This literature approximates the part of the invalidity that disappears as $\tau = \frac{c}{\sqrt{n}}$. This chapter combines two strands of the literature: the invalid IV selection literature and the literature on imperfect instruments.

The implications of this chapter are of interest for a variety of studies. First, a strand of literature on the effects of immigration on wages uses multiple origin-country-specific shares to instrument immigration. Here, it might be the case that shares of immigrants in the past have a direct effect on present wages. It might be that for some origin-country shares the correlation is big and persistent, while other origin-country shares are sufficiently lagged or belong to an immigrant community that did not have direct effects on wages. For the latter, there might still be small violations of the exclusion restriction which do not lead to big deviations from the exclusion restriction.

Second, in international economics a similar strategy is used. Here, the imports to high-income countries can be used as instruments when estimating the effect of import exposure on unemployment. When imports are correlated across countries this approach might fail. Even when there is a large-enough group of import variables that is not correlated with the unobserved shocks to the country of interest, the group of valid IVs might be split into smaller or even singleton groups by local invalidity.

Third, in epidemiology Mendelian Randomization is a quickly growing field. Here, genetic variants are used as instruments when estimating the effect of an exposure on a health outcome. These applications have motivated Kang, Zhang, Cai, and Small (2016) and WLHB. A case of concern is that the genetic variants have a direct effect, unknown to the epidemiologist. These direct effects are called *pleiotropy*. It is difficult to exclude these minor direct effects.

Fourth, studies that use lottery-based instruments, as in Angrist (1990) might be additional candidates for using valid IV selection methods. Here, the author uses 884 instruments. However, in this case strength of the first stage might be a concern. Overall, any application in economics and beyond where many instruments are used, many of which can be assumed to be valid, is suited for the methods discussed here. The implications discussed in this chapter are of interest for practitioners who want to apply these methods.

This chapter is structured as follows. In Section 2, I begin with the standard model, introduce the assumptions and then present the model with local violations of the exclusion restriction. In Section 3, I summarize three methods: the Confidence Interval Method by WLHB, the Agglomerative Hierarchical Clustering Method for IV selection by Apfel and Liang (2021) and the downward testing procedure of D. W. Andrews (1999) which the preceding two methods build upon. These methods make use of tests of overidentifying restrictions and hence we focus on them. The main desirable properties of these methods, the oracle properties, are also summarized. The common idea of these approaches is to consistently select the invalid IVs under appropriate conditions and to then control for the invalid IVs.

In Section 4, I allow for near exclusion of the instruments and study the properties of the introduced methods as well as of the Sargan test statistic. I show for which settings the results of the original studies still hold and in which not. In Section 5, I discuss whether these new results also imply that larger sequences of critical values are allowed in the downward testing procedure and whether there are situations in which the methods' key assumption can be relaxed. In Section 6, I provide a simulation study that confirms the points made in the preceding sections. Section 7 concludes.

2 MODEL AND ASSUMPTIONS

In the following, matrices will appear in uppercase bold, e.g. matrix $\mathbf{Z}$ and vectors will appear in lowercase bold, e.g. vector $\mathbf{d}$. For a full-column matrix $\mathbf{X}$, let $\mathbf{P}_X = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ the projection matrix on the column space and $\mathbf{M}_X = \mathbf{I} - \mathbf{P}_X$ the annihilator matrix. Let $\mathcal{V}$ the set of valid instruments, $\mathcal{I}$ the set of invalid instruments and $\mathcal{A}$ another set of instruments with $\mathcal{A} \neq \mathcal{I}$.

In this section, I present the model and assumptions as introduced in Windmeijer, Liang, Hartwig, and Bowden (2021, WLHB). Consider the linear model

$$\mathbf{y} = \mathbf{d}\beta + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{u} = \mathbf{X}\boldsymbol{\theta} + \mathbf{u} \tag{48}$$

with $\boldsymbol{\theta} = (\beta \;\; \boldsymbol{\alpha}')'$ and $\mathbf{X} = [\mathbf{d} \quad \mathbf{Z}]$. The first-stage equation is

$$\mathbf{d} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{49}$$

with n observations, J instruments and K endogenous regressors. The outcome is $\mathbf{y}$, the potentially endogenous treatment variable is $\mathbf{d}$, the matrix of instruments is $\mathbf{Z}$ and the error term is $\mathbf{u}$, with $E(u_i|\mathbf{z}_i)$. The coefficient of interest is β, $\boldsymbol{\alpha}$ models the direct correlation between IVs and outcome and hence IVs which are invalid are associated with a non-zero entry of α_j in $\boldsymbol{\alpha}$, as defined in the following. In the first-stage equation, $\boldsymbol{\gamma}$ collects the coefficients of the correlation between instruments and treatment. The mean-zero error

term is $\boldsymbol{\varepsilon}$, with $E(\mathbf{z}_i\varepsilon_i) = \mathbf{0}$. Here, I will focus on the case with one regressor ($K = 1$), for simplicity. Note that as before, Γ_j and γ_j are the j-th entries of the following vectors:

$$\boldsymbol{\Gamma} = E(\mathbf{z}_i\mathbf{z}_i')^{-1}E(\mathbf{z}_i y_i) \quad and \quad \boldsymbol{\gamma} = E(\mathbf{z}_i\mathbf{z}_i')^{-1}E(\mathbf{z}_i d_i).$$

Consider the model with the invalid instruments correctly included as controls (the oracle model):

$$\mathbf{y} = \mathbf{d}\beta + \mathbf{Z}_{\mathcal{I}}\boldsymbol{\alpha}_{\mathcal{I}} + \mathbf{u} = \mathbf{X}_{\mathcal{I}}\boldsymbol{\theta}_{\mathcal{I}} + \mathbf{u}, \tag{50}$$

with $\boldsymbol{\theta}_{\mathcal{I}} = (\beta \;\; \boldsymbol{\alpha}_{\mathcal{I}}')'$ and $\mathbf{X}_{\mathcal{I}} = [\mathbf{d} \quad \mathbf{Z}_{\mathcal{I}}]$. The model where some IVs have been wrongly included as valid is denoted as

$$\mathbf{y} = \mathbf{d}\beta + \mathbf{Z}_{\mathcal{A}}\boldsymbol{\alpha}_{\mathcal{A}} + \boldsymbol{\xi} = \mathbf{X}_{\mathcal{A}}\boldsymbol{\theta}_{\mathcal{A}} + \boldsymbol{\xi} \tag{51}$$

where $\boldsymbol{\xi} = \mathbf{Z}_1\boldsymbol{\alpha}_1 + \mathbf{u}$ and $\mathbf{Z}_1$ are invalid IVs which are wrongly assumed valid, and $\boldsymbol{\theta}_{\mathcal{A}} = (\beta \;\; \boldsymbol{\alpha}_{\mathcal{A}}')'$.

The assumptions are the same as in the preceding chapter, for the single endogenous regressor case, i.e. Assumptions 11 to 16. As before, the key assumption considered in Guo, Kang, Cai, and Small (2018), WLHB and Apfel and Liang (2021) is that the largest group of IVs is valid (*plurality assumption*).

Next, we consider an asymptotic setting in which the violations consist of a group-specific (*global*) part α_j and a (*local*) part τ_j that disappears asymptotically, with $\tau_j \to 0$ as $n \to \infty$. The model now becomes

$$\mathbf{y} = \mathbf{d}\beta + \mathbf{Z}_{\mathcal{A}}\boldsymbol{\alpha}_{\mathcal{A}} + \mathbf{Z}\boldsymbol{\tau} + \mathbf{u} \tag{52}$$

where both α_j and τ_j can differ by IV. All entries τ_j are collected in the vector $\boldsymbol{\tau}$. As in Caner (2014), I model $\tau_j = \frac{c}{n^\kappa}$, with $\kappa > 0$. Here, $\kappa = 1/2$ is a mild violation of an exclusion restriction, $1/2 < \kappa < \infty$ is the range of minor violations of the exclusion restriction and $0 < \kappa < 1/2$ is the range of major violations of an exclusion restriction. Conley, Hansen, and Rossi (2012) stress that this approximation should not be taken literally and serves as a thought experiment to illustrate a case where sampling error and exogeneity error may both affect the asymptotic distribution.

For $\kappa = 1/2$ without global invalidity, ($\alpha = 0$) Conley, Hansen, and Rossi (2012, Appendix A.2) and I. Andrews, Gentzkow, and Shapiro (2017) show that the standard 2SLS-estimator, $\hat{\beta}_{2SLS}$, is asymptotically biased:

$$\sqrt{n}(\hat{\beta}_{2SLS} - \beta) \xrightarrow{d} N(0, \mathbf{V}) + (\mathbf{Q}_{Zd}{}'\mathbf{Q}_{ZZ}\mathbf{Q}_{Zd})^{-1}\mathbf{Q}_{Zd}{}'\boldsymbol{\tau}$$

where $\mathbf{Z}'\mathbf{d} \xrightarrow{P} \mathbf{Q}_{Zd}$ and $\mathbf{V}$ is a variance-covariance matrix.

In this setting the standard t-test over- and under-rejects the Null hypothesis, making inference on the parameters of interest unreliable (Berkowitz, Caner, and Fang, 2008).

3 VALID INSTRUMENT SELECTION METHODS

A small statistical literature tries to select valid instruments from a large set of potentially invalid instruments. The first paper to do so is D. W. Andrews (1999). Kang, Zhang, Cai, and Small (2016) use the Lasso to shrink some entries in $\boldsymbol{\alpha}$ to zero. The key identifying assumption here is that the fraction of valid IVs needs to exceed one half. WFDS also rely on this majority assumption, but they also allow for correlated instruments and instruments with varying strength. Guo, Kang, Cai, and Small (2018) relax the majority assumption, basing their method on Assumption 16, the plurality assumption. WLHB and Apfel and Liang (2021) also rely on this plurality assumption. We focus on the last two papers, because these methods produce a selection path and select a model with a test of overidentifying restrictions. These two methods make use of the downward testing procedure proposed in D. W. Andrews (1999). Therefore, I also discuss the downward testing procedure.

The methods that I will present have desirable properties, which are called oracle properties as defined in Definition 2. This means that the methods select the invalid instruments as invalid and the post-selection estimator converges in distribution to the estimator that uses all invalid IVs as invalid and the valid ones as valid. In other words, oracle properties mean that asymptotically the method works as well as if one knew the true set of valid instruments. The CIM and AHC have oracle properties under the Assumptions detailed in the previous three chapters, which imply 11-16. This follows from Theorems 1 and 2 in WLHB and Apfel and Liang (2021).

The idea of these methods is to select those IVs as valid which produce similar estimates and to select the rest as invalid. In the following, I quickly present how two of the selection methods work.

3.1 *Confidence interval method*

The CIM by WLHB works as follows

1. Compute just-identified estimates using one IV at a time and control for the rest: $\frac{\hat{\Gamma}_j}{\hat{\gamma}_j}$. Calculate standard errors and CIs for each estimate, using the critical value ψ_n.
2. Order the CIs in ascending order.
3. For each CI, compare if it overlaps with the preceding intervals. Count with how many intervals it overlaps.
4. Select the largest group of IVs whose CIs overlap.

This corresponds to Algorithm 1 in WLHB. Sequentially reducing the critical value, the largest group and hence the selected model also varies and the researcher obtains a selection path. To select a specific step on the path, a significance level is pre-specified for the downward testing procedure and a test of the overidentifying restrictions is run at each

step. The procedure starts with the largest model, when all CIs overlap and all instruments are selected as valid and stops as soon as the test is not rejected anymore. The consistency of this procedure follows from the asymptotic properties of the test statistic.

One weakness of the CIM is that it is not clear how to deal with weak instruments. Weak instruments lead to wider CIs and therefore it is more likely that weak invalid IVs are selected as valid, because they overlap with the CIs of strong and valid IVs. Excluding weak instruments which do not pass the first-stage F-test but might contribute information jointly might not be optimal. Moreover, there is no straightforward extension to the multiple endogenous regressor case.

3.2 *Agglomerative hierarchical clustering*

The AHC by Apfel and Liang (2021) works similarly to the CIM. It is an application of Ward's algorithm (Ward, 1963) to this causal inference problem.

1. Compute $\frac{\hat{\Gamma}_j}{\hat{\gamma}_j}$. Each estimate constitutes its own cluster.
2. Join the two clusters which are closest in terms of their weighted Euclidian distance.
3. Recalculate the cluster means
4. Iterate steps 2. and 3.
5. Select the largest cluster as valid.

This is analogous to Algorithm 1 in Apfel and Liang (2021). Repeating step 2 until all estimates are in a single cluster, again produces a selection path. In this case, the selection path is reversed and the number of clusters is again selected with a downward testing procedure. There are two key advantages of this method as compared to the CIM: First, there is a straightforward extension to accommodate multiple regressors. In this case, the plurality assumption has to be generalized to a family plurality assumption: the largest group of just-identified estimates should use valid instruments. Second, weak instruments do not deteriorate the performance of the method when a plurality of IVs is strong and valid. Intuitively, this holds because estimators which use weak IVs are inconsistent and hence are treated as invalid by the algorithm. Moreover, the computational complexity is lower. The disadvantages are that the information contained in the standard errors is discarded and the entire selection path needs to be computed, while the CIM can stop as soon as the test is not rejected at the pre-specified significance level.

3.3 *Downward testing procedure*

Both methods just presented make use of the downward testing procedure by D. W. Andrews (1999). To discuss the methods' performance with local violations it is therefore necessary to introduce how the downward testing procedure works. If future methods will also be built on top of this downward testing procedure, my results will also be of interest.

In the downward testing procedure, the Sargan test is used to select the valid and invalid instruments. It starts by evaluating the Sargan test for the model with all J IVs assumed to be valid. If the test rejects, it continues testing all possible models with one IV used as control and the rest valid. If the minimum Sargan statistic rejects, the procedure continues with all models that assume two of the IVs are invalid. This can go on until the downward testing procedure gets to evaluate models with $K+1$ IVs. The downward testing procedure stops when the Sargan test is not rejected at a pre-specified level of significance.

This procedure is said to be *consistent* when the probability of selecting the valid IVs as valid and the invalid IVs as invalid goes to 1. The downward testing procedure is consistent if the critical values of the χ^2_{J-K-1} distribution satisfy $\gamma_n \to \infty$ for $n \to \infty$ and $\gamma_n = o(n)$ (Theorem 2 in D. W. Andrews, 1999). Instead of choosing the critical value γ_n, typically WLHB set the p-value of the Sargan-test to $0.1/log(n)$, following Belloni, Chen, Chernozhukov, and Hansen (2012).

4 ALLOWING FOR NEAR EXCLUSION

In Section 2 I considered two asymptotic settings: One in which the violations consist of a global part α_j and a local part τ_j which disappears asymptotically. In this section, I allow for local violations. First, I look at the performance of the selection methods, then at how the Sargan test's behavior changes. This also determines the behavior of the downward testing procedure.

4.1 *Selection methods under near exclusion*

First, I establish that when the local violation $\boldsymbol{\tau}$ is $o_p(1)$, the just-identified estimators will cluster as before and hence the AHC also works as before. The first question to ask is: When will $\hat{\beta}_j$ cluster at the groups with $\beta + q$, where $q = \frac{\alpha_j}{\gamma_j}$ is only determined by the global part?

$$\hat{\beta}_j - \beta = \frac{\hat{\Gamma}_j}{\hat{\gamma}_j} - \beta \xrightarrow{P} \frac{\alpha_j}{\gamma_j} + \frac{\tau_j}{\gamma_j} \tag{53}$$

When $\tau_j = o_p(1)$, $plim(\hat{\beta}_j - \beta) = q$, as without violations. In this setting, a group is defined by q similarly to Definition 4 with the difference that there is a τ_j-term that disappears asymptotically. Therefore, in large samples the point estimates cluster at the respective $\beta + q$, the distances between the $\hat{\beta}_j$ become zero and the AHC can still find the correct cluster.

For the CIM to lead to the same results in this new setting, the instrument-specific estimates have to retain the same asymptotic distribution as without local violations. For

this to hold, the asymptotic distribution of $(\hat{\mathbf{\Gamma}} \quad \hat{\boldsymbol{\gamma}})'$ should be the same as without local violations. Since

$$\begin{aligned}\hat{\mathbf{\Gamma}} &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{Z}(\boldsymbol{\gamma}\beta + \boldsymbol{\alpha} + \boldsymbol{\tau}) + \mathbf{u} + \boldsymbol{\varepsilon}\beta) \\ &= \boldsymbol{\gamma}\beta + \boldsymbol{\alpha} + \boldsymbol{\tau} + (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{u} + \boldsymbol{\varepsilon}\boldsymbol{\beta})\end{aligned}$$

and

$$\mathbf{\Gamma} = E(\mathbf{z}_i\mathbf{z}_i')^{-1}E(\mathbf{z}_i y_i) = E(\mathbf{z}_i\mathbf{z}_i')^{-1}E(\mathbf{z}_i(\mathbf{z}_i(\boldsymbol{\gamma}\beta + \boldsymbol{\alpha} + \boldsymbol{\tau}) + \mathbf{u} + \boldsymbol{\varepsilon}\beta)) = \boldsymbol{\gamma}\beta + \boldsymbol{\alpha} + \boldsymbol{\tau}$$

the following still holds, because $\boldsymbol{\tau}$ cancels out

$$\sqrt{n}\left(\begin{pmatrix}\hat{\mathbf{\Gamma}} \\ \hat{\boldsymbol{\gamma}}\end{pmatrix} - \begin{pmatrix}\mathbf{\Gamma} \\ \boldsymbol{\gamma}\end{pmatrix}\right) \xrightarrow{d} N(0, \mathbf{\Lambda}) \tag{54}$$

where $\mathbf{\Lambda}$ is a variance-covariance matrix. Based on this, WLHB derive the asymptotic distribution of $\hat{\beta}_j$. Hence, with $\tau_j = o_p(1)$, asymptotic normality is untouched.

In their Lemma 1, WLHB show that the CIs of $\hat{\beta}_j$ with two instruments from the same group overlap asymptotically and with two instruments from different groups the IVs don't overlap. This now needs to hold in the setting of the model with local invalidity. The vectors $\boldsymbol{\alpha}_V$ and $\boldsymbol{\alpha}_I$ denote the global violations for valid and invalid IVs, while $\boldsymbol{\tau}_V$ and $\boldsymbol{\tau}_I$ denote the respective local violations. The local violations are modeled as $\boldsymbol{\tau} = \frac{\mathbf{c}}{n^\kappa}$ with $\kappa > 0$. The local violation can be IV-specific with $\tau_q = \frac{c_q}{n^\kappa}$ and $\tau_k = \frac{c_k}{n^\kappa}$, and c_q and c_k being IV-specific constants. Larger κ mean smaller local violations and vice versa. Next, I show under which conditions for ψ_n the CIs of the same group overlap and those of different groups don't overlap.

Proposition 2. *CIs of same group instruments still overlap.*
For $\psi_n = o(n^\delta)$, $\delta > 0$, the CIs of instruments from the same group overlap with probability 1 when $\kappa > 1/2 - \delta > 0$.

Proposition 3. *CIs of different-group instruments don't overlap.*
For $\psi_n \to \infty$ and $\psi_n = o(n^{1/2})$, the CIs of IVs from different groups don't overlap when $\boldsymbol{\tau} = \frac{\mathbf{c}}{n^\kappa}$ with $\kappa > 0$.

All of the proofs for the following propositions can be found in the Appendix. The next corollary follows directly from the two propositions. Now, the rates at which local violations vanish can differ for globally valid and globally invalid instruments. I model this by different κ_V (the κ for the group of IVs with $\alpha = 0$) and κ_I (the κ for other groups of IVs):

Corollary 4. *Special case*
When $\psi_n = o(n^\delta)$, and the local violations are such that $\kappa_V > 1/2 - \delta \geq \kappa_I$ the CIs of valid groups overlap while those of the invalid IVs do not, not even for the same group.

