Weak Identification with Many Instruments

Anna Mikusheva[1] and Liyang Sun[2]

Abstract

Linear instrumental variable regressions are widely used to estimate causal effects. Many instruments arise from the use of "technical" instruments and more recently from the empirical strategy of "judge design". This paper surveys and summarizes ideas from recent literature on estimation and statistical inferences with many instruments. We discuss how to assess the strength of the instruments and how to conduct weak identification-robust inference under heteroscedasticity. We establish new results for a jack-knifed version of the Lagrange Multiplier (LM) test statistic. Many exogenous regressors arise often in practice to ensure the validity of the instruments. We extend the weak-identification-robust tests to settings with both many exogenous regressors and many instruments. We propose a test that properly partials out many exogenous regressors while preserving the re-centering property of the jack-knife. The proposed tests have uniformly correct size and good power properties.

1 Introduction

In linear instrumental variables (IV) regression, when there are many instruments, the consistency of the estimation for the first stage coefficients becomes questionable. If the uncertainty about the first stage coefficients has a first order importance, conventional approximations to the distribution of IV estimators are generally unreliable. Recognizing this problem, Bekker (1994) formally modeled the issue of many instruments by considering asymptotic approximations that assume the number of instruments grows to infinity with the sample size. Specifically, Bekker (1994) is the first paper that pointed out the standard two-stage least squares (TSLS) estimator can be badly biased under many instruments.

[1]Department of Economics, M.I.T. Address: 77 Massachusetts Avenue, E52-526, Cambridge, MA, 02139. Email: amikushe@mit.edu.

[2]UCL and CEMFI. Email: lsun20@cemfi.es Liyang Sun gratefully acknowledges support from Ayudas Juan de la Cierva Formación. We are grateful to Mikkel Sølvsten and Tiemen Woutersen for advice, to John C. Chao and Brigham Frandsen for sharing code for simulations. The accompanying Stata package `manyweakiv` is available at `https://github.com/lsun20/manyweakiv`

Our paper provides an exposition of the challenges discovered and some solutions proposed in the three decades since the original paper of Bekker (1994) of fast-growing econometric literature on estimation and statistical inferences (testing, confidence sets construction) in linear IV models with many potentially weak instruments. We first describe the trade-offs that arise from using many instruments. The benefit of using more instruments is obvious — they bring additional exogenous information that can help to estimate the structural parameter of interest and may lead to a more efficient estimator. The challenges of using many instruments arise from the need to find an optimal way to combine them (the task done by the first stage) and from the growing complexity of such a task. When the information from additional instruments grows slower than the complexity of the first stage, the additional instruments might be detrimental and lead to a worse estimator. Specifically, the uncertainty from the first stage tends to translate into the bias of the estimator for a structural parameter, and may lead to an inconsistency of the structural estimator.

We survey some influential ideas attempting to properly use information from an increasing number of instruments. These ideas include jack-knifing and sample splitting, that motivate some estimators having a superior performance in comparison to the TSLS in settings with many instruments. We then discuss the definition of weak identification, a situation when the information contained in the instruments is low relative to the number of instruments to the extend that conventional approximations to the distribution of IV estimators become invalid. We describe an empirically relevant pre-test for weak identification as well as identification robust tests.

In addition to a survey of the existing literature, in this paper we also establish two new results related to identification robust tests. First of all, we present a new form of identification robust LM test that uses a new estimator of variance. Our test has superior power properties in comparison with the LM test recently proposed in Matsushita and Otsu (2022). The second new result is an identification robust AR test that is valid under both many instruments and many exogenous regressors (controls).

The remainder of this paper is organized as follows. Section 2 summarizes results on estimation with many instruments and defines weak identification. Section 3 describes

a pre-test for weak identification. Section 4 describes existing weak identification robust tests and introduces the new LM test. Section 5 discusses the challenges and solutions of having many exogenous regressors. Section 6 concludes with open questions.

We finish the introduction by presenting two well-known examples of many instruments. These examples demonstrate that whenever there is a good exogenous variation allowing the identification of causal effect, many instruments arise naturally.

Example 1: Angrist and Krueger (1991) contains one of the most well-known applications of linear IV regression. This paper estimated the return to education using the quarter of birth as instruments, and was prominently mentioned in the press-release about the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2021. The structural parameter of interest is the coefficient on the educational attainment in the structural equation:

$$wage_i = \beta\ education_i + controls + e_i.$$

Due to compulsory educational laws, children are required to stay in school until a certain physical age, and yet children typically start their schooling in September when they are at a different physical age. Thus, the quarter of birth, which can be considered as randomly assigned in the population, produces an exogenous and observable variation in educational attainment. Since the compulsory educational laws vary by state, the effects of the quarter of birth on education are heterogeneous across the states. The effects are possibly heterogeneous by birth cohort as well. Therefore, Angrist and Krueger (1991) considered specifications that interact quarter of birth dummies with either state or year of birth dummies or both. In the main specifications used by Angrist and Krueger (1991), interactions of the quarter of birth with the year of birth dummies yield 30 instruments. Adding interactions with state of birth dummies yields 180 instruments. Finally, adding fully saturated three-way interactions yields 1530 instruments.

In this setting, one starts with a single variable producing the exogenous variation, but creates multiple "technical" instruments based on it in order to extract information from heterogeneous and/or non-linear first stage. Such a setting is extremely common in empirical practice and is one of the main sources of examples with many instruments.