For example, when $\psi_n = o(n^{1/4})$, i.e. $\delta = 1/4$, and for the valid IVs, $\boldsymbol{\tau}_V = \frac{\mathbf{c}}{n^{2/3}}$, i.e. $\kappa_V = 2/3$, Propositions 1 and 2 still hold. For the invalid IVs, however, take $\kappa_I = 1/6$, i.e.

$\tau_I = \frac{c}{n^{1/6}}$, it needs to hold that $lim\frac{\psi_n}{n^{1/3}} \neq 0$. Hence, if $\psi_n = o(n^{1/4})$, ψ_n is also $o(n^{1/3})$, the CIs of the valid group asymptotically overlap, while those of the invalid IVs do not, not even for instruments of the same group.

What does this special case mean intuitively? The local violations of the globally invalid instruments are larger than those of the globally valid IVs (i.e. $\kappa_V > \kappa_I$). The critical values are $\psi_n \to \infty$, $\psi_n = o(n^{\delta})$ and the local invalidity of globally valid IVs is small (κ_V is large) enough relative to a given δ and the local invalidity of globally invalid IVs is large (κ_I is small) enough. Then, the CIs overlap for globally valid instruments but they do not overlap for globally invalid ones, even though they are from the same group.

Figure 10 summarises the above results. The light gray set shows combinations of growth rates of the critical values, δ, and orders of convergence of the local violations, κ, for which Propositions 1 and 2 and hence Lemma 1 in WLHB still hold. That is for points in this area the CIs that use IVs from the same group still overlap, as they should and those from different groups don't overlap. The dark gray set shows combinations for which CIs from the same group will not overlap and those from different groups still do not overlap. In this area, the results in WLHB do not hold anymore. For a given $\delta = \bar{\delta}$, when κ is on line b (excluding the border point), Proposition 1 still holds. When κ is on line a, Proposition 1 does not hold and CIs from different groups can overlap. The special case discussed is given when κ_I lies on line a and κ_V lies on line b. Then, the CIs of IVs from the valid group overlap while the CIs of the same invalid group will not overlap. This special case can help relaxing the plurality assumption, later on.

4.2 *Properties of Sargan-test*

Next, to study the properties of the downward testing procedure under near exclusion I look at the Sargan test statistic, when there are local violations. For the following, take Assumptions 11 - 15 and the model referred to in each subsection.

4.2.1 *Invalid instruments from same group*

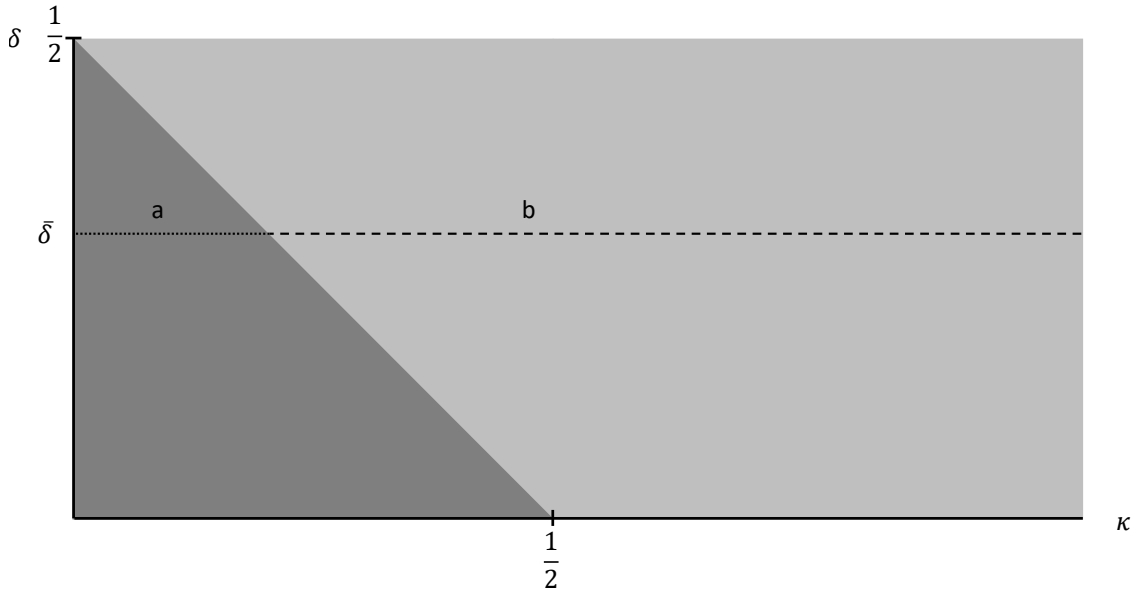
First, I show that if we test overidentifying restrictions with instruments from the same group, the Sargan test will behave the same as if there were no global violations. Hence, the results about the Sargan statistic for instruments from the valid group in sections 4.2.2 and 4.2.3 also hold for IVs from the same group.

The test of overidentification restrictions tests the validity of moment conditions $E(\mathbf{z}_i'\epsilon_i) = 0$ in the model

$$\mathbf{y} = \mathbf{d}\beta + \boldsymbol{\epsilon}. \tag{55}$$

Define $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \mathbf{d}\hat{\beta}$. The empirical moments $E(\mathbf{z}_i\hat{\epsilon}_i) = 0$ can also be fulfilled when $E(\mathbf{z}_i\epsilon_i) \neq 0$, as cautioned in Parente and Silva (2012). Note that $plim\hat{\beta} = \beta_s$. In my example $\boldsymbol{\epsilon} = \mathbf{Z}\boldsymbol{\alpha} + \mathbf{u}$. Also note that by Assumption 5 in WLHB the *plim*s of the terms in

Figure 10: Visualization of results for CIM



Note: This graph visualizes for which combinations of κ (horizontal axis) in $\boldsymbol{\tau} = \frac{\mathbf{c}}{n^{\kappa}}$, for a given δ (on vertical axis) in $\psi_n = o(n^{\delta})$, same-group CIs overlap and different-group instruments don't overlap. The light gray set is closed to the top, open to the right and does not include the border line $\delta = 1/2 - \kappa$. For a given δ if κ is in the light (dark) gray set, Proposition 1 (does not hold) holds. The dashed line (b) visualizes all possible values for κ for which it holds. If $\kappa \leq 1/2 - \bar{\delta}$, i.e. κ is on the dotted line (a), including the border of the dark and light gray set, Proposition 1 does not hold anymore.

expectations, below, are equal to their expectations, hence in the following I directly look at the expectations. From Assumption 14, it holds that $E(\mathbf{z}_i u_i) = 0$. Then,

$$\begin{aligned} E(\mathbf{z}_i \hat{\epsilon}_i) &= E(\mathbf{z}_i (d_i \beta + \mathbf{z}_i' \boldsymbol{\alpha} + u_i - d_i \hat{\beta})) = 0 \\ E(\mathbf{z}_i d_i)(\beta_s - \beta) &= E(\mathbf{z}_i \mathbf{z}_i') \boldsymbol{\alpha} \\ E(\mathbf{z}_i \mathbf{z}_i')^{-1} E(\mathbf{z}_i d_i)(\beta_s - \beta) &= \boldsymbol{\alpha} \\ \boldsymbol{\gamma}(\beta_s - \beta) &= \boldsymbol{\alpha} \end{aligned}$$

This holds for each element of the vectors when $\hat{\beta} \xrightarrow{P} \beta + \frac{\alpha_j}{\gamma_j}$. That's also exactly how a group is defined. But this $\hat{\beta}$ refers to the 2SLS estimate, and not to the just-identified estimates. The 2SLS estimate can be written as a weighted average of the just-identified estimates (controlling for the rest of instruments), with weights that sum to one. Given that the β_j are all the same by definition when they are in the same group, the 2SLS estimand will also be equal to the β_j.

4.2.2 *Globally valid, locally valid*

In the remainder of this section, I will derive the asymptotic distribution of the Sargan test statistic for four cases: globally valid IVs with and without local invalidity as well as globally invalid IVs with and without local invalidity.

Assume a model where we know which instruments are globally valid and invalid and $\boldsymbol{\tau} = 0$:

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

with $\mathbf{X}_{\mathcal{I}} = [\mathbf{d} \quad \mathbf{Z}_{\mathcal{I}}]$ and $\boldsymbol{\theta}_{\mathcal{I}} = (\beta \quad \boldsymbol{\alpha}_{\mathcal{I}}')'$.
The Sargan test statistic is

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

The standard result is that the Sargan test statistic converges in distribution to a chi-squared distribution.

Proposition 4. *If* $\boldsymbol{\tau} = 0$*, then*

$$Sar\left(\widehat{\boldsymbol{\theta}}_{\mathcal{I}}\right) \xrightarrow{d} \boldsymbol{\zeta}'\mathbf{A}\boldsymbol{\zeta} \sim \chi^2_{J-K-1} \tag{56}$$

where $\boldsymbol{\zeta} \sim N(0, \mathbf{I})$.

4.2.3 *Globally valid, locally invalid*

Next, I assume that the globally valid IVs have been selected but there is local invalidity as in the model in equation 52.

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

Proposition 5. *If* $\boldsymbol{\tau} = o_p(n^{-1/2})$*, then*

$$Sar\left(\widehat{\boldsymbol{\theta}}_{\mathcal{I}}\right) \xrightarrow{d} \chi^2_{J-K_{\mathcal{I}}-1} \ .$$

MILD VIOLATIONS: $\boldsymbol{\kappa} = \mathbf{1/2}$ Next, I'll look at the border case of mild violations, when $\boldsymbol{\tau} = \frac{\mathbf{c}}{\sqrt{n}}$. This case was the object of a lot of attention in the nearly exogenous instruments literature, because exogeneity error and sampling error both play a role contemporaneously for the asymptotic distribution.

Proposition 6. *If* $\boldsymbol{\tau} = \frac{\mathbf{c}}{\sqrt{n}}$*, then*

$$Sar\left(\widehat{\boldsymbol{\theta}}_{\mathcal{I}}\right) \xrightarrow{d} \chi^2_{J-K_{\mathcal{I}}-1}\left(\frac{1}{\sigma_u^2}\mathbf{c}'\mathbf{Q}_{ZZ}^{1/2\prime}\mathbf{A}\mathbf{Q}_{ZZ}^{1/2}\mathbf{c}\right) \ . \tag{57}$$

This result corresponds to the ones in Newey (1985, Theorem 2.1) and Hayakawa (2014, Theorem 2) who look at general violations of the moment conditions. The Sargan statistic converges in distribution and therefore is $O_p(1)$.

MAJOR VIOLATIONS: $\boldsymbol{\kappa < 1/2}$ Now, consider the case where $\boldsymbol{\tau} = \frac{\mathbf{c}}{n^{\kappa}}$ with $0 < \kappa < 1/2$, i.e. $\boldsymbol{\tau} > \frac{\mathbf{c}}{\sqrt{n}}$.

Proposition 7. *If* $\boldsymbol{\tau} > \frac{\mathbf{c}}{\sqrt{n}}$, *then*

$$\frac{Sar\left(\widehat{\boldsymbol{\theta}}_{\mathcal{I}}\right)}{n^{1-2\kappa}} \xrightarrow{P} \frac{1}{\sigma_u^2}\mathbf{c}'\mathbf{Q}_{ZZ}^{1/2'}\mathbf{A}\mathbf{Q}_{ZZ}^{1/2}\mathbf{c} \tag{58}$$

Hence as $Sar\left(\widehat{\boldsymbol{\theta}}_{\mathcal{I}}\right) = O_p(n^{1-2\kappa})$, for critical values that fulfill $\gamma_n = o(n^{\delta})$ with $\delta > 1 - 2\kappa$ and $\gamma_n = o(n)$, the Sargan test still accepts asymptotically.

4.2.4 *Mixture of globally valid and invalid IVs, locally valid*

Next, assume that the incorrect set of instruments has been selected as valid. A mixture of instruments from different groups has been selected instead. For example, some invalid instruments have been selected as valid and some valid IVs have been correctly selected as valid. Recall the model when some invalid IVs are wrongly chosen as valid, as in Equation (51):

$$\mathbf{y} = \mathbf{X}_{\mathcal{A}}\boldsymbol{\theta}_{\mathcal{A}} + \mathbf{Z}_1\boldsymbol{\alpha}_1 + \mathbf{u}.$$

Proposition 8. *When testing a mixture of globally valid and invalid IVs, if* $\boldsymbol{\tau} = o_p(1)$, *then*

$$Sar\left(\widehat{\boldsymbol{\theta}}_{\mathcal{A}}\right) = O_p(n)$$

4.2.5 *Mixture of globally valid and invalid IVs, with local violations*

Next, assume we have selected instruments incorrectly, and hence we have a mixture of valid and invalid instruments, but there are local violations τ_j. The error in the model with too few IVs selected as valid is now

$$\boldsymbol{\xi} = \mathbf{Z}_1\boldsymbol{\alpha}_1 + \mathbf{Z}\boldsymbol{\tau} + \mathbf{u}. \tag{59}$$

Proposition 9. *When testing a mixture of globally valid and invalid IVs, if* $\boldsymbol{\tau} = o_p(1)$, *then*

$$Sar\left(\widehat{\boldsymbol{\theta}}_{\mathcal{A}}\right) = O_p(n)$$

4.2.6 *Summary of results for Sargan test*

Table 18 includes a summary of the results from the preceding subsections. The first column shows the size of local violations $\boldsymbol{\tau}$, the second column shows the behavior of the

Table 18: Summary of Sargan results

	Same group		Mixture	
Local violations	Sargan	Needed to accept	Sargan	Needed to reject
None	$Sar\left(\widehat{\boldsymbol{\theta}}_{\mathcal{I}}\right) \xrightarrow{d} \chi^2_{J-K-1}$	$\delta_\gamma > 0$	$O_p(n)$	$\gamma_n = o(n)$
Minor	$Sar\left(\widehat{\boldsymbol{\theta}}_{\mathcal{I}}\right) \xrightarrow{d} \chi^2_{J-K-1}$	$\delta_\gamma > 0$	$O_p(n)$	$\gamma_n = o(n)$
Mild	$Sar\left(\widehat{\boldsymbol{\theta}}_{\mathcal{I}}\right) \xrightarrow{d} O_p(1)$	$\delta_\gamma > 0$	$O_p(n)$	$\gamma_n = o(n)$
Major	$Sar\left(\widehat{\boldsymbol{\theta}}_{\mathcal{I}}\right) = O_p(n^{1-2\kappa})$	$\gamma_n > o(n^{1-2\kappa})$	$O_p(n)$	$\gamma_n = o(n)$

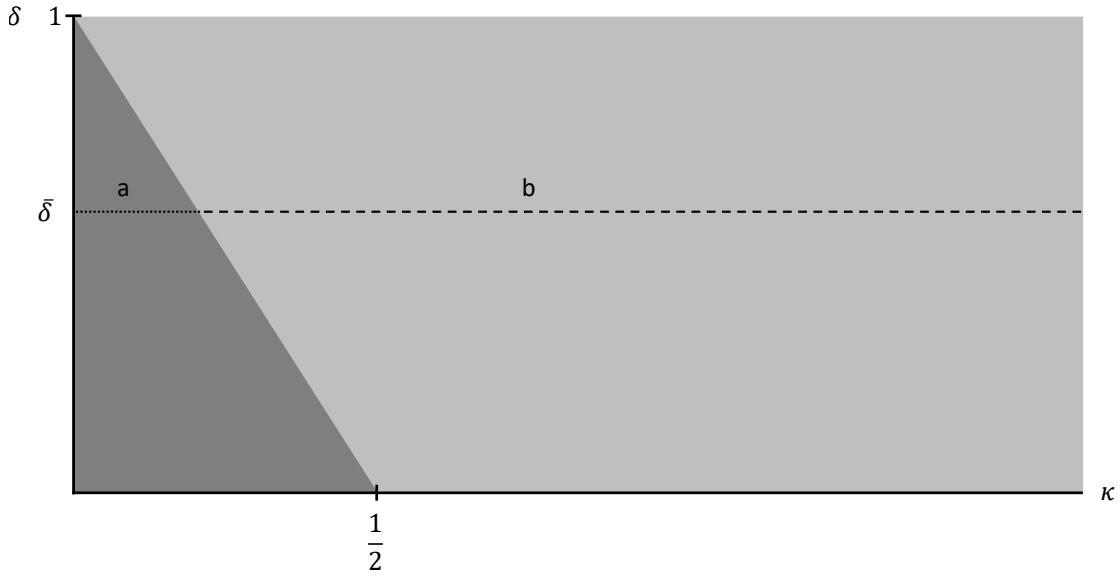
Note: This table summarizes the behavior of the Sargan test with local violations $\boldsymbol{\tau} = \frac{\mathbf{c}}{n^\kappa}$. The first column denotes the size of violations, depending on κ. Minor stands for $1/2 < \kappa < \infty$, mild stands for $\kappa = 1/2$ and major stands for $0 < \kappa < 1/2$. The second and third columns concern the behavior of the Sargan test when IVs are in the same group. The second column denotes the asymptotic behavior of the Sargan test. The third column shows how large the critical values need to be in order for the test to correctly reject. The fourth and fifth columns concern the Sargan test when there is a mixture of IVs from different groups.

Sargan test when all IVs are globally valid and the third column shows the order of growth needed for the critical values of the Sargan test in order for the test to correctly accept. In columns 4 and 5 I repeat the last two columns, for a mixture of instruments. Now the order of the critical values denotes how large they need to be in order for the test to correctly reject. With the help of these results, we can establish under which circumstances the downward testing procedure continues to work correctly, even in presence of local violations.

In brief, with instruments from the same group, as long as $\boldsymbol{\tau} \leq \frac{\mathbf{c}}{\sqrt{n}}$ and $\gamma_n \to \infty$ the Sargan test accepts asymptotically. For mixtures, the statistic will always be $O_p(n)$ and as long as $\gamma_n = o(n)$, the test rejects asymptotically. For major violations, i.e. $\boldsymbol{\tau} > \frac{\mathbf{c}}{\sqrt{n}}$ the critical values need to be large enough in order to avoid to wrongly reject asymptotically.

Figure 11 summarizes the above results for the case with *major* violations (last row in table 18). This visualization is very similar to Figure 10. The horizontal axis still displays the size of local violations. The vertical axis shows the δ for the critical values $\gamma_n = o(n^\delta)$ of the Sargan statistic. The gray areas denote the combinations of convergence rates of γ_n and local violations that are allowed in order for the downward testing procedure to be consistent. One special case is given when $\kappa_I \leq \frac{1-\delta}{2} < \kappa_V$ and $\gamma_n = o(n^\delta)$: then the Sargan test will also reject for invalid IVs of the same group. This case is given when the local invalidity parameter of the globally invalid instruments lies in the dark gray area, for example on line a, and the local validity parameter of the globally valid IVs lies in the light gray area, for example on line b. Jointly with the special case from Corollary 4, this special case will help us determining when the plurality assumption can be relaxed.

Figure 11: Visualization of results for Sargan



Note: This graph visualizes for which values of κ (horizontal axis) in $\tau = \frac{c}{n^{\kappa}}$, for a given δ (on vertical axis) in $\gamma_n = o(n^{\delta})$, consistency of the downward testing procedure still holds. The light gray area is closed to the top, open to the right and does not include the border line $\delta = 1 - 2 \cdot \kappa$. For a given δ, if κ is in the light gray area, the downward testing procedure is consistent - the Sargan test accepts for instruments from the same group and rejects for IVs from different groups. The dashed line (b) visualizes all possible values for κ for which that holds. If it is in the dark gray area, the downward testing procedure is not consistent anymore. If $\kappa \leq \frac{1-\delta}{2}$, i.e. κ is on the dotted line (a), including the border of the dark and light grey areas, the testing procedure is not consistent anymore.

5 IMPLICATIONS

In this section, I discuss two possible implications of these results. The first one is that the presence of local violations does not make it possible to allow for larger sequences of critical values in the CIM and in the downward testing procedure. The second one is that local violations make it possible to relax the plurality assumption in some cases.

5.1 *Larger sequences of critical values not allowed*

One might think that in presence of local violations larger sequences of critical values ψ_n or γ_n are allowed, without invalidating the proofs of (no) overlap and of consistency of the downward testing procedure. In this subsection, I discuss that this is not the case and that the standard result still applies. D. W. Andrews (1999) shows that downward testing procedures are consistent when $\gamma_n = o(n)$ and hence this rate of convergence is also relevant for the IV selection methods. In WLHB, critical values of the CIs have to be $\psi_n = o(n^{1/2})$. Therefore, I am interested to have $\gamma_n > o(n)$ or $\psi_n > o(n^{1/2})$.

Larger critical values are not allowed in presence of local invalidity ($\boldsymbol{\tau} \neq 0$). This becomes clear from columns 4 and 5 of Table 18: if $\gamma_n > o(n)$, the Sargan test will falsely

accept, as $Sar\left(\widehat{\boldsymbol{\theta}}_{\mathcal{A}}\right) = O_p(n)$. The downward testing procedure can not be consistent under such circumstances. This is the standard result from the literature and cases with $\boldsymbol{\tau} \neq 0$ do not affect this result.

Also, following Proposition 3, ψ_n has to be $o(n^{1/2})$ so that the CIs of different groups don't overlap. Overall there is no scope for larger sequences of critical values, neither in the CIM algorithm nor in the downward testing procedure.

5.2 *Relaxing the plurality assumption*

In presence of local violations, under some conditions, the plurality assumption can be violated and the procedures will still select the group of valid instruments as valid. To avoid confusion, I refer to the δ discussed in the CIM as δ_ψ and the one discussed in the downward testing procedure as δ_γ. Fixed values for δ are denoted by $\bar{\delta}$.

Proposition 10. *Relaxing plurality*
Under Assumptions 11-15, when $\kappa_I \leq \frac{1}{2} - \bar{\delta}_\psi < \kappa_V$ *and* $\kappa_I \leq \frac{1-\bar{\delta}_\gamma}{2} < \kappa_V$ *hold, the CIM consistently selects the valid IVs.*

The first assumption on κ_I and κ_V guarantees that the CIs of the same invalid group do not overlap, while those of the valid group do. The second part guarantees the consistency of the downward testing procedure, while the Sargan test rejects for IVs from the same invalid group.

The key thing to note here is that Assumption 16 is not needed anymore now. This is the case, because the CIs of invalid IVs from the same group do not overlap and the Sargan test also rejects for them, making the plurality assumption superfluous. For the AHC method, the assumption on δ_ψ is not needed, because the valid group is on the path as long as $\boldsymbol{\tau} = o_p(1)$.

6 SIMULATIONS

In this section, I run Monte Carlo simulations to illustrate the relevance of the points made in the preceding sections. The simulations begin with the standard case, without local violations of the exclusion restriction. Then, local violations are allowed and I can verify whether the predictions with respect to the selection results hold. In particular, I can also verify whether in the special case with minor local violations for globally valid and major local violations for globally invalid IVs the plurality assumption can be relaxed.

As discussed in Conley, Hansen, and Rossi (2012) the local invalidity asymptotics should not be taken as the true data-generating process, but as an approximation. Nevertheless, in a simulation, settings with mild, minor and major local violations can be compared to each other so as to get a feeling for the performance of IV selection methods in more realistic settings, where there are no exact groups.
The data are generated from

Table 19: Simulation results

Violation	n	Bias	Coverage	Or Bias	Or Cover	Nr inv	Freq inv	Freq or
$\tau = 0$	500	0.0258	0.708	0.015	0.957	2.032	0.246	0.237
	10000	0.0035	0.947	0.0035	0.952	3.011	1	0.989
$\kappa_V = 0.8$	500	0.031	0.628	0.0163	0.929	2.011	0.249	0.244
	10000	0.0034	0.953	0.0034	0.954	3.008	1	0.992
$\kappa_V = 0.5$	500	0.071	0.095	0.0455	0.318	1.355	0.08	0.075
	10000	0.0099	0.351	0.01	0.347	3.009	1	0.991
$\kappa_V = 0.2$	500	0.2544	0	0.2868	0	1.062	0.003	0.002
	10000	0.1928	0	0.1587	0	3.817	0.017	0
$\kappa_V = 0.9$	500	0.7363	0.003	0.0232	0.945	5.287	0.005	0.005
$\kappa_I = 0.1$	10000	0.0057	0.945	0.0052	0.952	7.011	0.999	0.989

Note: This table shows the results of Monte Carlo simulations. The first column (Violation) shows the order of local violations with $\boldsymbol{\tau} = \frac{\mathbf{c}}{n^{\kappa}}$. The second column (n) shows the sample size, the third (Bias) shows the mean absolute error of the post-CIM 2SLS, the fourth column (Coverage) shows the coverage rate of the 95 % CI of post-CIM 2SLS. The fifth column (Or Bias) displays the mean absolute error of the oracle 2SLS. The sixth column (Or Cover) shows the coverage rate of the CI for oracle 2SLS. The seventh column (Nr inv) shows the number of instruments selected as invalid. The correct number of invalid IVs is 3 in all lines except for the last two, where it is 7 (see details in text). The eighth column (Freq inv) shows the frequency with which all instruments have been selected as invalid. The ninth column (Freq or) shows the frequency with which the oracle model has been selected, i.e. all valid IVs have been selected as valid and all invalid have been selected as invalid.

$$d_i = \mathbf{z}_i'\boldsymbol{\gamma} + \varepsilon_i$$
$$y_i = d_i\beta + \mathbf{z}_i'(\boldsymbol{\alpha} + \boldsymbol{\tau}) + u_i$$

with

$$\begin{pmatrix} u \\ \varepsilon \end{pmatrix} \sim N\left(\mathbf{0}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right),$$

the instruments are $\mathbf{z}_i = N(0, \mathbf{I}_J)$, the true coefficient is the scalar $\beta = 0$ and the first-stage coefficient vector is $\boldsymbol{\gamma} = (1, 1, 1, ..., 1)'$. The number of observations is $n \in \{500; 10,000\}$. The number of repetitions used in the simulation is 1000.

I simulate ten IVs and three of these are globally invalid, with $\alpha_j = 0.2$. This means that three IVs are in a group with $q = 0.2$ and the remaining instruments are in a group with $q = 0$. The results of the simulation are displayed in Table 19. In the standard case as in WLHB there is no local invalidity, i.e. $\boldsymbol{\tau} = \mathbf{0}$. Here, we expect the selection to work well and the bias of post-CIM 2SLS to be low. In the first two lines of Table 19 the results for this case are shown. When the number of observations is 500, the bias is close to the oracle bias but the coverage rate of the 90% CI is too low, at 0.7. The oracle coverage rate is correct, at 0.95. The CIM underselects, with only two IVs selected as invalid on average and the oracle model selected in only 24 percent of cases. As expected, when the number of

observations is 10,000 we can see clearly that the correct number of IVs (three) is selected as invalid on average, all invalid instruments are always selected as invalid and the oracle model is selected in 99 percent of all repetitions. The bias is low for the post-selection 2SLS and it is equal to that of the oracle 2SLS (after rounding). The coverage rate of the 2SLS confidence intervals are close to 95%.

In the third and fourth lines I simulate a setting with $\kappa_V = 0.8$, which means that violations are minor for globally valid instruments. The vector $\mathbf{c}_V$ in $\boldsymbol{\tau}_V = \frac{c}{n^{\kappa_V}}$ is set to $\mathbf{c}_V = (0.7, 0.8, 0.9, ..., 1.2, 1.3)'$. According to Proposition 2, for a ψ_n that goes to infinity with any growth rate the CIs of globally valid IVs should still overlap and those of different groups should not overlap. With $n = 500$, the results are comparable to before: the mean absolute error is still low, at 0.03, but almost double that of the oracle. The method underselects and again in only 24 percent of the repetitions the oracle model has been selected. For a large sample size, the expected behavior is in fact reflected in the results: the bias is very low at 0.0034 and equal to the oracle bias. The coverage rate again is at about 0.95. The number of IVs is 3.008 and the oracle model is selected in 992 out of 1000 cases. This confirms that the CIM retains its oracle properties with minor violations of the exclusion restriction.

The fifth and sixth row show results where $\kappa_V = 0.5$. With $n = 500$, the selection results clearly deteriorate along all of the reported dimensions as compared to the case without or with minor violations. For example in only 75 out of 1000 repetitions the oracle model is correctly selected. For $n = 10,000$, the selection performance is still almost perfect, with 3.009 IVs selected on average and the oracle model being selected in 99% of the cases. However, the 2SLS-estimator now has an asymptotic bias. Therefore, it might be the case that the globally valid instruments are selected, but due to the asymptotic bias, the results are still misleading. Especially, the coverage rate of the post-selection 2SLS is at about 0.35. This is also the case for the oracle 2SLS.

In the seventh and eighth lines κ_V is set to 0.2 which is a case of a major violation of the exclusion restriction. Now, the exogeneity error completely dominates the asymptotic behavior and we expect a poor performance, even for large sample sizes. Unsurprisingly, for $n = 500$, selection results deteriorate even further as compared to the preceding settings. Oracle bias and post-CIM bias are both high and the coverage rates of the CI from the oracle model is zero. Moreover, the oracle model is selected only two times. With 10,000 observations, more than three instruments are selected as invalid, on average only 17 out of 1000 times all invalid IVs have been selected as invalid and the oracle model is never selected. This leads to a bias which is at about 0.19 and is higher than the oracle bias. The coverage rates of the CIs are still zero. In sum, as expected in this setting the method breaks down.

The ninth and tenth lines illustrate the case where globally invalid instruments have major local violations and the globally valid IVs have only minor violations. Proposition 10 states that when the critical values of the CIM and the downward testing procedure lie in the range described by the proposition, the largest group in terms of $\frac{\alpha_j}{\gamma_j}$ can consist of invalid instruments and the CIM will still stop at the smaller group which consists of

invalid IVs. To achieve this, three IVs are set to be globally valid with local invalidity $\boldsymbol{\tau}_V$ such that $\kappa_V = 0.9$ with $\mathbf{c}_V = (0.7, 1, 1.3)'$ and seven IVs globally invalid with $\frac{\alpha_j}{\gamma_j} = 0.2$, κ_I is set to 0.1 with $\mathbf{c}_I = (0.7, 0.8, 0.9, ..., 1.2, 1.3)'$. For $n = 500$, the method still performs poorly. The results for $n = 10000$, however, confirm the implication of Prediction 10: on average 7.01 IVs are selected as invalid, the instruments selected as invalid always include the invalid ones and in 99% percent of the cases the oracle model is selected. The bias of post-CIM 2SLS is very low at 0.0057 and it is very close to that of the oracle 2SLS. The coverage rate of the CIs is again close to 0.95.

Overall, the simulations confirm the implications from the preceding sections. Minor violations do not deteriorate the performance of the CIM and major violations can severely impair the quality of selection. In the special case where the invalid instruments also suffer major local violations and the valid IVs only have minor local violations, the plurality assumption might be relaxed.

7 CONCLUSION

In this chapter, I have shown under which circumstances the results of WLHB, Apfel and Liang (2021) and D. W. Andrews (1999) still hold. This is the case when the local invalidity that I modeled in addition to the global invalidity parameters is not too large, as compared to the critical values in the CIM and in the downward testing procedure. The results imply that when local invalidity of globally invalid instruments is large, the plurality rule can be violated, without affecting the selection result of the discussed algorithms. A Monte Carlo simulation confirms these findings.

These results underscore the practical importance of IV selection methods. Even when IVs do not perfectly fulfill the exclusion restriction, and there are no exact groups, the methods are still guaranteed to have oracle properties when the local violations of globally valid instruments are not too large in relation with the critical values of the CIM and the downward testing procedure.

In this setting, I have weakened the strict exclusion restriction but upheld the relevance condition (Assumption 11). Future work could consider the behavior of the methods in presence of weak instruments and near exclusion.

Appendices

F PROOFS

Proof of Proposition 2

Proof: In order for same-group CIs to overlap with probability 1 in the limit, in the proof of Lemma 1 in Appendix A.1 in WLHB, the following needs to hold

$$\lim_{n\to\infty} P\left(\sqrt{n}\frac{((\hat{\beta}_q-\hat{\beta}_k)-(\beta_q-\beta_k))}{\sqrt{\sigma_q+\sigma_k-2\sigma_{qk}}} > -\frac{\psi_n(\hat{\sigma}_q+\hat{\sigma}_k)-\sqrt{n}(\beta_q-\beta_k)}{\sqrt{\sigma_q+\sigma_k-2\sigma_{qk}}}\right)=1 \tag{60}$$

where k and q denote instruments from the valid group. Without local invalidity, i.e. with $\boldsymbol{\tau}=0$ it follows $\beta_q-\beta_k=0$ and the result is the same expression as in WLHB. As $\psi_n\to\infty$, the inequality holds. Note that by Assumption 4 in WLHB, the denominators can not be zero.

With $\boldsymbol{\tau}_V\neq 0$, however, we get $\beta_q-\beta_k=\frac{\tau_q}{\gamma_q}-\frac{\tau_k}{\gamma_k}\neq 0$ and hence the term on the right-hand side can become positive. Since $-\psi_n(\hat{\sigma}_q+\hat{\sigma}_k)-\sqrt{n}(\frac{c_q/(n^\kappa)}{\gamma_q}-\frac{c_k/(n^\kappa)}{\gamma_k})=-\psi_n(\hat{\sigma}_q+\hat{\sigma}_k)-n^{1/2-\kappa}(\frac{c_q}{\gamma_q}-\frac{c_k}{\gamma_k})$, this is guaranteed *not* to be the case when $lim\frac{\psi_n}{n^{1/2-\kappa}}\neq 0$, i.e. $\psi_n=O(n^\delta)$ with $\delta>1/2-\kappa$. This means that ψ_n diverges faster than $n^{1/2-\kappa}$. When $\boldsymbol{\tau}=\mathbf{c}/(n^\kappa)$, where $\kappa>0$, and when $\psi_n=O(n^\delta)$, $\delta>1/2-\kappa$ and $\psi_n\neq o(n^{1/2-\kappa})$, the CIs of the same group still overlap. This also implies that in order for the groups to overlap $\kappa>1/2-\delta$, by rearranging. □

Proof of Proposition 3

Proof: Let q a valid IV and s an invalid IV or two instruments from different groups. For $\beta_s>\beta_q$ (i.e. $\alpha_s>0$) without loss of generality the following needs to hold in order for the CIs to not overlap

$$\lim_{n\to\infty} P\left(\sqrt{n}\frac{((\hat{\beta}_q-\hat{\beta}_s)-(\beta_q-\beta_s))}{\sqrt{\sigma_q+\sigma_k-2\sigma_{qk}}} < \frac{-\psi_n(\hat{\sigma}_q+\hat{\sigma}_s)+\sqrt{n}(\beta_s-\beta_q)}{\sqrt{\sigma_q+\sigma_k-2\sigma_{qk}}}\right)=1 \tag{61}$$

with $\boldsymbol{\tau}=0$ this holds because the left-hand side is $N(0,1)$ and $\psi_n=o(n^{1/2})$ and hence the right-hand side goes to positive infinity. With $\boldsymbol{\tau}\neq 0$, $\beta_s-\beta_q=\frac{\alpha_s}{\gamma_s}+\frac{\tau_s}{\gamma_s}-\frac{\tau_q}{\gamma_q}$ and in order for this to not change the sign of the RHS, note that $-\psi_n(\hat{\sigma}_q+\hat{\sigma}_s)+\sqrt{n}(\frac{\alpha_s}{\gamma_s})+\sqrt{n}(\frac{c_s/(n^\kappa)}{\gamma_s}-\frac{c_q/(n^\kappa)}{\gamma_q})=-\psi_n(\hat{\sigma}_q+\hat{\sigma}_s)+\sqrt{n}(\frac{\alpha_s}{\gamma_s})+n^{1/2-\kappa}(\frac{c_s}{\gamma_s}-\frac{c_q}{\gamma_q})$ and the third summand will always be outweighed by the second one, except when $\kappa=0$, which is not possible, as otherwise the violations would not be local anymore. Hence, as $\psi_n\to\infty$, $\psi_n=o(n^{1/2})$ and $\boldsymbol{\tau}=\frac{\mathbf{c}}{n^\kappa}$, the CIs of different groups don't overlap. □

Proof of Proposition 4

We know that $\frac{1}{n}\mathbf{Z}'\mathbf{u} \xrightarrow{P} 0$, $\frac{1}{n}\mathbf{Z}'\mathbf{Z} \xrightarrow{P} \mathbf{Q}_{ZZ}$, $\frac{1}{n}\mathbf{Z}'\mathbf{X} \xrightarrow{P} \mathbf{Q}_{ZX}$ and $\frac{1}{\sqrt{n}}\mathbf{Z}'\mathbf{u} \xrightarrow{d} N(0, \sigma_u^2\mathbf{Q}_{ZZ})$, as $n \to \infty$.

Note that if all invalid instruments are selected correctly and controlled for, the estimator of $\boldsymbol{\theta}$ is consistent:

$$\widehat{\boldsymbol{\theta}}_{\mathcal{I}} \xrightarrow{P} \boldsymbol{\theta}_{\mathcal{I}}. \tag{62}$$

The residual can be written as:

$$\begin{aligned}\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}}) &= \mathbf{y} - \mathbf{X}_{\mathcal{I}}\widehat{\boldsymbol{\theta}} = \mathbf{X}_{\mathcal{I}}\boldsymbol{\theta}_{\mathcal{I}} + \mathbf{u} - \mathbf{X}_{\mathcal{I}}\widehat{\boldsymbol{\theta}} \\ &= \mathbf{X}_{\mathcal{I}}\boldsymbol{\theta}_{\mathcal{I}} + \mathbf{u} - \mathbf{X}_{\mathcal{I}}(\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{I}})^{-1}\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{y} \\ &= \mathbf{X}_{\mathcal{I}}\boldsymbol{\theta}_{\mathcal{I}} + \mathbf{u} - \mathbf{X}_{\mathcal{I}}(\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{I}})^{-1}\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z(\mathbf{X}_{\mathcal{I}}\boldsymbol{\theta}_{\mathcal{I}} + \mathbf{u}) \\ &= \mathbf{u} - \mathbf{X}_{\mathcal{I}}(\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{I}})^{-1}\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{u}.\end{aligned} \tag{63}$$

Note that in $\mathbf{u}'\mathbf{P}_Z\mathbf{X}_{\mathcal{I}}(\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{I}})^{-1}\mathbf{X}_{\mathcal{I}}'\mathbf{X}_{\mathcal{I}}(\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{I}})^{-1}\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{u}$ we have that $plim(\mathbf{u}'\mathbf{Z}) = 0$, while $plim(\mathbf{Z}'\mathbf{Z})$, $plim(\mathbf{X}_{\mathcal{I}}'\mathbf{Z})$ and $plim(\mathbf{X}_{\mathcal{I}}'\mathbf{X}_{\mathcal{I}})$ all converge against finite, non-zero matrices. In this term, we have an uneven number of terms that converge, one of them to zero. Therefore, dividing by n in the limit leads to a product of finite matrices with zero. All terms except for $\mathbf{u}'\mathbf{u}/n$ disappear for this reason. Hence

$$\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}})'\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}})/n \xrightarrow{P} \sigma_u^2.$$

Also, the numerator of the Sargan statistic can be written as

$$\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}})'\mathbf{P}_Z\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}}) = \widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}})'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1/2}(\mathbf{Z}'\mathbf{Z})^{-1/2}\mathbf{Z}'\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}}). \tag{64}$$

Plugging in (63), I get

$$(\mathbf{Z}'\mathbf{Z})^{-1/2}\mathbf{Z}'\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}}) = (\mathbf{I} - (\mathbf{Z}'\mathbf{Z})^{-1/2}\mathbf{Z}'\mathbf{X}_{\mathcal{I}}(\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{I}})^{-1}\mathbf{X}_{\mathcal{I}}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1/2})(\mathbf{Z}'\mathbf{Z})^{-1/2}\mathbf{Z}'\mathbf{u} \tag{65}$$

With the assumptions shown at the beginning, it can be seen that

$$(\mathbf{Z}'\mathbf{Z})^{-1/2}\mathbf{Z}'\mathbf{u} \xrightarrow{d} N(0, \sigma_u^2\mathbf{I}) \tag{66}$$

as $(\frac{\mathbf{Z}'\mathbf{Z}}{n})^{-1/2} = \mathbf{Q}_{ZZ}^{-1/2}$, $\frac{\mathbf{Z}'\mathbf{u}}{\sqrt{n}} \xrightarrow{d} N(0, \sigma_u^2\mathbf{Q}_{ZZ})$ and hence

$$\frac{(\mathbf{Z}'\mathbf{Z})^{-1/2}\mathbf{Z}'\mathbf{u}}{\sqrt{\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}})'\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}})/n}} \xrightarrow{d} N(0, \mathbf{I}) \tag{67}$$

Consider $\mathbf{I} - (\mathbf{Z}'\mathbf{Z})^{-1/2}\mathbf{Z}'\mathbf{X}_{\mathcal{I}}(\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{I}})^{-1}\mathbf{X}_{\mathcal{I}}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1/2}$, a symmetric, idempotent projection matrix that goes to a finite square matrix $\mathbf{A}$ in probability and hence:

$$Sar\left(\widehat{\boldsymbol{\theta}}_{\mathcal{I}}\right) \xrightarrow{d} \boldsymbol{\zeta}'\mathbf{A}\boldsymbol{\zeta} \sim \chi^2_{J-K-1} \tag{68}$$

where $\boldsymbol{\zeta} \sim N(0, \mathbf{I})$. This holds by Theorem 2 in Searle (1971, p.57-59). □

Proof of Proposition 5

The 2SLS estimator is

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_{\mathcal{I}} &= (\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{I}})^{-1}\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{y} \\ &= (\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{I}})^{-1}\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z(\mathbf{X}_{\mathcal{I}}\boldsymbol{\theta}_{\mathcal{I}} + \mathbf{Z}\boldsymbol{\tau} + \mathbf{u}) \\ &= \boldsymbol{\theta}_{\mathcal{I}} + (\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{I}})^{-1}\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z(\mathbf{Z}\boldsymbol{\tau} + \mathbf{u}) \\ &= \boldsymbol{\theta}_{\mathcal{I}} + (\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{I}})^{-1}\mathbf{X}_{\mathcal{I}}'\mathbf{Z}\boldsymbol{\tau} + (\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{I}})^{-1}\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{u}\end{aligned} \tag{69}$$

If $\boldsymbol{\tau} = o_p(1)$, then again $\widehat{\boldsymbol{\theta}}_{\mathcal{I}} \xrightarrow{P} \boldsymbol{\theta}_{\mathcal{I}}$. Then, the residual becomes

$$\begin{aligned}\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}}) &= \mathbf{X}_{\mathcal{I}}\boldsymbol{\theta}_{\mathcal{I}} + \mathbf{Z}\boldsymbol{\tau} + \mathbf{u} - \mathbf{X}_{\mathcal{I}}\widehat{\boldsymbol{\theta}}_{\mathcal{I}} \\ &= \mathbf{Z}\boldsymbol{\tau} + \mathbf{u} - \mathbf{X}_{\mathcal{I}}(\boldsymbol{\theta}_{\mathcal{I}} - \widehat{\boldsymbol{\theta}}_{\mathcal{I}}) \\ &= \mathbf{Z}\boldsymbol{\tau} + \mathbf{u} - \mathbf{X}_{\mathcal{I}}(\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{I}})^{-1}\mathbf{X}_{\mathcal{I}}'\mathbf{Z}\boldsymbol{\tau} - \mathbf{X}_{\mathcal{I}}(\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{I}})^{-1}\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{u}\end{aligned} \tag{70}$$

Hence:

$$\begin{aligned}\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}})'\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}})/n = &(\mathbf{Z}\boldsymbol{\tau} + \mathbf{u} - \mathbf{X}_{\mathcal{I}}(\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{I}})^{-1}\mathbf{X}_{\mathcal{I}}'\mathbf{Z}\boldsymbol{\tau} - \mathbf{X}_{\mathcal{I}}(\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{I}})^{-1}\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{u})' \\ &(\mathbf{Z}\boldsymbol{\tau} + \mathbf{u} - \mathbf{X}_{\mathcal{I}}(\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{I}})^{-1}\mathbf{X}_{\mathcal{I}}'\mathbf{Z}\boldsymbol{\tau} - \mathbf{X}_{\mathcal{I}}(\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{I}})^{-1}\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{u}) \\ = &\frac{\boldsymbol{\tau}'\mathbf{Z}'\mathbf{Z}\boldsymbol{\tau}}{n} + \frac{\mathbf{u}'\mathbf{u}}{n} - 2\frac{\mathbf{u}'\mathbf{X}_{\mathcal{I}}(\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{I}})^{-1}\mathbf{X}_{\mathcal{I}}'\mathbf{Z}\boldsymbol{\tau}}{n} \\ &- 2\frac{\boldsymbol{\tau}'\mathbf{Z}'\mathbf{X}_{\mathcal{I}}(\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{I}})^{-1}\mathbf{X}_{\mathcal{I}}'\mathbf{Z}\boldsymbol{\tau}}{n} \\ &+ \frac{\boldsymbol{\tau}'\mathbf{Z}'\mathbf{X}_{\mathcal{I}}(\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{I}})^{-1}\mathbf{X}_{\mathcal{I}}'\mathbf{X}_{\mathcal{I}}(\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{I}})^{-1}\mathbf{X}_{\mathcal{I}}'\mathbf{Z}\boldsymbol{\tau}}{n} \\ &+ o_p(1).\end{aligned} \tag{71}$$

The terms that involve ($\frac{\mathbf{u}'\mathbf{Z}}{n} \xrightarrow{P} 0$) go to zero in probability and are subsumed by $o_p(1)$. This term goes to σ_u^2 in probability, when $\boldsymbol{\tau} = o_p(1)$.

The the root of the numerator of the Sargan test, the term from equation 65, now becomes

$$\begin{aligned}(\mathbf{Z}'\mathbf{Z})^{-1/2}\mathbf{Z}'\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}}) &= (\mathbf{Z}'\mathbf{Z})^{-1/2}\mathbf{Z}'(\mathbf{Z}\boldsymbol{\tau} + \mathbf{u} - \mathbf{X}_{\mathcal{I}}(\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{I}})^{-1}\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z(\mathbf{Z}\boldsymbol{\tau} + \mathbf{u})) \\ &= (\frac{\mathbf{Z}'\mathbf{Z}}{n})^{-1/2}\frac{\mathbf{Z}'\mathbf{Z}}{\sqrt{n}}\boldsymbol{\tau} + (\mathbf{Z}'\mathbf{Z})^{-1/2}\mathbf{Z}'\mathbf{u} \\ &\quad - (\mathbf{Z}'\mathbf{Z})^{-1/2}\mathbf{Z}'\mathbf{X}_{\mathcal{I}}(\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{I}})^{-1}\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z(\mathbf{Z}\boldsymbol{\tau} + \mathbf{u}).\end{aligned} \tag{72}$$

In order for this term to converge to the same normal distribution as before all summands involving $\boldsymbol{\tau}$ have to disappear. This is the case when there are minor violations, i.e. the local violation vector is $\boldsymbol{\tau} = o_p(n^{-1/2})$. If that is true, it still holds that

$$Sar\left(\widehat{\boldsymbol{\theta}}_{\mathcal{I}}\right) \xrightarrow{d} \chi^2_{J-K_{\mathcal{I}}-1} \quad \square$$

Proof of Proposition 6

As $\boldsymbol{\tau} = o_p(1)$, as before

$$\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}})'\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}})/n \to \sigma_u^2$$

Following equation (72), the root of the numerator with local invalidity becomes

$$(\mathbf{Z}'\mathbf{Z})^{-1/2}\mathbf{Z}'\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}}) = \mathbf{M}(\mathbf{u} + \mathbf{Z}\frac{\mathbf{c}}{\sqrt{n}}) \tag{73}$$

where as before $\mathbf{M} = \mathbf{I} - (\mathbf{Z}'\mathbf{Z})^{-1/2}\mathbf{Z}'\mathbf{X}_{\mathcal{I}}(\mathbf{X}_{\mathcal{I}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{I}})^{-1}\mathbf{X}_{\mathcal{I}}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1/2}$ is an idempotent projection matrix that goes to $\mathbf{A}$ in probability. Still,

$$\frac{(\mathbf{Z}'\mathbf{Z})^{-1/2}\mathbf{Z}'\mathbf{u}}{\sqrt{\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}})'\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}})/n}} \xrightarrow{d} N(0, \mathbf{I})$$

and

$$\frac{(\mathbf{Z}'\mathbf{Z})^{-1/2}\mathbf{Z}'\mathbf{Z}\frac{\mathbf{c}}{\sqrt{n}}}{\sqrt{\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}})'\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}})/n}} = \frac{(\frac{\mathbf{Z}'\mathbf{Z}}{n})^{-1/2}\frac{\mathbf{Z}'\mathbf{Z}}{n}\mathbf{c}}{\sqrt{\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}})'\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}})/n}} \xrightarrow{P} \frac{1}{\sigma_u}\mathbf{Q}_{ZZ}^{1/2}\mathbf{c}$$

and hence the second part of equation (73) converges by Slutsky's Theorem:

$$\frac{(\mathbf{Z}'\mathbf{Z})^{-1/2}\mathbf{Z}'(\mathbf{u} + \mathbf{Z}\frac{\mathbf{c}}{\sqrt{n}})}{\sqrt{\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}})'\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}})/n}} \xrightarrow{d} N\left(\frac{1}{\sigma_u}\mathbf{Q}_{ZZ}^{1/2}\mathbf{c}, \mathbf{I}\right)$$

By Theorem 2 in Searle (1971, p. 57-59): when $\mathbf{x} \sim N(\boldsymbol{\mu}, \mathbf{I})$ and $\mathbf{A}$ is idempotent, $\mathbf{x}'\mathbf{A}\mathbf{x} \sim \chi^2_{J-K-1}(\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu})$, where $\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$ is the non-centrality parameter. For the Sargan statistic, this implies:

$$Sar\left(\widehat{\boldsymbol{\theta}}_{\mathcal{I}}\right) \xrightarrow{d} \chi^2_{J-K_{\mathcal{I}}-1}\left(\frac{1}{\sigma_u^2}\mathbf{c}'\mathbf{Q}_{ZZ}^{1/2\prime}\mathbf{A}\mathbf{Q}_{ZZ}^{1/2}\mathbf{c}\right) \quad \square$$

Proof of Proposition 7

As before, since $\boldsymbol{\tau} = o_p(1)$:

$$\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}})'\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}})/n \xrightarrow{P} \sigma_u^2.$$

The root of the numerator becomes

$$\frac{(\mathbf{Z}'\mathbf{Z})^{-1/2}\mathbf{Z}'(\mathbf{u} + \mathbf{Z}\frac{\mathbf{c}}{n^\kappa})}{\sqrt{\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}})'\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}})/n}} = \frac{(\frac{\mathbf{Z}'\mathbf{Z}}{n})^{-1/2}(\frac{\mathbf{Z}'\mathbf{u}}{\sqrt{n}} + \frac{\mathbf{Z}'\mathbf{Z}}{n}\frac{\mathbf{c}}{n^{\kappa-1/2}})}{\sqrt{\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}})'\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}})/n}} = n^{1/2-\kappa}\frac{(\frac{\mathbf{Z}'\mathbf{Z}}{n})^{-1/2}(\frac{\mathbf{Z}'\mathbf{u}}{n^{1-\kappa}} + \frac{\mathbf{Z}'\mathbf{Z}}{n}\mathbf{c})}{\sqrt{\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}})'\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}})/n}}$$

$(\frac{\mathbf{Z}'\mathbf{Z}}{n})^{-1/2}(\frac{\mathbf{Z}'\mathbf{u}}{n^{1-\kappa}})$ would converge in distribution when $\kappa = 1/2$ and goes to zero in probability when $\kappa < 1/2$. This follows because it converges to a normally distributed random variable scaled by $n^{1-\kappa}$. The term $\frac{(\frac{\mathbf{Z}'\mathbf{Z}}{n})^{-1/2}(\frac{\mathbf{Z}'\mathbf{Z}}{n}\mathbf{c})}{\sqrt{\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}})'\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{\mathcal{I}})/n}}$ converges in probability to $\frac{1}{\sigma_u}\mathbf{Q}_{ZZ}^{1/2}\mathbf{c}$ and therefore

$$\frac{Sar\left(\widehat{\boldsymbol{\theta}}_{\mathcal{I}}\right)}{n^{1-2\kappa}} \xrightarrow{P} \frac{1}{\sigma_u^2}\mathbf{c}'\mathbf{Q}_{ZZ}^{1/2\prime}\mathbf{A}\mathbf{Q}_{ZZ}^{1/2}\mathbf{c} \quad \square \tag{74}$$

Proof of Proposition 8

Under model (51) the estimator becomes

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_{\mathcal{A}} &= (\mathbf{X}_{\mathcal{A}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}_{\mathcal{A}}'\mathbf{P}_Z\mathbf{y} = (\mathbf{X}_{\mathcal{A}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}_{\mathcal{A}}'\mathbf{P}_Z(\mathbf{X}_{\mathcal{A}}\boldsymbol{\theta}_{\mathcal{A}} + \boldsymbol{\xi}) \\ &= \boldsymbol{\theta}_{\mathcal{A}} + (\mathbf{X}_{\mathcal{A}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}_{\mathcal{A}}'\mathbf{P}_Z\boldsymbol{\xi}\end{aligned}$$

The residual becomes

$$\hat{\boldsymbol{\xi}} = \widehat{\boldsymbol{\xi}}(\widehat{\boldsymbol{\theta}}_{\mathcal{A}}) = \boldsymbol{\xi} - \mathbf{X}_{\mathcal{A}}(\mathbf{X}_{\mathcal{A}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}_{\mathcal{A}}'\mathbf{P}_Z\boldsymbol{\xi}$$

To see to what the inner product of the residual converges, we look at each addend of the product: Note that $\boldsymbol{\xi}'\boldsymbol{\xi}/n \xrightarrow{P} \boldsymbol{\alpha}_1'\mathbf{Q}_{Z_1Z_1}\boldsymbol{\alpha}_1 + \sigma_u^2$. The terms $\boldsymbol{\xi}'\mathbf{Z}_{\mathcal{A}}/n$ and $\boldsymbol{\xi}'\mathbf{X}_{\mathcal{A}}/n$ all converge in probability to finite vectors and hence

$$\begin{aligned}&\boldsymbol{\xi}'\mathbf{X}_{\mathcal{A}}(\mathbf{X}_{\mathcal{A}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}_{\mathcal{A}}'\mathbf{P}_Z\boldsymbol{\xi}/n \\ &\text{and} \\ &\boldsymbol{\xi}'\mathbf{P}_Z\mathbf{X}_{\mathcal{A}}(\mathbf{X}_{\mathcal{A}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}_{\mathcal{A}}'\mathbf{X}_{\mathcal{A}}(\mathbf{X}_{\mathcal{A}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}_{\mathcal{A}}'\mathbf{P}_Z\boldsymbol{\xi}/n\end{aligned} \tag{75}$$

also do. Denote the first term as C_1 and the second term as C_2. Let $C = -2C_1 + C_2$. Then, the denominator of the Sargan-statistic converges in probability to σ_u^2 plus an inconsistency

$$\widehat{\boldsymbol{\xi}}(\widehat{\boldsymbol{\theta}}_{\mathcal{A}})'\widehat{\boldsymbol{\xi}}(\widehat{\boldsymbol{\theta}}_{\mathcal{A}})/n \xrightarrow{P} \sigma_u^2 + \boldsymbol{\alpha}_1'\mathbf{Q}_{Z_1Z_1}\boldsymbol{\alpha}_1 + C \tag{76}$$

The root of the numerator is

$$\begin{aligned}(\mathbf{Z}'\mathbf{Z})^{-1/2}\mathbf{Z}'\hat{\boldsymbol{\xi}} &= (\mathbf{Z}'\mathbf{Z})^{-1/2}\mathbf{Z}'(\boldsymbol{\xi} - \mathbf{X}_{\mathcal{A}}(\mathbf{X}_{\mathcal{A}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}_{\mathcal{A}}'\mathbf{P}_Z\boldsymbol{\xi}) \\ &= (\mathbf{I} - (\mathbf{Z}'\mathbf{Z})^{-1/2}\mathbf{Z}'\mathbf{X}_{\mathcal{A}}(\mathbf{X}_{\mathcal{A}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}_{\mathcal{A}}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1/2})(\mathbf{Z}'\mathbf{Z})^{-1/2}\mathbf{Z}'\boldsymbol{\xi} \\ &= (\mathbf{I} - (\mathbf{Z}'\mathbf{Z})^{-1/2}\mathbf{Z}'\mathbf{X}_{\mathcal{A}}(\mathbf{X}_{\mathcal{A}}'\mathbf{P}_Z\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}_{\mathcal{A}}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1/2})(\mathbf{Z}'\mathbf{Z})^{-1/2}\mathbf{Z}'(\mathbf{Z}_1\boldsymbol{\alpha}_1 + \mathbf{u}).\end{aligned}$$

For the expression behind the projection matrix it holds

$$\frac{1}{\sqrt{n}}(\mathbf{Z}'\mathbf{Z})^{-1/2}\mathbf{Z}'(\mathbf{Z}_1\boldsymbol{\alpha}_1 + \mathbf{u}) \xrightarrow{P} \mathbf{Q}_{ZZ}^{-1/2}\mathbf{Q}_{ZZ_1}\boldsymbol{\alpha}_1$$

Therefore, after dividing by n (because the last term with $\frac{1}{\sqrt{n}}$ appears twice), all of the matrices in the denominator and numerator of the Sargan test statistic converge in probability:

$$\frac{Sar\left(\widehat{\boldsymbol{\theta}}_{\mathcal{A}}\right)}{n} \xrightarrow{P} \frac{\boldsymbol{\alpha}_1' \mathbf{Q}_{ZZ_1}{}' \mathbf{Q}_{ZZ}^{-1/2\prime} \mathbf{A} \mathbf{Q}_{ZZ}^{-1/2} \mathbf{Q}_{ZZ_1} \boldsymbol{\alpha}_1}{\sigma_u^2 + \boldsymbol{\alpha}_1' \mathbf{Q}_{Z_1 Z_1} \boldsymbol{\alpha}_1 + C} \tag{77}$$

This means that $Sar\left(\widehat{\boldsymbol{\theta}}_{\mathcal{A}}\right) = O_p(n)$. □

Proof of Proposition 9

The inner product of the error term divided by n is $\boldsymbol{\xi}'\boldsymbol{\xi}/n \xrightarrow{P} \boldsymbol{\alpha}_1' \mathbf{Q}_{Z_1 Z_1} \boldsymbol{\alpha}_1 + \boldsymbol{\tau}' \mathbf{Q}_{ZZ} \boldsymbol{\tau} + \sigma_u^2$. Note that if $\boldsymbol{\tau} = o_p(1)$, this inner product as well as the terms $\boldsymbol{\xi}'\mathbf{Z}_{\mathcal{A}}/n$ and $\boldsymbol{\xi}'\mathbf{X}_{\mathcal{A}}/n$ and the terms in (75) still converge to the same finite matrices and scalars as before. Equation (76) therefore still holds. Also, if $\boldsymbol{\tau}$ vanishes, for the root of the numerator it still holds that

$$\frac{1}{\sqrt{n}} (\mathbf{Z}'\mathbf{Z})^{-1/2} \mathbf{Z}' (\mathbf{Z}_1 \boldsymbol{\alpha}_1 + \mathbf{Z}\boldsymbol{\tau} + \mathbf{u}) \xrightarrow{P} \mathbf{Q}_{ZZ}^{-1/2} \mathbf{Q}_{ZZ_1} \boldsymbol{\alpha}_1.$$

Hence it again holds that the Sargan test statistic divided by n converges in probability to the same expression as in (77) and the Sargan statistic is $O_p(n)$ □

V

FALSIFICATION ADAPTIVE SETS AND VALID INSTRUMENT SELECTION METHODS - A COMPARISON

1 INTRODUCTION

What should researchers do when an IV model with more instruments than endogenous regressors is falsified? If the data is not consistent with the overidentifying restrictions, the model is said to be falsified. When the model is falsified it is not clear how to proceed.

Recently, two approaches have been developed to deal with falsified IV models. The first one has been presented by Masten and Poirier (2021, MP) who propose to report sets which indicate the severity of falsification of the model: the falsification frontier and the falsification adaptive set. If the extent of falsification is small, then even though the model is falsified, standard results might still be trusted. The second approach is to select the invalid IV under appropriate conditions and to then control for the IVs. Kang, Zhang, Cai, and Small (2016) have first proposed such a selection method, which has inspired further methods, such as WLHB and Apfel and Liang (2021).

In this chapter, I set out to compare and combine the insights from both approaches. The research questions are: What can be learned from each strand of the literature? Can some settings be illuminated with the help of a combination of both approaches?

The key results are the following. First, there are a few connection points of the methods: both are based on the just-identified IV estimates which use remaining instruments as controls. Both make use of a first-stage screening of the weak and strong instruments. Moreover, the CIM can be restated in terms of quantities introduced in MP. Also, in their setting, the plurality assumption which is central for selection methods translates to a sparsity assumption for the vector of partial relaxations. The outcome of selection methods then is the point on the falsification frontier with the fewest non-zero relaxations.

Second, even though there are similarities, the two methods proceed differently. The falsification frontier looks at all minimal deviations, while selection methods rely on the assumption that a big enough group of IVs is valid. These differences make the methods useful complements: when the selection methods select a very small group of instruments as valid a researcher might want to fall back to the falsification adaptive set, which does not need the relatively strict plurality assumption.

Third, problems discussed in the IV selection literature might be also helpful for a better understanding of some issues in MP. Fourth, when using a model with partial relaxations of exogeneity, the results of MP do not hold. In fact, the analogous treatment leads to a different falsification frontier and a falsification adaptive set. This can be the point of departure for future research.

In Section 2, I start by summarizing both methods. I first present the model and the approach of MP and focus on the falsification frontier and the falsification adaptive set. Then, I present the instrument selection methods. I exemplify this with two recent methods: the CIM (Windmeijer, Liang, Hartwig, and Bowden, 2021) and the AHC (Apfel and Liang, 2021). These methods make use of tests of overidentifying restrictions, connecting them closely to the idea of falsified models, and hence I focus on them.

In Section 3, the two methods are compared. I first discuss similarities and differences between the two methods. Then, the CIM is restated in the framework of MP. I then

compare the two approaches in general, frame their differences from the perspective of the theory of science and present the two different angles from which the same problem is approached. I then discuss advantages and disadvantages of the two approaches to then propose the complementary use of both. In Section 4, I discuss problems which have been raised in the IV selection literature and that can offer insight when following the analysis proposed by MP.

In Section 5, I discuss violations of the exogeneity assumption and how these might affect the falsification adaptive set proposed by MP. I compare their FF and FAS with alternative quantities which can be derived analogously. This gives rise to the question of whether the original falsification frontier and falsification adaptive set can effectively account for all types of violations of the exogeneity assumption. These considerations can help to further improve analyses which use the falsification adaptive set. Section 6 concludes.

2 PRESENTATION OF METHODS

I first introduce notational conventions used throughout this chapter. Scalars are in lowercase, vectors are in lowercase bold and matrices in uppercase bold. The number of instruments is denoted by J. Instruments are indexed by j. The symbol $|\cdot|$ denotes the cardinality of a set, when used around a set, and the absolute value, when used around a quantity. $||\cdot||_0$ denotes the number of non-zero elements of a vector. $\xrightarrow{P}$ denotes convergence in probability. As before

$$\mathbf{\Gamma} = E(\mathbf{z}_i\mathbf{z}_i')^{-1}E(\mathbf{z}_i y_i) \quad \textit{and} \quad \boldsymbol{\gamma} = E(\mathbf{z}_i\mathbf{z}_i')^{-1}E(\mathbf{z}_i d_i).$$

with entries Γ_j and γ_j.

2.1 *Model and assumptions*

Consider the linear model

$$\mathbf{y} = \mathbf{d}\beta + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{u} \tag{78}$$

where $\mathbf{y}$ is the outcome variable, $\mathbf{d}$ is the treatment variable, β is the coefficient of interest, $\mathbf{Z}$ is a matrix of instruments, $\boldsymbol{\alpha}$ is the vector of direct effects and $\mathbf{u}$ is an idiosyncratic error. For ease of exposition I treat the case where the number of regressors P is one. The first-stage equation is

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

The assumptions are

Assumption 21. *Relevance: The $J \times K$ matrix $cov(\mathbf{z}_i, d_i)$ has rank K.*

Assumption 22. *Sufficient variation: The $J \times J$ matrix $var(\mathbf{z}_i)$ is invertible.*

Assumptions 1 and 2 are equivalent to Assumptions 1 and 2 in Windmeijer, Liang, Hartwig, and Bowden (2021, WLHB), i.e. Assumptions 11 and 12 in this manuscript.

Assumption 23. *Exogeneity:* $cov(\mathbf{z}_{ij}, u_i) = 0$ *for all* j.

This is a part of assumption 5 in WLHB. The next assumption is the center piece of the analysis in MP.

Assumption 24. *Partial exclusion:*
There are constant $\delta_j \geq 0$ *such that*

$$|\alpha_j| \leq \delta_j$$

for all $j \in \{1, ..., J\}$.

The assumption considered in WLHB, Guo, Kang, Cai, and Small (2018) and Apfel and Liang (2021) is that the largest group of IVs is valid (*plurality assumption*). This is Assumption 16. For the selection methods, moreover, assumptions 11 to 15 hold, as in the preceding two chapters.

The identified set $\mathcal{B}(\boldsymbol{\delta})$ is the subspace of the parameter space of β that could have produced the data under the given model and the relaxation $\boldsymbol{\delta}$. Under Assumption 4, it follows that in this case, the identified set is

Definition 6. *Identified set:*

$$\mathcal{B}(\boldsymbol{\delta}) = \{\beta \in \mathbb{R}^K : \ -\boldsymbol{\delta} \leq var(\mathbf{Z})^{-1}(cov(\mathbf{Z}, \mathbf{y}) - cov(\mathbf{Z}, \mathbf{d})\beta) \leq \boldsymbol{\delta}\}$$

Note that point identification means that it is possible to solve for unique values of the parameters of the structural equation. The parameter β can be uniquely determined from the data. An identified set is the set of all β that are observationally equivalent, but there are some β that are not observationally equivalent. Observational equivalence means that there exist multiple (unobserved) parameter values from the parameter space that lead to the same data distribution.

2.2 *Masten and Poirier (2021) - Salvaging falsified models*

I first introduce the study of Masten and Poirier (2021).

2.2.1 *Falsification*

The key concept in MP is falsification. Therefore it is helpful to understand what falsification means in this context. MP discuss settings in which the exclusion restriction is violated and this leads to falsification of the model. They study population level identification, that is they consider the case with $n \to \infty$. In this case, when the just-identified estimands from using one IV at a time are not the same, the overidentification tests will reject even for minimal violations of the exclusion restriction $\alpha_j = 0$. Often researchers try to make sense of 2SLS estimates, even when these tests reject the Null hypothesis.

The following statements are used equivalently by the authors:

- The model is falsified.
- $\frac{\Gamma_j}{\gamma_j} \neq \frac{\Gamma_l}{\gamma_l}$, for some j and l.
- $\frac{\alpha_j}{\gamma_j} \neq 0$ for at least one j.
- The identified set $\mathcal{B}(\boldsymbol{\delta} = 0)$ is empty.

The following statements are also equivalent

- The model is not falsified.
- $\frac{\Gamma_j}{\gamma_j} = \frac{\Gamma_l}{\gamma_l}$, for all j and l.
- $\frac{\alpha_j}{\gamma_j} = 0$ for all j.
- The identified set $\mathcal{B}(\boldsymbol{\delta} = 0)$ is not empty and is a singleton.

In MP, a more detailed definition of falsification is given.[1]

2.2.2 *Falsification frontier*

The falsification frontier is the minimal set of δ_js that lead to a non-empty identified set. MP suggest to report this frontier to better understand how far away from *non*-falsification the falsified baseline model is.

Definition 7. *Falsification frontier:*

$$FF = \left\{ \boldsymbol{\delta} \in \mathbb{R}^J_{\geq 0} : \delta_j = |\Gamma_j - \beta\gamma_j|,\ j = 1, ..., J,\ \beta \in \left[\min_j \frac{\Gamma_j}{\gamma_j},\ \max_j \frac{\Gamma_j}{\gamma_j} \right] \right\}$$

In other words, the falsification frontier answers the following question: How large must the δ_j at least be to produce the given $\frac{\Gamma_j}{\gamma_j}$? In the case with $J = 2$, the falsification frontier can be visualized as a line in $\mathbb{R}^2$. A falsification frontier which is far away from the origin, means that the deviations from a non-falsified model are large.

2.2.3 *Falsification adaptive set*

The second, important diagnostic introduced by MP is the falsification adaptive set. The falsification adaptive set is the identified set $\mathcal{B}(\boldsymbol{\delta})$, under the assumption that one of the points on the falsification frontier is true.

Definition 8. *Falsification adaptive set:*

$$\bigcup_{\boldsymbol{\delta} \in FF} \mathcal{B}(\boldsymbol{\delta}) = \left[\min_j \frac{\Gamma_j}{\gamma_j},\ \max_j \frac{\Gamma_j}{\gamma_j} \right]$$

If this set is wide, this is taken as evidence that the degree of falsification is large, while if it is small the degree of falsification is less severe.

[1] "*A given model $\mathcal{M}$ is falsifiable if there are some observed distributions F_W which could not have been generated by the model. If such a cdf is observed, we say the model is falsified (equivalently, refuted). For a given model, let $\mathcal{F}_f$ denote the set of cdfs F_W which falsify the model. Let $\mathcal{F}_{nf}$ denote the set of cdfs F_W which do not falsify the model.*" (Masten and Poirier, 2021, p. 1451)

2.3 *Valid IV selection methods*

A small statistical literature tries to select valid instruments from a large set of potentially invalid instruments. The first paper to do so is Kang, Zhang, Cai, and Small (2016) who use the Lasso. The key identifying assumption here is that the fraction of valid IVs needs to exceed one half. WFDS use the adaptive Lasso and also rely on this majority assumption, but they also allow for correlated instruments. Guo, Kang, Cai, and Small (2018) relax the majority assumption, basing their method on Assumption 16. They use the HT method which is based on a pairwise testing procedure. WLHB and Apfel and Liang (2021) also rely on this plurality assumption. I focus on the last two papers, because these methods produce a selection path and select a model with a test for overidentifying restrictions. The two methods have been presented in more detail in the preceding chapters.

All of these methods have oracle properties as defined in Definition 2, that is for $n \to \infty$ they work as well as if one knew the true set of valid instruments. This means that the methods select the invalid instruments as invalid and the post-selection estimator converges in distribution to the estimator that uses all invalid IVs as invalid and the valid ones as valid. The idea of these methods is to select those IVs as valid which produce the largest group of similar estimates and to select the rest as invalid. In the following, I present how two of the selection methods work.

3 COMPARISON OF THE METHODS

In the following, I compare the two ways in which the methods proceed.

3.1 *Analogies and differences*

To do so, I start by pointing out an analogy and a difference.

3.1.1 *Equivalence of just-identified estimates*

First, consider a direct point of connection between the two approaches. Both approaches are based on the just-identified estimates which use one IV at a time and control for the remaining IVs. Both MP and WLHB independently show that there is an alternative representation for these, which is $\hat{\beta}_j = \frac{\hat{\Gamma}_j}{\hat{\gamma}_j}$. MP show this in their Lemma 1, while WLHB show this in their Proposition A1. For the just-identified estimates it holds that

Property 5. *Property of just-identified estimates*

$$plim(\hat{\beta}_j) = \beta + \frac{\alpha_j}{\gamma_j}$$

3.1.2 *Different notions of falsification*

Both methods look at models which are rejected by the data. However, rejection is defined differently in the two approaches. In MP, the baseline model without relaxations ($\boldsymbol{\delta} = 0$) is

falsified when $\frac{\hat{\Gamma}_j}{\hat{\gamma}_j} \neq \frac{\hat{\Gamma}_l}{\hat{\gamma}_l}$ for any j, l. Generally, when $|\frac{\delta_{min}}{\gamma_{\min}}| + |\frac{\delta_{max}}{\gamma_{max}}| < max(\frac{\hat{\Gamma}_j}{\hat{\gamma}_j}) - min(\frac{\hat{\Gamma}_j}{\hat{\gamma}_j})$, where the variables indexed by *min* and *max* refer to the minimal and maximal estimates $\frac{\hat{\Gamma}_j}{\hat{\gamma}_j}$, the identified set is empty and the model is refuted. This follows from MP's Corollary 1, because then the intersection is empty.

In WLHB, the model is refuted when the Sargan test rejects the Null hypothesis that $\frac{\hat{\Gamma}_j}{\hat{\gamma}_j} = \frac{\hat{\Gamma}_l}{\hat{\gamma}_l} \quad \forall j, l$. This is why a model that is not falsified in WLHB might still be falsified in MP.

3.2 *Restate selection methods in terms of MP*

Next, I restate the selection methods in terms of the framework of MP. To do so, I first restate the plurality assumption with the help of the partial exclusion restriction in MP and discuss how the two methods proceed. With the help of this, I establish a connection between the outputs of both methods. Then, I restate the CIM with the help of quantities introduced in MP.

3.2.1 *Assumptions*

MP consider the $\boldsymbol{\delta}$ so that the model is marginally non-rejected, given the data. In IV selection, additionally it is assumed that a large enough group of IVs is valid in the sense that it fulfills the exclusion restriction.
In the language of MP, the existence of groups translates as:

Assumption 25. *Sparsity*

$$\exists \mathcal{F}_{nf}(\boldsymbol{\delta}') \ s.t. \ ||\boldsymbol{\delta}'||_0 < J - 1$$

In other words, this means that IV selection methods assume that there exist a non-falsified model whose assumption vector $\boldsymbol{\delta}$ is sparse in the sense that at least for two j it holds that $\delta_j = 0$. This just means that at least one group of at least two IVs exists.

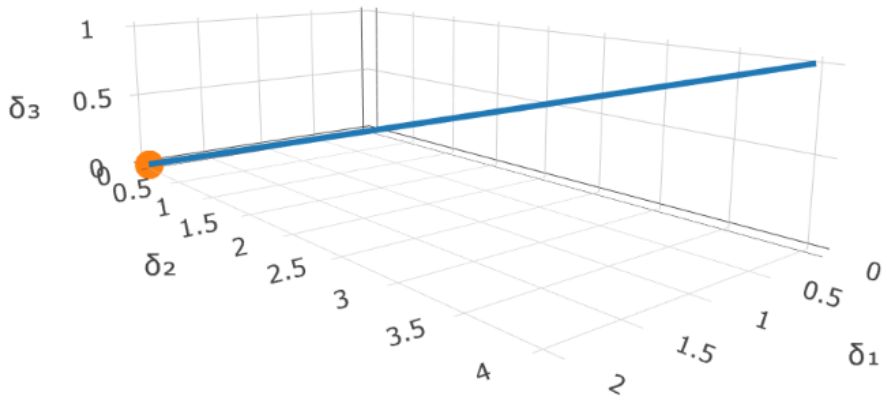
Plurality means that the largest group of IVs is valid, or equivalently that in presence of multiple sparse assumption vectors on the FF, the identified set of the assumption vector with the most valid IVs is unique and is at the true value:

Assumption 26. *Plurality*
For $\boldsymbol{\delta}^* = argmin||\boldsymbol{\delta}'||_0$, $\boldsymbol{\delta}^*$ is unique and $\mathcal{B}(\boldsymbol{\delta}^*) = \beta$.

3.2.2 *Outcome of selection methods*

The outcome of selection methods is the model that maximizes the number of zero-violations (or minimizes the number of non-zero violations) among the minimally non-falsified models. In other words, they find the model with the most IVs which exactly fulfill the exclusion restriction among the models that are just non-refuted.

Figure 12: Visualization of example with $J = 3$



Note: This graph visualizes the falsification frontier of the example from the text, with three IVs.

The minimally non-falsified relaxations of the baseline model on the falsification frontier are $\boldsymbol{\delta}(b) \in \mathcal{D}_{nf}$. The selection methods select the model with the minimal number of δ_j with a non-zero value:

$$\boldsymbol{\delta}_{Sel} = argmin||\boldsymbol{\delta}||_0.$$

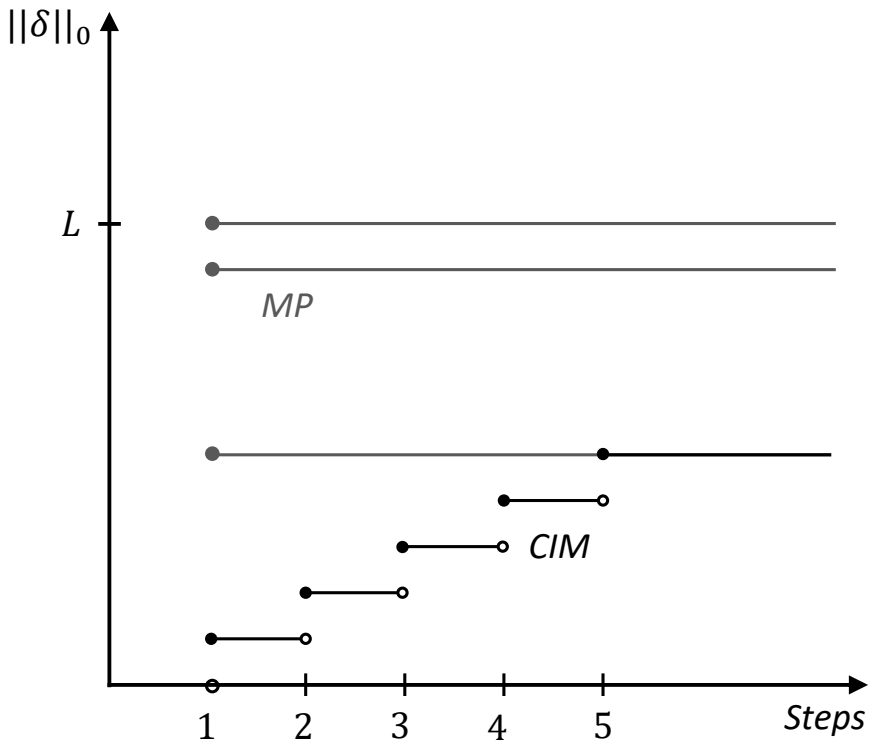
If there is only one marginally non-falsified model with a minimal number of relaxations set to non-zero, i.e. if the minimum is unique, the identified set $\mathcal{B}(\boldsymbol{\delta}_{Sel})$ is a singleton. If the plurality assumption holds, this identified set coincides with the true value of β.

Note that the outcome of the selection methods will be a corner/end point of the falsification frontier. To better understand this, take a case with $J = 3$ with $\frac{\Gamma_1}{\gamma_1} \neq \frac{\Gamma_2}{\gamma_2} = \frac{\Gamma_3}{\gamma_3}$. Assume $\Gamma_1 = 3, \Gamma_2 = 2, \Gamma_3 = 0.5$ and $\gamma_1 = 1, \gamma_2 = 2, \gamma_3 = 0.5$. There is a non-falsified model with $\delta_1 = 2$, $\delta_2 = 0$ and $\delta_3 = 0$. In this case, the falsification frontier is a line. The corner of the FF with $\delta_1 \neq 0$, $\delta_2 = 0$ and $\delta_3 = 0$ is the outcome of the selection methods. Another set that includes a binary relaxation is the one where $\delta_1 = 0$ and $\delta_2, \delta_3 \neq 0$ but this is not the model with the minimal number of non-zero relaxations. Figure 12 visualizes this falsification frontier. The blue line represents the falsification frontier and the orange dot at the lower left end of the frontier is the model that the selection methods search for and find in large samples.

3.2.3 *Continuous versus binary relaxations*

Given this understanding of plurality and of the outcome of the methods, another way to compare the two approaches is to notice that the relaxations of the baseline model are different. MP propose to look at the minimal deviations from the baseline model so that the model is marginally not falsified, i.e. such that a smaller relaxation would lead to a refutation of the model. The δ_j on the falsification frontier are non-zero potentially for

Figure 13: Visualization: Relaxations in MP and CIM



Note: This graph visualizes the different types of relaxations of the exclusion restriction in MP and WLHB for the example mentioned in the text. The horizontal axis denotes the steps and the vertical axis denotes the number of non-zero relaxations. MP start with the baseline model at step 0, where all IVs fulfill the exclusion restriction with $\delta_j = 0$. At step 1, a minimum of five instruments and potentially all J instruments violate the exclusion restriction, hence the gray lines denote the potential number of invalid instruments according to MP. The CIM selects an increasing number of IVs as violating the exclusion restriction and stops at the model with the minimal number of invalid IVs on the falsification frontier.

all j. In one step, the FAS gives the identified set for all minimally non-falsified models including all *continuous* relaxations which are marginally non-falsified.

In contrast, IV selection methods propose to find a model where for a group of IVs it holds that $\delta_j = 0$, while for the remaining IVs $\delta_j \neq 0$. Therefore, one can think of these relaxations as *binary* relaxations: an IV either violates the exclusion restriction or it does not. These methods try to find a set of assumptions for which the model is not rejected and the identified set is still a singleton. Some of the methods relax the exclusion restriction for one IV at a time and give a path of models which are sequentially falsified.

Figure 13 visualizes the two different procedures. Consider an example where $J = 10$. Five of the IVs have $\frac{\Gamma_j}{\gamma_j} = 1$ and the remaining five IVs have $\frac{\Gamma_j}{\gamma_j} \neq 1$, which are all unequal to each other. The graph shows that the method of MP allows for 5, 9 or 10 IVs to be invalid, i.e. potentially all IVs.

3.2.4 *Overlap analogy*

To further stress the similarity between the two methods, the CIM can be restated in terms of some quantities mentioned in MP's working paper.

MP introduce directional falsification points. When the researcher can specify the direction $\mathbf{d}$ of the relaxation of the exclusion restriction, the directional falsification point is the smallest m, so that $\mathcal{B}(\boldsymbol{\delta})$ is not empty for $\boldsymbol{\delta} = m \cdot \mathbf{d}$. The directional falsification point then is

$$m^* = \max_{j,j' \in \{1,...,J\}} \frac{\frac{\Gamma_j}{\gamma_j} - \frac{\Gamma_{j'}}{\gamma_{j'}}}{\frac{d_j}{|\gamma_j|} - \frac{d_{j'}}{|\gamma_{j'}|}} \tag{79}$$

In the CIM, the breaking points are defined as

$$\hat{\psi}^*_{j,j'} = \frac{|\frac{\hat{\Gamma}_j}{\hat{\gamma}_j} - \frac{\hat{\Gamma}_{j'}}{\hat{\gamma}_{j'}}|}{\hat{\sigma}_j - \hat{\sigma}_{j'}}, \tag{80}$$

where $\hat{\sigma}_j = \sqrt{\hat{Var}(\hat{\beta}_j)}$. If the direction of falsification is specified as $\mathbf{d} = (\hat{\sigma}_1|\hat{\gamma}_1|, ..., \hat{\sigma}_j|\hat{\gamma}_j|)$, then asyptotically the directional falsification points in MP are equivalent to the breaking points in WLHB, $m^* = \max_{j,j' \in \{1,...,J\}} \hat{\psi}^*_{j,j'}$. If the tuning parameter ψ is larger than this value in the CIM, then all CIs overlap and all IVs are selected as valid. Hence, at this first breaking point, one IV will be selected as invalid. Relaxing all restrictions by $m^* \cdot \mathbf{d}$, the identified set is not empty anymore.

According to Theorem 1 in MP, the intersection between the intervals $\left[\frac{\Gamma_j}{\gamma_j} - \frac{\delta_j}{\gamma_j};\ \frac{\Gamma_j}{\gamma_j} + \frac{\delta_j}{\gamma_j}\right]$ is the identified set. With the set of minimal deviations $\boldsymbol{\delta}$, the identified set is then given. The analysis stops here.

WLHB also look at intersections of intervals. The difference is that they look at intersections of CIs $\left[\frac{\hat{\Gamma}_j}{\hat{\gamma}_j} - \hat{\sigma}_j\psi;\ \frac{\hat{\Gamma}_j}{\hat{\gamma}_j} + \hat{\sigma}_j\psi\right]$. If for a given ψ the CIs overlap, this is taken as evidence that the $\frac{\Gamma_j}{\gamma_j}$ are indeed the same and the instruments are valid. The method then goes on to test this hypothesis via the Sargan test. If the Sargan test rejects the model, the next breaking point ψ where a lower number of IVs fulfills the exclusion restriction is evaluated. This goes on until the model is not falsified by the Sargan test anymore. Hence, the CIM path can be seen as a sequence of directional falsification points where the direction of the violation is determined by the variance estimate and the first-stage estimate and falsification is assessed by the Sargan test.

3.3 *Advantages and disadvantages of the two approaches*

I continue by stating the respective limitations and advantages of each of the two approaches. I start with the MP approach. The main advantage of MP is to give a range of smallest possible relaxations so that the model is not refuted. The identified set for this set of relaxations (the FF) is the FAS. No prior knowledge about the sparsity of the assumption vector $\boldsymbol{\delta}$ is needed.

This approach also has a few disadvantages: First, the falsification frontier is difficult to visualize when $J > 3$. One could visualize the FF for two instruments at a time, but that can also mean a prohibitively large number of visual diagnoses when the number of IVs is large. Second, in finite sample settings, it is unclear whether the variation in $\frac{\hat{\Gamma}_j}{\hat{\gamma}_j}$ comes from violations of the exclusion restriction ($\alpha_j \neq 0$) or from the variance of the 2SLS estimator. Third, the method lacks a clear-cut indication of how wide an FAS is too wide.

The main disadvantage of MP is that no specific indication is given as to which specific model is more likely and no additional information on the location of the coefficient of interest, β, is obtained. Moreover, the title suggests that a model is salvaged, while in fact the baseline model with $\boldsymbol{\delta} = 0$ is replaced with a range of new ones which might make more sense in light of the data.

The main advantage of WLHB is that the outcome is a specific model where β is point-identified and which makes it possible to consistently estimate β. This method hence allows to retrieve β in a setting where the exclusion restriction is violated. Arguably, this approach has the same model in place all along and truly salvages it, by finding the correct vector of assumptions.

The main disadvantage of the method is that the plurality assumption (Ass. 26) has to be fulfilled and it is still strict. If that assumption does not hold, the method might output some misleading point on the FF and FAS.

3.4 *Synthesis*

The approach of MP is less restrictive, but adds little knowledge about β, while IV selection methods need assumptions which are more restrictive but give a corrected estimate of β as an output. Therefore, both approaches can be helpful in different situations.

MP allow for continuous relaxations, are agnostic about the origin of violations and allow all IVs to be invalid at the same time. MP's analysis is more helpful when there is no IV for which the exclusion restriction holds ($\alpha_j \neq 0$, for all IVs). In this case, the FF and FAS provide diagnostics about how far away the model is from non-falsification. However, it is not possible to learn anything new about β, the final object of interest of the analysis. A very wide FAS would leave the researcher with little insight.

Instrument selection methods like CIM and AHC are helpful when the exclusion restriction ($\alpha_j = 0$) holds for some but not all IVs. IV selection procedures retrieve the set of valid IVs and allow to learn the parameter β. This comes at the cost of assuming that a plurality of IVs is valid. The issue is that the exclusion restriction restriction might not hold in practice and therefore the IV selection methods are not applicable.

After comparing the two methods and their respective advantages, it seems sensible to use both methods in a complementary way: researchers can report the FAS *and* the result of the selection methods. The FAS gives the identified set for the minimal relaxations that are not falsified and a selection method gives the identified value for a *specific point* on the falsification frontier and hence a specific point on the FAS. If the selection method continues until the end without stopping, this is an indication that the Sparsity Assumption

is not fulfilled. Then, one should resort to the entire FAS. If the selection method stops, however, it might offer some helpful insights.

3.5 *Conceptual differences*

Finally, I compare the two methods on a higher level, from the perspective of the philosophy of science. MP relate their approach to *falsificationism*, a viewpoint introduced by Popper (1959). Thereafter, a model can be falsified and hence rejected. A non-rejection means a corroboration of the hypothesis but never a verification, because there always could be a further observation which falsifies the model.

Dogmatic falsificationism relies on the truth of a (single) observational statement, able to falsify a model. However, this statement also relies on some theoretical background assumptions, which can themselves not be verified. Therefore, dogmatic falsificationism leads to a dead-end road in which no scientific discovery is possible any longer. More sophisticated forms of falsificationism judge a model based on the extent of observations that it is able to explain and whether there is a model that can explain more observations and not on whether there is a single contradictory observation.

Lakatos (1976) devises a methodology of scientific research according to which scientists proceed along series of theories that are sequentially formulated, tested, rejected and renewed. Assumptions are split into a hard core of untested assumptions and a belt of auxiliary assumptions.

The methodological approaches that try to deal with falsified IV models can be understood before the background of these theories of science. The current approach taken by researchers who use IVs is that of dogmatic falsificationism: once the model with all instruments assumed to be valid is rejected, the model is falsified and this stands in the way of the researcher as an apparently insurmountable obstacle.

MP propose to reconcile the observed data with the model by allowing for violations of the exclusion restriction by potentially all models. The hard assumptions are stated in Assumptions 21 to 23. The auxiliary assumptions are on the size of direct effects in 24.

The IV selection methods propose a testing path: one starts with a very restrictive model and relaxes the exclusion restriction for one IV at a time. If the first model is rejected, one continues with the next one and enters an iteration of proposing a model and rejection until a model is not rejected anymore, corroborating the proposed model. This gives a longer path of models resembling the path of scientific discovery proposed by Lakatos (1976). The hard assumption here is plurality. Auxiliary assumptions are the assumptions on whether a specific instrument is valid.

4 ISSUES RAISED IN IV SELECTION LITERATURE

In this subsection I discuss issues that surfaced in the IV selection literature but were not discussed in MP. This discussion can also be of interest for readers of MP.

4.1 *Weak instruments*

Both approaches propose to screen for weak instruments beforehand. The IVs are tested for relevance individually. If the j-th IV surpasses a given cutoff, as proposed in Stock and Yogo (2002), the IV is estimated to belong to the set of relevant IVs $\mathcal{L}$:

$$\hat{\mathcal{L}} = \{j \in \{1, ..., J\} : F_j \geq C_n\},$$

where F_j is the first-stage F-statistic for a given IV, while controlling for the remaining IVs, $\mathbf{Z}_{-j}$, and C_n is a cut-off that grows as n goes to infinity. This procedure is used in MP, Guo, Kang, Cai, and Small (2018) and WLHB and is a further common feature.

However, after having pre-screened strong IVs, it is not clear what should best be done with the weak IVs. If they are invalid, one should control for them, and if they are valid using them might still add some information. MP choose to include the weak IVs as controls (see MP, Table I, Panel B). Still, it is not at all clear whether this is the best practice.

An alternative would be to leave out the pre-selection and proceed directly to the selection of valid instruments. However, in the CIM it could be that weak and invalid instruments are always selected, because their CIs are very wide and always overlap with the valid ones - ending up with a worst case scenario. Indeed Apfel and Liang (2021) show in their simulations that with weak valid instruments, the weak valid ones are selected out by HT and with weak invalid ones, the weak invalid ones are selected as valid by CIM.

In contrast, selecting with the AHC leads to good selection performance in both settings. AHC does not automatically throw out all weak IVs, but still includes them as valid, when the just-identified estimate using a weak IV is still included in the largest cluster and passes the Sargan test. This is taken as an indication that the weakness of the first-stage does not lead to a too large bias. Hence, in settings with potentially weak instruments, AHC might be the best option.

4.2 *Correlated instruments*

A problem that has been pointed out by WLHB is that not all types of violations of exogeneity are permitted. Appendix G of MP states that general violations of the exclusion restriction are also allowed, but this statement is not qualified more closely.

Assume there are two types of instruments $\mathbf{z}_1$ and $\mathbf{z}_3$: $\mathbf{z}_1$ is a valid instrument and $\mathbf{z}_3$ is an invalid instrument. Moreover, $\mathbf{z}_3$ is a collider, which means that it is affected by $\mathbf{z}_1$ and some unobservable factor $\mathbf{u}$ (which is also correlated with the outcome $\mathbf{y}$). Not controlling for the collider, there is no correlation between $\mathbf{z}_1$ and $\mathbf{u}$. However, once a collider is controlled for, $\mathbf{z}_1$ is invalidated, because a correlation with the unobservable factor was induced. For an example of this, see also Cole et al. (2010). MP do not discuss how this affects the width and location of the FAS.

5 VIOLATIONS OF EXOGENEITY

MP claim that their setting allows for arbitrary violations of exclusion and exogeneity and that in case of a violation of exogeneity, only the interpretation of the coefficient of the direct effect would change. In this section I show a setting with an unobserved confounder and how it affects the results.

5.1 *Model and assumptions*

Let the model

$$\mathbf{y} = \mathbf{d}\beta + \mathbf{u} \tag{81}$$

with three instruments $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$. The first-stage equation is

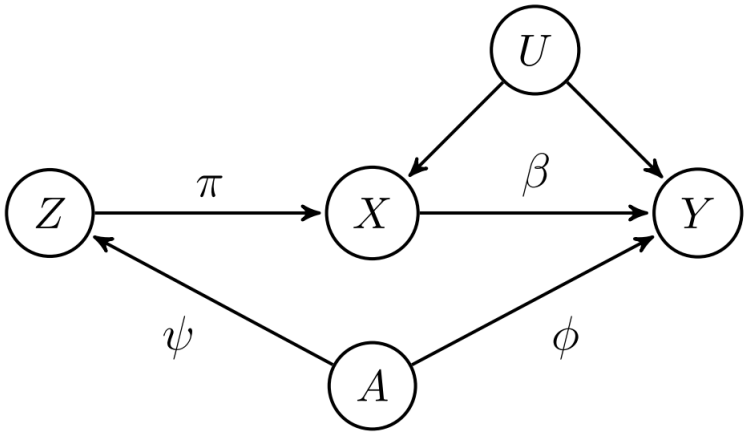
$$\mathbf{d} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

There is one invalid IV $\mathbf{z}_3$ which is confounded by $\mathbf{u}$:

$$\mathbf{z}_3 = \kappa_3 \mathbf{u} + \boldsymbol{\epsilon}_3 \tag{82}$$

with $\frac{1}{n}\mathbf{z}_3'\mathbf{u} \xrightarrow{P} \sigma_u^2 \kappa_3 = \eta_3$ and $\frac{1}{n}\mathbf{z}_3'\boldsymbol{\epsilon}_3 \xrightarrow{P} 0$. Here, the third instrument is determined by the error term $\mathbf{u}$. In fact, MP illustrate this case in the DAG in Figure 9 of their working paper. The DAG is shown in Figure 14. Note that in my model A is $\mathbf{u}$ and ψ is κ, to avoid confusion with the tuning parameter from the CIM. I assume partial exogeneity

Figure 14: Directed acyclic graph from MP



Note: Directed Acyclic Graph from Masten and Poirier (2021, p.82), Figure 9 (b).

Assumption 27. *Partial exogeneity*
There are known constants $\delta_j \geq 0$ *such that* $|\eta_j| \leq \delta_j \;\; \forall j \in \{1, ..., J\}$.

When using each IV at a time and controlling for the rest, the result is the FAS proposed by MP. Consider the reduced form estimator

$$\begin{aligned}\hat{\mathbf{\Gamma}} &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} \\ &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{Z}\boldsymbol{\gamma}\beta + \boldsymbol{\varepsilon}\beta + \mathbf{u}) \\ &= \boldsymbol{\gamma}\beta + \left(\frac{1}{n}\mathbf{Z}'\mathbf{Z}\right)^{-1}\frac{1}{n}\mathbf{Z}'(\boldsymbol{\varepsilon}\beta + \mathbf{u}) \\ &\xrightarrow{P} \boldsymbol{\gamma}\beta + \mathbf{Q}_{ZZ}^{-1}\begin{pmatrix} 0 \\ 0 \\ \eta_3 \end{pmatrix} \\ &= \boldsymbol{\gamma}\beta + \eta_3\begin{pmatrix} \sigma^{13} \\ \sigma^{23} \\ \sigma^{33} \end{pmatrix}\end{aligned}$$

with $\mathbf{Q}_{ZZ}^{-1}$ having entries $\sigma_{jj'}$, with $j, j' \in \{1, ..., J\}$. The just-identified estimators then converge to:

$$\frac{\hat{\Gamma}_j}{\hat{\gamma}_j} \xrightarrow{P} \beta + \frac{\sigma_{j3}\eta_3}{\gamma_j}$$

where σ_{j3} is the j-3 entry of $\mathbf{Q}_{ZZ}^{-1}$. Note that the violation of the exclusion restriction for the third instrument affects the estimators for each single IV, irrespective of whether it violates the exclusion restriction. In this case,

$$\boldsymbol{\sigma}_{\cdot 3}\eta_3 = \mathbf{\Gamma} - \boldsymbol{\gamma}\beta$$

where $\boldsymbol{\sigma}_{\cdot 3}$ is the third column of $\mathbf{Q}_{ZZ}^{-1}$. The identified set therefore is

$$\mathcal{B}(\boldsymbol{\delta}) = \{b : -\sigma_{j3}\delta_j \leq \Gamma_j - \gamma_j b \leq \sigma_{j3}\delta_j\}$$

and analogously to Corollary 1 in MP it is the intersection of $\mathcal{B}_j(\delta_j)$, where

$$\mathcal{B}_j(\delta_j) = \left[\frac{\Gamma_j}{\gamma_j} - \frac{\sigma_{j3}\delta_j}{|\gamma_j|}, \frac{\Gamma_j}{\gamma_j} + \frac{\sigma_{j3}\delta_j}{|\gamma_j|}\right]$$

Therefore, the identified set is not as specified in Theorem 1 and Corollary 1, anymore. To see which implications this has for the analysis, I follow MP's paper in developing analogous results but with violations of exogeneity.

5.2 *Results with violations of exogeneity*

When leaving out the controls for the remaining IVs in each just-identified estimator, there is a set of J just-identified estimators

$$\begin{aligned}\tilde{\beta}_j &= (\mathbf{z}_j'\mathbf{d})^{-1}\mathbf{z}_j'\mathbf{y} \\ &= (\mathbf{z}_j'\mathbf{d})^{-1}\mathbf{z}_j'(\mathbf{d}\beta+\mathbf{u}) \\ &= (\mathbf{z}_j'\mathbf{d})^{-1}\mathbf{z}_j'(\mathbf{d}\beta+\mathbf{u}) \\ &= \beta+\left(\frac{1}{n}\mathbf{z}_j'\mathbf{d}\right)^{-1}\left(\frac{\mathbf{z}_j'\mathbf{u}}{n}\right).\end{aligned} \tag{83}$$

Let $\gamma_j^* := \text{plim}\left(\frac{1}{n}\mathbf{z}_j'\mathbf{d}\right) = \text{plim}\left(\frac{1}{n}\mathbf{z}_j'\mathbf{Z}\boldsymbol{\gamma}\right) = \mathbf{q}_{j\cdot}\boldsymbol{\gamma}$, where $\mathbf{q}_{j\cdot}$ is the j-th line of $\mathbf{Q}_{ZZ}$ then

$$\begin{aligned}\tilde{\beta}_1 &\xrightarrow{P} \beta \\ \tilde{\beta}_2 &\xrightarrow{P} \beta \\ \tilde{\beta}_3 &\xrightarrow{P} \beta+\frac{\eta_3}{\gamma_3^*}\end{aligned} \tag{84}$$

$$\tilde{\beta}_j = \frac{\mathbf{z}_j'\mathbf{y}}{\mathbf{z}_j'\mathbf{d}} \xrightarrow{P} \beta+\frac{\eta_3}{\gamma_j^*}$$

The identified set as in Theorem 1 becomes

$$\mathcal{B}(\boldsymbol{\delta}) = \{b : -\delta_j \leq (Cov(\mathbf{z}_j,\mathbf{y}) - Cov(\mathbf{z}_j,\mathbf{d})b) \leq \delta_j\}$$

Analogously to Corollary 1 the identified set is the intersection of

$$\mathcal{B}_j(\delta_j) = \left[\frac{Cov(\mathbf{z}_j,\mathbf{y})}{Cov(\mathbf{z}_j,\mathbf{d})} - \frac{\delta_j}{\mathbf{q}_{j\cdot}\boldsymbol{\gamma}}, \frac{Cov(\mathbf{z}_j,\mathbf{y})}{Cov(\mathbf{z}_j,\mathbf{d})} + \frac{\delta_j}{\mathbf{q}_{j\cdot}\boldsymbol{\gamma}}\right]$$

Results, propositions, lemmas and theorems analogous to those in MP can be easily shown by replacing Γ_j and γ_j in MP with $\tilde{\Gamma}_j = Cov(\mathbf{z}_j,\mathbf{y})$ and $\tilde{\gamma}_j = Cov(\mathbf{z}_j,\mathbf{d})$. For the sake of completeness, I reproduce the proofs in the Appendix.

The following Lemma says that if the $\boldsymbol{\delta}$ are as specified, for each $\boldsymbol{\delta}$, the identified set is a singleton.

Lemma 2. *MP's Lemma 2*
Suppose Assumptions 21 - 23 hold. Suppose the joint distribution of $(\mathbf{y},\mathbf{d},\mathbf{Z})$ *is known. Suppose* $K = 1$. *Let* $b \in \left[min\tilde{\beta}_j, max\tilde{\beta}_j\right]$.
Define $\delta(b) = (|Cov(\mathbf{z}_1,\mathbf{y}) - Cov(\mathbf{z}_1,\mathbf{d})b|, ..., |Cov(\mathbf{z}_J,\mathbf{y}) - Cov(\mathbf{z}_J,\mathbf{d})b|)$. *Then* $\mathcal{B}(\boldsymbol{\delta}(b)) = \{b\}$.

This Lemma is used to prove that the falsification frontier is the one proposed in the following Proposition.

Proposition 11. *MP's Proposition 2*
Suppose Assumptions 21 - 23 hold. Suppose the joint distribution of $(\mathbf{y}, \mathbf{d}, \mathbf{Z})$ *is known. Suppose* $K = 1$*. Then the FF is the set*

$$FF = \left\{ \boldsymbol{\delta} \in \mathbb{R}^J_{\geq 0} : \ \delta_j = |Cov(\mathbf{z}_j, \mathbf{y}) - Cov(\mathbf{z}_j, \mathbf{d})b|, \ j = 1, ..., J, \ b \in \left[min\tilde{\beta}_j, max\tilde{\beta}_j \right] \right\}$$

The next Theorem introduces the falsification adaptive set.

Theorem 4. *MP's Theorem 2*
Suppose Assumptions 21 - 23 hold. Suppose the joint distribution of $(\mathbf{y}, \mathbf{d}, \mathbf{Z})$ *is known. Suppose* $K = 1$*. Then*

$$\bigcup_{\boldsymbol{\delta} \in FF} \mathcal{B}(\boldsymbol{\delta}) = \left[min \frac{Cov(\mathbf{z}_j, \mathbf{y})}{Cov(\mathbf{z}_j, \mathbf{d})}, max \frac{Cov(\mathbf{z}_j, \mathbf{y})}{Cov(\mathbf{z}_j, \mathbf{d})} \right]$$

is the falsification adaptive set.

The proof of the latter is the same as the one in MP and is not reproduced separately. This shows that the partial exclusion restriction and the partial exogeneity assumption result in two different falsification frontiers and two different falsification adaptive sets: one which uses just-identified estimates, controlling for the remaining IVs (MP) and the other not controlling for the remaining IVs (the new FAS).

5.3 *Comparison of FF*

To better understand the outcomes of the analyses under the different types of violations, I first compare the resulting falsification frontiers. The FF from MP is denoted FF_{MP} and the one from this document is denoted by FF_{Exog}.

$$\begin{aligned} FF_{Exog} : \ & \boldsymbol{\delta}' = |Cov(\mathbf{Z}, \mathbf{y}) - Cov(\mathbf{Z}, \mathbf{d})b|, \quad b \in \left[min \frac{Cov(\mathbf{z}_j, \mathbf{y})}{Cov(\mathbf{z}_j, \mathbf{d})}, max \frac{Cov(\mathbf{z}_j, \mathbf{y})}{Cov(\mathbf{z}_j, \mathbf{d})} \right] \\ & \text{and} \\ FF_{MP} : \ & \boldsymbol{\delta} = |\boldsymbol{\Gamma} - \boldsymbol{\gamma} b|, \quad b \in \left[min \frac{\Gamma_j}{\gamma_j}, max \frac{\Gamma_j}{\gamma_j} \right] \end{aligned} \tag{85}$$

If $\left[min \frac{Cov(\mathbf{z}_j, \mathbf{y})}{Cov(\mathbf{z}_j, \mathbf{d})}, max \frac{Cov(\mathbf{z}_j, \mathbf{y})}{Cov(\mathbf{z}_j, \mathbf{d})} \right]$ and $\left[min \frac{\Gamma_j}{\gamma_j}, max \frac{\Gamma_j}{\gamma_j} \right]$ overlap, i.e. if some range of b is in both intervals, the following holds

$$FF_{MP} : \boldsymbol{\delta} = |Var(\mathbf{Z})^{-1}[Cov(\mathbf{Z}, \mathbf{y}) - Cov(\mathbf{Z}, \mathbf{d})b]| = |Var(\mathbf{Z})^{-1}|\boldsymbol{\delta}'$$

For given b, it might be that $|Cov(\mathbf{z}_j, \mathbf{y}) - Cov(\mathbf{z}_j, \mathbf{d})b| < |\Gamma_j - \gamma_j b|$ i.e. $\delta'_j < \delta_j$. This means that the deviation in FF_{MP} is not *minimal*. But by definition the FF summarizes the set of minimal deviations so that the model is not falsified. Therefore, δ_j can not be on the FF. As an example, take the case with violations of exogeneity and uncorrelated IVs. Then, $\delta_j = |Var(\mathbf{z}_j)^{-1}|\delta_j{}'$. If for some j, $Var(\mathbf{z}_j) < 1$, then $Var(\mathbf{z}_j)^{-1} > 1$ and $\delta_j > \delta'_j$.

5.4 *FAS are the same*

Next, I compare the two resulting falsification adaptive sets. I show a special case when the FAS are identical. This is the case, when the IVs are uncorrelated, such that $Cov(\mathbf{z}_j, \mathbf{z}_{j'}) = 0$ for $j \neq j'$ and the inconsistency comes only from violations of the exclusion restriction. The estimand, including the other IVs as controls is

$$\beta_j = \beta + \frac{\alpha_j}{\gamma_j}$$

In the example in Section 5.1:

$$\beta_1 = \beta + \frac{\alpha_1 + \sigma_{j3}\eta_3}{\gamma_1}.$$

Assuming that the violations come only from violations of the exclusion restriction means that $\eta_3 = 0$, and the estimands are $\beta + \frac{\alpha_j}{\gamma_j}$. The estimands, excluding the remaining IVs are

$$\begin{aligned} \tilde{\beta}_j &= (\mathbf{z}_j{}'\mathbf{d})^{-1}\mathbf{z}_j{}'\mathbf{y} = (\mathbf{z}_j{}'\mathbf{d})^{-1}\mathbf{z}_j{}'(\mathbf{d}\beta + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{u}) \\ &= \beta + (\mathbf{z}_j{}'\mathbf{Z}\boldsymbol{\gamma})^{-1}\mathbf{z}_j{}'\mathbf{Z}\boldsymbol{\alpha} \\ &= \beta + (\mathbf{e}_j\boldsymbol{\gamma})^{-1}\mathbf{e}_j\boldsymbol{\alpha} \\ &= \beta + \frac{\alpha_j}{\gamma_j} \end{aligned} \tag{86}$$

where $\mathbf{e}_j$ is the j-th standard basis vector. The last two equalities follow, as $\mathbf{z}_j{}'\mathbf{Z} = Var(\mathbf{z}_j)\mathbf{e}_j$, because $Cov(\mathbf{z}_j, \mathbf{z}_{j'}) = 0$ for $j \neq j'$. The $Var(\mathbf{z}_j)$ cancel out in equation 86. In this case, the spans of just-identified estimates are the same and the two FAS are identical.

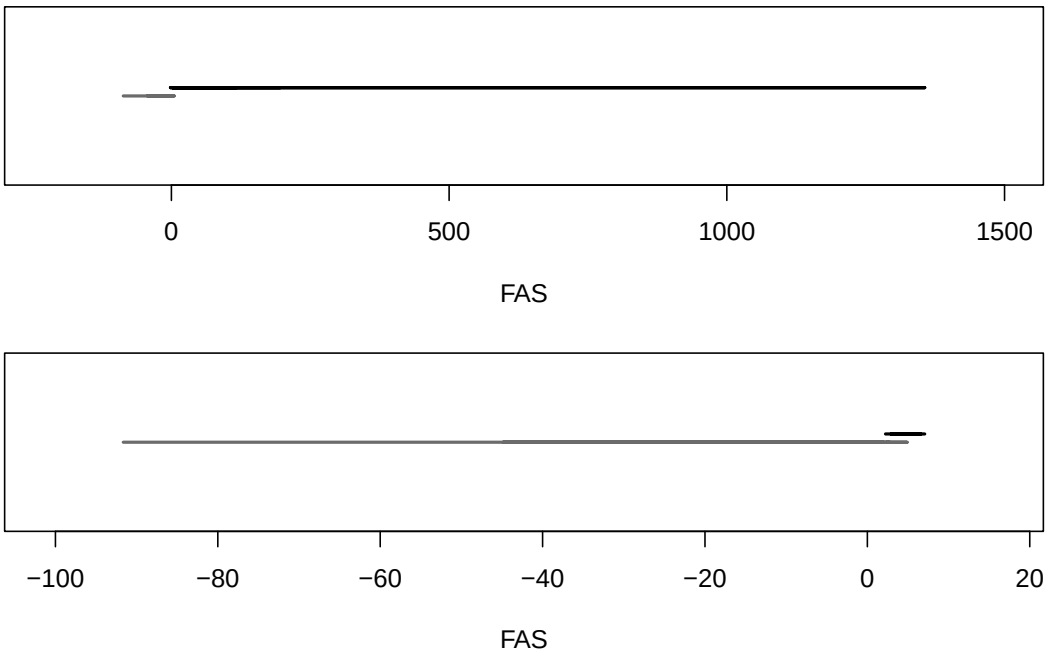
From the preceding subsection it follows that if $Var(\mathbf{Z})^{-1} = \mathbf{I}$, the two falsification frontiers are the same. Hence, when the violations come from violations of the exclusion restriction, the IVs are uncorrelated and have variance one, the two FF and FAS are identical.

5.5 *Two examples*

To illustrate the difference between FAS_{MP} and FAS_{Exog}, in the following two examples are shown. One could think that one of the FAS is a subset of the other. This examples shows that this need not be the case.

In the examples there are ten IVs, all of which are invalid. The true β is 0 and $\mathbf{z}_i \sim N(0, \mathbf{\Sigma}_z)$ with $\mathbf{\Sigma}_{z,jk} = 0.5^{|j-k|}$. The $\boldsymbol{\gamma}$ is drawn from a uniform distribution, $Unif(0, 2)$. The direct effects $\boldsymbol{\alpha} = (0, 1, 2, 3, ..., 9)$ and the $\boldsymbol{\varepsilon} = N(0, 1)$ and $\mathbf{u} = 0.5 \cdot \boldsymbol{\varepsilon} + N(0, 1)$. In my first example, $\kappa_j = 0.2$ for all IVs. The black line is the FAS_{Exog} and the grey line is FAS_{MP}. In the first example (upper graph in 15), FAS_{Exog} is much wider than FAS_{MP}. In the second example (lower graph), $\kappa_j = 1$ for all IVs. Now, FAS_{MP} is wider than the FAS_{Exog}. This shows that none of the FAS needs to be a subset of the other. The puzzling

Figure 15: Illustration of FAS



Note: Illustration of the FAS from MP (lower line, in gray) and the one in this chapter (upper line, in black). First graph, simulated FAS with $\kappa = 0.2$, lower graph with $\kappa = 1$.

result that a relaxation of the exogeneity assumption leads to different FF and FAS calls for more thorough research. The question that should be investigated in future is: Which FF and FAS are the correct ones? Might it be a union of the two sets of just-identified estimates or a union of even more sets of estimates? One way to think about this difference is that the different FF and FAS are defined for relaxations of slightly different assumptions and the choice of FAS has to do with which partial relaxations one considers.

6 CONCLUSION

In this chapter I have connected valid IV selection methods with the analysis of MP. The latter paper is interested in the minimal deviations from the baseline assumption that do not lead to a falsification of the model. It introduces the concept of the falsification frontier and the falsification adaptive set for IV models. Instrument selection methods estimate the set of valid IVs and then control for them in a post-selection 2SLS.

I have shown analogies and differences of the methods. Importantly, MP's analysis relaxes potentially all IVs contemporaneously, while IV selection methods relax the exclusion restriction one IV at a time. I then go on to restate the selection methods in terms of the framework introduced by MP and discuss how the selection methods find a specific point on the falsification frontier. This point maximizes the number of zero-violations, i.e. it finds the most sparse relaxation vector on the falsification frontier. Moreover, one specific selection method, the CIM can be restated in terms of the directional falsification points, another concept introduced by MP to the IV context. Advantages and disadvantages of the

two methods and a way to use them as complements are discussed. Moreover, I describe issues from the selection literature that can also be of equal interest for readers of MP.

Finally, it is shown that allowing for violations of exogeneity instead of exclusion leads to different outcomes for the falsification frontier and the falsification adaptive set. These differences call for more thorough research on the nature of violation allowed by MP. Future work should look at how these two FAS relate and if there is a single FAS which can allow for all possible types of violations.

Overall, when tests of overidentifying restrictions reject there are different possible tools that could be used. This chapter discusses the connections of two approaches that treat the same problem and shows how they can work together. If one believes that a large group of IVs is valid, methods such as CIM or AHC can help obtain results which correct for the presence of invalid IVs. If one does not believes so, the FAS still gives an idea about the extent of falsification. This chapter has therefore shown how new types of analyses can work together and enter the econometrician's toolbox. In this way, even in situations where IV estimation does not seem well-suited these new tools can still help researchers gain some insight on the important questions that they are investigating.

Appendices

G PROOFS

Proof of Lemma 2

Proof:

$$
\begin{aligned}
\mathcal{B}(\boldsymbol{\delta}(b)) &= \bigcap_{\gamma_j \neq 0} \left[\tilde{\beta}_j - \frac{|Cov(\mathbf{z}_j, \mathbf{y}) - Cov(\mathbf{z}_j, \mathbf{d})b|}{|Cov(\mathbf{z}_j, \mathbf{d})|}, \tilde{\beta}_j + \frac{|Cov(\mathbf{z}_j, \mathbf{y}) - Cov(\mathbf{z}_j, \mathbf{d})b|}{|Cov(\mathbf{z}_j, \mathbf{d})|} \right] \\
&= \bigcap_{\tilde{\gamma}_j \neq 0} \left[\tilde{\beta}_j - |\frac{Cov(\mathbf{z}_j, \mathbf{y})}{Cov(\mathbf{z}_j, \mathbf{d})} - b|, \tilde{\beta}_j + |\frac{Cov(\mathbf{z}_j, \mathbf{y})}{Cov(\mathbf{z}_j, \mathbf{d})} - b| \right] \\
&= \bigcap_{\tilde{\gamma}_j \neq 0} \left[\tilde{\beta}_j - |\tilde{\beta}_j - b|, \tilde{\beta}_j + |\tilde{\beta}_j - b| \right] \\
&= \left(\bigcap_{\tilde{\beta}_j > b, \tilde{\gamma}_j \neq 0} \left[\tilde{\beta}_j - |\tilde{\beta}_j - b|, \tilde{\beta}_j + |\tilde{\beta}_j - b| \right] \right) \bigcap \left(\bigcap_{\tilde{\beta}_j < b, \tilde{\gamma}_j \neq 0} \left[\tilde{\beta}_j - |\tilde{\beta}_j - b|, \tilde{\beta}_j + |\tilde{\beta}_j - b| \right] \right) \\
&= \left(\bigcap_{\tilde{\beta}_j > b, \tilde{\gamma}_j \neq 0} \left[b, 2\tilde{\beta}_j - b \right] \right) \bigcap \left(\bigcap_{\tilde{\beta}_j < b, \tilde{\gamma}_j \neq 0} \left[2\tilde{\beta}_j - b, b \right] \right) = \{b\}
\end{aligned}
$$

The last equality follows, because the only point of intersection is b. □

Proof of Proposition 11

Proof:

The procedure is the same as in MP's Appendix:

1. If $\boldsymbol{\delta} \in \text{FF}^{guess}$, then $\mathcal{B}(\boldsymbol{\delta})$ is not empty, by Lemma 2.

2. Part two of the proof also holds. The idea is that when the identified set is a singleton, for a smaller $\boldsymbol{\delta}' < \boldsymbol{\delta}$, the identified set must be empty.
   The first two parts together prove that $\text{FF}^{guess} \subseteq \text{FF}$.

3. Part 3 of MP's proof goes on to show that $\text{FF}^{guess} \supseteq \text{FF}$. This is shown by showing that $\boldsymbol{\delta} \notin \text{FF}^{guess} \Rightarrow \boldsymbol{\delta} \notin \text{FF}$. The proof proceeds in two parts: the first assuming that the identified set is a subset of the range of just-identified estimands, the second assumes that the identified set might comprise elements outside of that range.

   - *(a)* If $\mathcal{B}(\boldsymbol{\delta}) \subseteq \left[min \frac{Cov(\mathbf{z}_j, \mathbf{y})}{Cov(\mathbf{z}_j, \mathbf{d})}, max \frac{Cov(\mathbf{z}_j, \mathbf{y})}{Cov(\mathbf{z}_j, \mathbf{d})} \right]$, this part of the proof assumes a $\boldsymbol{\delta}$ which is larger than that on the FF. By Lemma 2, there is a smaller $\boldsymbol{\delta}'$ which leads to a non-empty identified set. Hence, by definition of the FF, if $\boldsymbol{\delta}$ does not lie on FF^{guess}, $\boldsymbol{\delta}$ also does not lie on FF.

- *(b)* Assume $\boldsymbol{\delta} = \boldsymbol{\delta}(b)$ with $b \notin \left[min\frac{Cov(\mathbf{z}_j,\mathbf{y})}{Cov(\mathbf{z}_j,\mathbf{d})}, max\frac{Cov(\mathbf{z}_j,\mathbf{y})}{Cov(\mathbf{z}_j,\mathbf{d})}\right]$ and $\boldsymbol{\delta}' = \boldsymbol{\delta}(b_{max})$ where $b_{max} = max\frac{Cov(\mathbf{z}_j,\mathbf{y})}{Cov(\mathbf{z}_j,\mathbf{d})}$. Here, $\boldsymbol{\delta} \notin \text{FF}^{guess}$ and $\boldsymbol{\delta}' \in \text{FF}^{guess}$. Assume $b_{max} < b$, then

$$\begin{aligned}\delta'_j &= |Cov(\mathbf{z}_j,\mathbf{y}) - Cov(\mathbf{z}_j,\mathbf{d})b_{max}| \\ &= |Cov(\mathbf{z}_j,\mathbf{d})|\left(b_{max} - \frac{Cov(\mathbf{z}_j,\mathbf{y})}{Cov(\mathbf{z}_j,\mathbf{d})}\right) < |Cov(\mathbf{z}_j,\mathbf{d})|\left(b - \frac{Cov(\mathbf{z}_j,\mathbf{y})}{Cov(\mathbf{z}_j,\mathbf{d})}\right) \leq \delta_j\end{aligned}$$

  Therefore, $\boldsymbol{\delta}' < \boldsymbol{\delta}(b)$. By Lemma 2, $\mathcal{B}(\boldsymbol{\delta}') \neq \emptyset$. This means that there is a $\boldsymbol{\delta}'$ for which the model is not falsified. Therefore, by the definition of FF, $\boldsymbol{\delta}$ can't lie on the FF, $\boldsymbol{\delta} \notin$ FF.

□

VI

CONCLUSION

In this dissertation, I have introduced and developed methods which allow to select valid IVs from a set of potentially invalid instruments. I have shown how these methods can be helpful when trying to answer the question: How does an exposure affect an economic outcome? The new approaches demonstrate how IV selection methods can be helpful when trying to answer this question. I have connected the econometric and statistical literature to applied economic research and introduced methods that are suitable for situations in which previous methods do not work.

In Chapter 1, I have shown how the methods can be applied to a large number of studies which use shift-share instruments. Chaper 2 introduces the AHC method. This new method can be easily applied in presence of multiple endogenous regressors, it allows more flexible selection of weak instruments and also offers helpful insights in presence of heterogeneous treatment effects. Chapter 3 shows that the selection methods can still offer helpful insights when there are small violations of the exclusion restriction but these violations are not too large. In some specific settings when the local component of the violations are large for globally invalid IVs, the plurality assumption can even be relaxed. Chapter 4 shows that the selection methods can be understood in the framework proposed in a new econometric study (Masten and Poirier, 2021). Importantly, I propose to use the two methods as complements.

This thesis has contributed to the study of IV selection methods, in directions which were not studied before. Still, there are many interesting open questions which deserve more thorough research: For example, it is unclear how to apply the selection methods in settings with heterogeneous treatment effects *and* invalid instruments. Furthermore, the AHC method could readily be applied to genetic studies using Mendelian Randomization. Also, it could be interesting to extend the current methods to nonlinear models. Finally, another direction of research could be high-dimensional settings where the number of instruments grows with the sample size, when the number of IVs exceeds the number of observations or when there is a large number of covariates. These questions are left for future research.

Overall, this thesis has shown how to tackle a causal inference problem with the help of methods developed in machine learning. With improved processing power of modern computers and massive amounts of data available, methods from statistical learning will keep having great relevance for empirical researchers in economics. The methods applied here have their relevance not only in economics, but also in biostatistics and other fields of research. Therefore, the push towards improved statistical methods should take developments in other disciplines into account. Moreover, the development of theory and software for these methods will continue to be of central importance. Another aspect that deserves great attention is how to connect the functioning of these methods to economic ideas and how to spread them effectively, so that they are in fact accepted by researchers. I hope that this thesis has helped contributing to these tasks.

BIBLIOGRAPHY

Adão, Rodrigo, Michal Kolesár, and Eduardo Morales (2019). "Shift-Share Designs: Theory and Inference". In: *The Quarterly Journal of Economics* 134.4, pp. 1949–2010.

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.

Amior, Michael (Feb. 2020). *The Contribution of Immigration to Local Labor Market Adjustment*. Tech. rep. CEP Discussion Paper No 1678.

Amorim, Renato Cordeiro de, Vladimir Makarenkov, and Boris Mirkin (2016). "A-Ward$_{p\beta}$: Effective Hierarchical Clustering Using the Minkowski Metric and a Fast K-Means Initialisation". In: *Information Sciences* 370, pp. 343–354.

Andrews, Donald WK (1999). "Consistent Moment Selection Procedures for Generalized Method of Moments Estimation". In: *Econometrica* 67.3, pp. 543–563.

Andrews, Isaiah, Matthew Gentzkow, and Jesse M Shapiro (2017). "Measuring the Sensitivity of Parameter Estimates to Estimation Moments". In: *The Quarterly Journal of Economics* 132.4, pp. 1553–1592.

Angrist, Joshua D (1990). "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence From Social Security Administrative Records". In: *The American Economic Review* 80.3, pp. 313–336.

Angrist, Joshua D and Ivan Fernandez-Val (2010). "Extrapolate-ing: External Validity and Overidentification in the LATE Framework". In: *Advances in Economics and Econometrics Tenth World Congress*. Edited by Daron Acemoglu, Manuel Arellano, Eddie Dekel.

Apfel, Nicolas and Xiaoran Liang (2021). "Agglomerative Hierarchical Clustering for Selecting Valid Instrumental Variables". In: *arXiv preprint arXiv:2101.05774*.

Athey, Susan and Guido W Imbens (2019). "Machine Learning Methods That Economists Should Know About". In: *Annual Review of Economics* 11, pp. 685–725.

Autor, David H, David Dorn, and Gordon H Hanson (2013). "The China Syndrome: Local Labor Market Effects of Import Competition in the United States". In: *American Economic Review* 103.6, pp. 2121–68.

Aydemir, Abdurrahman B and Murat G Kirdar (2017). "Quasi-Experimental Impact Estimates of Immigrant Labor Supply Shocks: The Role of Treatment and Comparison Group Matching and Relative Skill Composition". In: *European Economic Review* 98, pp. 282–315.

Bartik, Timothy J (1991). "Who Benefits from State and Local Economic Development Policies?" In: *WE Upjohn Institute for Employment Research*.

Basso, Gaetano and Giovanni Peri (2015). "The Association Between Immigration and Labor Market Outcomes in the United States". In: *IZA Discussion Paper 9436*.

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov, and Christian Hansen (2012). "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain". In: *Econometrica* 80.6, pp. 2369–2429.

Berkowitz, Daniel, Mehmet Caner, and Ying Fang (2008). "Are "Nearly Exogenous Instruments" Reliable?" In: *Economics Letters* 101.1, pp. 20–23.

Borjas, George J (2003). "The Labor Demand Curve Is Downward Sloping: Reexamining the Impact of Immigration on the Labor Market". In: *The Quarterly Journal of Economics* 118.4, pp. 1335–1374.

Borusyak, Kirill, Peter Hull, and Xavier Jaravel (2021). "Quasi-Experimental Shift-Share Research Designs". In: *Review of Economic Studies* Forthcoming.

Bosetti, Valentina, Cristina Cattaneo, and Elena Verdolini (2015). "Migration of Skilled Workers and Innovation: A European Perspective". In: *Journal of International Economics* 96.2, pp. 311–322.

Bowden, Jack, George Davey Smith, and Stephen Burgess (2015). "Mendelian Randomization With Invalid Instruments: Effect Estimation and Bias Detection Through Egger Regression". In: *International Journal of Epidemiology* 44.2, pp. 512–525.

Bowsher, Clive G (2002). "On Testing Overidentifying Restrictions in Dynamic Panel Data Models". In: *Economics Letters* 77.2, pp. 211–220.

Bratti, Massimiliano and Chiara Conti (2018). "The Effect of Immigration on Innovation in Italy". In: *Regional Studies* 52.7, pp. 934–947.

Caner, Mehmet (2014). "Near Exogeneity and Weak Identification in Generalized Empirical Likelihood Estimators: Many Moment Asymptotics". In: *Journal of Econometrics* 182.2, pp. 247–268.

Caner, Mehmet, Xu Han, and Yoonseok Lee (2018). "Adaptive Elastic Net GMM Estimation with Many Invalid Moment Conditions: Simultaneous Model and Moment Selection". In: *Journal of Business & Economic Statistics* 36.1, pp. 24–46.

Card, David (2001). "Immigrant Inflows, Native Outflows, and the Local Labor Market Impacts of Higher Immigration". In: *Journal of Labor Economics* 19.1, pp. 22–64.

Card, David (2009). "Immigration and Inequality". In: *American Economic Review: Papers and Proceedings* 99.2, pp. 1–21.

Cattaneo, Cristina, Carlo V Fiorio, and Giovanni Peri (2015). "What Happens to the Careers of European Workers When Immigrants "Take Their Jobs"?" In: *Journal of Human Resources* 50.3, pp. 655–693.

Cole, Stephen R, Robert W Platt, Enrique F Schisterman, Haitao Chu, Daniel Westreich, David Richardson, and Charles Poole (2010). "Illustrating Bias Due to Conditioning on a Collider". In: *International Journal of Epidemiology* 39.2, pp. 417–420.

Conley, Timothy G, Christian B Hansen, and Peter E Rossi (2012). "Plausibly Exogenous". In: *Review of Economics and Statistics* 94.1, pp. 260–272.

Cortes, Patricia and José Tessada (2011). "Low-Skilled Immigration and the Labor Supply of Highly Skilled Women". In: *American Economic Journal: Applied Economics* 3.3, pp. 88–123.

D'Amuri, Francesco and Giovanni Peri (2014). "Immigration, Jobs, and Employment Protection: Evidence From Europe Before and During the Great Recession". In: *Journal of the European Economic Association* 12.2, pp. 432–464.

Dustmann, Christian, Francesca Fabbri, and Ian Preston (2005). "The Impact of Immigration on the British Labour Market". In: *The Economic Journal* 115.507, F324–F341.

Dustmann, Christian, Tommaso Frattini, and Ian P Preston (2013). "The Effect of Immigration Along the Distribution of Wages". In: *Review of Economic Studies* 80.1, pp. 145–173.

Dustmann, Christian and Albrecht Glitz (2015). "How Do Industries and Firms Respond to Changes in Local Labor Supply?" In: *Journal of Labor Economics* 33.3, pp. 711–750.

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.

Edo, Anthony, Yvonne Giesing, Jonathan Öztunc, and Panu Poutvaara (2019). "Immigration and Electoral Support for the Far-Left and the Far-Right". In: *European Economic Review* 115, pp. 99–143.

Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani (2004). "Least Angle Regression". In: *The Annals of Statistics* 32.2, pp. 407–499.

Farbmacher, Helmut (2017). "SIVREG: Stata Module to Perform Adaptive Lasso with Some Invalid Instruments". In: *Statistical Software Components S458394, Boston College Department of Economics*.

Farré, Lídia, Libertad González, and Francesc Ortega (2011). "Immigration, Family Responsibilities and the Labor Supply of Skilled Native Women". In: *The BE Journal of Economic Analysis & Policy* 11.1.

Foged, Mette and Giovanni Peri (2016). "Immigrants' Effect on Native Workers: New Analysis on Longitudinal Data". In: *American Economic Journal: Applied Economics* 8.2, pp. 1–34.

Freedom House (2017). *Freedom of the Press 2017 Press Freedom's Dark Horizon*. Freedom House.

— (2020). *Freedom in the World 2020: A Leaderless Struggle for Democracy*. Freedom House.

Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift (2020). "Bartik Instruments: What, When, Why, and How". In: *American Economic Review* 110.8, pp. 2586–2624.

Guo, Zijian, Hyunseung Kang, Tony T Cai, and Dylan S Small (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 (Statistical Methodology)* 80.4, pp. 793–815.

Han, Chirok (2008). "Detecting Invalid Instruments Using L1-GMM". In: *Economics Letters* 101.3, pp. 285–287.

Hayakawa, Kazuhiko (2014). *Alternative Over-Identifying Restriction Test in GMM with Grouped Moment Conditions*. Tech. rep. Mimeo.

Hunt, Jennifer (2017). "The Impact of Immigration on the Educational Attainment of Natives". In: *Journal of Human Resources* 52.4, pp. 1060–1118.

Imbens, Guido W and Joshua D Angrist (1994). "Identification and Estimation of Local Average Treatment Effects". In: *Econometrica: Journal of the Econometric Society*, pp. 467–475.

Jaeger, David A (2007). "Green Cards and the Location Choices of Immigrants in the United States, 1971-2000". In: *Research in Labor Economics* 27, pp. 131–183.

Jaeger, David A, Joakim Ruist, and Jan Stuhler (2020). "Shift-Share Instruments and the Impact of Immigration". In: *NBER Working Paper No. 24285*.

Kang, Hyunseung, Anru Zhang, Tianwen T Cai, and Dylan S Small (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.

Kerr, Sari Pekkala, William R Kerr, and William F Lincoln (2015). "Skilled Immigration and the Employment Structures of US Firms". In: *Journal of Labor Economics* 33.S1, S147–S186.

Kolesár, Michal, Raj Chetty, John Friedman, Edward Glaeser, and Guido W Imbens (2015). "Identification and Inference with Many Invalid Instruments". In: *Journal of Business & Economic Statistics* 33.4, pp. 474–484.

Lakatos, Imre (1976). "Falsification and the Methodology of Scientific Research Programmes". In: *Can Theories Be Refuted?* Springer, pp. 205–259.

Lin, Wei, Rui Feng, and Hongzhe Li (2015). "Regularization Methods for High-Dimensional Instrumental Variables Regression with an Application to Genetical Genomics". In: *Journal of the American Statistical Association* 110.509, pp. 270–288.

Llull, Joan (2017). "The Effect of Immigration on Wages: Exploiting Exogenous Variation at the National Level". In: *Journal of Human Resources*, 0315–7032R2.

Masten, Matthew A and Alexandre Poirier (2021). "Salvaging falsified instrumental variable models". In: *Econometrica* 89.3, pp. 1449–1469.

Moreno-Galbis, Eva and Ahmed Tritah (2016). "The Effects of Immigration in Frictional Labor Markets: Theory and Empirical Evidence From EU Countries". In: *European Economic Review* 84, pp. 76–98.

Mullainathan, Sendhil and Jann Spiess (2017). "Machine Learning: An Applied Econometric Approach". In: *Journal of Economic Perspectives* 31.2, pp. 87–106.

Newey, Whitney K (1985). "Generalized Method of Moments Specification Testing". In: *Journal of Econometrics* 29.3, pp. 229–256.

Orrenius, Pia M and Madeline Zavodny (2015). "Does Immigration Affect Whether Us Natives Major in Science and Engineering?" In: *Journal of Labor Economics* 33.S1, S79–S108.

Ottaviano, Gianmarco and Giovanni Peri (2005). *Rethinking the Gains From Immigration: Theory and Evidence From the US*. Tech. rep.

Parente, Paulo MDC and JMC Santos Silva (2012). "A Cautionary Note on Tests of Overidentifying Restrictions". In: *Economics Letters* 115.2, pp. 314–317.

Peri, Giovanni, Kevin Shih, and Chad Sparber (2015). "Stem Workers, H-1B Visas, and Productivity in US Cities". In: *Journal of Labor Economics* 33.S1, S225–S255.

Popper, Karl Raimund (1959). *The Logic of Scientific Discovery*. Hutchinson.

Roodman, David (2009). "A Note on the Theme of Too Many Instruments". In: *Oxford Bulletin of Economics and Statistics* 71.1, pp. 135–158.

Ruggles, Steven, Katie Genadek, Ronald Goeken, Josiah Grover, and Matthew Sobek (2015). "Integrated Public Use Microdata Series (IPUMS): Version 6.0 [Dataset]". In: *University of Minnesota, Minneapolis, available at http://usa. ipums. org/usa*.

Sargan, John D (1958). "The Estimation of Economic Relationships Using Instrumental Variables". In: *Econometrica: Journal of the Econometric Society* 26.3, pp. 393–415.

Searle, Shayle R (1971). *Linear models*. John Wiley & Sons.

Smith, Christopher L (2012). "The Impact of Low-Skilled Immigration on the Youth Labor Market". In: *Journal of Labor Economics* 30.1, pp. 55–89.

Staiger, Doug and James H Stock (1997). "Instrumental Variables Regression with Weak Instruments". In: *Econometrica* 65.3, pp. 557–586.

Stock, James H and Francesco Trebbi (2003). "Retrospectives: Who Invented Instrumental Variable Regression?" In: *Journal of Economic Perspectives* 17.3, pp. 177–194.

Stock, James H and Motohiro Yogo (2002). *Testing For Weak Instruments in Linear IV Regression*. Tech. rep. NBER Working Paper No. 284.

Tabellini, Marco (2020). "Gifts of the Immigrants, Woes of the Natives: Lessons From the Age of Mass Migration". In: *The Review of Economic Studies* 87.1, pp. 454–486.

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, Pei Gao, Stephen Burgess, Stephen Kaptoge, Emanuele Di Angelantonio, Tina Shah, James C Engert, Robert Clarke, George Davey-Smith, Børge G Nordestgaard, et al. (2011). "Association Between C Reactive Protein and Coronary Heart Disease: Mendelian Randomisation Analysis Based on Individual Participant Data". In: *BMJ: British Medical Journal* 342.Feb 15, p. 548.

Windmeijer, Frank (2019). "Two-Stage Least Squares as Minimum Distance". In: *The Econometrics Journal* 22.1, pp. 1–9.

Windmeijer, Frank, Helmut Farbmacher, Neil Davies, and George D Smith (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.

Windmeijer, Frank, Xiaoran Liang, Fernando Hartwig, and Jack Bowden (2021). "The Confidence Interval Method for Selecting Valid Instrumental Variables". In: *Journal of the Royal Statistical Society: Series B* Forthcoming.

Wozniak, Abigail and Thomas J Murray (2012). "Timing Is Everything: Short-Run Population Impacts of Immigration in US Cities". In: *Journal of Urban Economics* 72.1, pp. 60–78.

Wright, Philip G (1928). *The Tariff on Animal and Vegetable Oils*. Macmillan Company, New York.

Zou, Hui (2006). "The Adaptive Lasso and Its Oracle Properties". In: *Journal of the American Statistical Association* 101.476, pp. 1418–1429.