Not only did Angrist and Krueger (1991) inspire the literature on weak identification, as it was the main example in Staiger and Stock (1997), it was also a motivating example for the literature on many instruments including Hansen et al. (2008).

Example 2. "Judge design" is a common name for empirical settings that use the exogenous assignment of cases to different decision makers as instruments for a treatment in an attempt to estimate important causal effects of interest. Fueled by rich administrative data, recent applications of "judge design" include Maestas et al. (2013); Dobbie et al. (2018); Sampat and Williams (2019), and Bhuller et al. (2020). For example, Bhuller et al. (2020) estimate the effect of incarceration on recidivism using random assignments of criminal cases to judges as a source of exogenous variation. Judges express different leniencies producing an exogenous variation in incarceration decisions. The instruments here are dummies for individual judges. Since each judge can only process a certain number of cases out of the total court cases, the number of judges (the number of instruments) increases fast with the sample size.

2 Estimation with many instruments

To describe the main ideas, we consider a simplified setup with one endogenous regressor and no included exogenous regressors (controls). We will add exogenous regressors in Section 5. Assume we observe an i.i.d. sample $\{(X_i, Y_i, Z_i), i = 1, ..., N\}$ satisfying a linear IV model:

$$\begin{cases} Y_i = \beta X_i + e_i; \\ X_i = \pi' Z_i + v_i, \end{cases}$$

where X_i is a one-dimensional endogenous regressor, $Z_i \in \mathbb{R}^K$ are instruments satisfying the exclusion restriction $\mathbb{E}[e_i|Z_i] = \mathbb{E}[v_i|Z_i] = 0$. We allow errors to be heteroskedastic with $0 < c < \mathbb{E}[e_i^2|Z_i] < C$. In this setting we are interested in estimation of and statistical inferences on the structural coefficient β, while π is a set of nuisance parameters. The first equation is often referred to as the structural equation, while the second is called the first stage. We will discuss estimation and statistical inference conditional on the realization of Z_i, and thus will treat instruments as fixed. For simplicity of notation, we drop the

conditioning sign and all expectation signs should be read as conditional on Z_i's.

2.1 Asymptotic bias of TSLS

The TSLS is the most widely known and used estimator in this case. To implement the TSLS, one first runs the OLS regression of the first stage by regressing X_i on Z_i which obtains the estimated coefficients $\widehat{\pi}$. Then one runs the second stage regression of Y_i on $\widehat{X}_i = \widehat{\pi}' Z_i$, where the regression coefficient estimate $\widehat{\beta}_{TSLS}$ is the TSLS estimate.

The TSLS is popular due to its asymptotic optimality under homoskedasticity in a setting with a small number of instruments. The notion of asymptotic efficiency for GMM appeared in Chamberlain (1987). Since the unknown parameter β is a scalar, a single instrument is sufficient for identification. If we have K instruments and can use any linear combination of them as the single instrument, the question is which linear combination provides an estimator with the smallest asymptotic variance. In the homoskedastic model the optimal combination is $\mathbb{E}[X_i|Z_i] = \pi' Z_i$ as it delivers the asymptotic efficiency. An alternative interpretation of the TSLS through the lens of "optimal instrument", is that the TSLS uses the first stage to combine multiple instruments Z_i into a single estimated "optimal" instrument $\widehat{X}_i$. One can show that $\widehat{\beta}_{TSLS}$ is equal to the IV estimate in a just-identified IV regression of Y_i on X_i using $\widehat{X}_i$ as the single instrument. Furthermore, the asymptotic variance of the TSLS estimator as well as of the infeasible optimal IV estimator is inversely proportional to $\pi' Z' Z \pi$, the part of the endogenous regressor X explained by Z. Thus, if we know how to construct the optimal instrument from the available data, we can expect additional instruments to improve efficiency of the TSLS through increasing the explained part of the endogenous regressor.

In practice we do not know π, the coefficient of the optimal instrument combination, and have to estimate it in the first stage. As we show, this typically leads to a bias in the estimation of the structural coefficient that increases with the number of instruments. Specifically, the estimated optimal instrument

$$\widehat{X}_i = X' Z (Z'Z)^{-1} Z_i = \pi' Z_i + v' Z (Z'Z)^{-1} Z_i$$

contains not only the true optimal instrument $\pi' Z_i$ but also the estimation mistake, which makes the estimated optimal instrument endogenous. For the next derivation only, assume that the errors are homoskedastic with $\sigma_{ev} = \mathbb{E}[e_i v_i] \neq 0$. Parameter σ_{ev} measures the degree of endogeneity of the regressor, and therefore the bias of the OLS. Then

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\widehat{X}_i - \pi' Z_i)e_i\right] = \frac{1}{N}\sum_{i=1}^{N} Z_i'(Z'Z)^{-1}Z_i\mathbb{E}[v_i e_i] = \frac{tr(Z(Z'Z)^{-1}Z')}{N}\sigma_{ev} = \frac{K}{N}\sigma_{ev}.$$

As shown above, the estimated instrument $\widehat{X}_i$ is endogenous, and consequently the TSLS has bias proportional to a $\frac{K}{N}$. Specifically, the bias of TSLS is increasing in the number of instruments K, and if the number of instruments is a large fraction of the sample size, then the TSLS is inconsistent. This result first appeared in Bekker (1994), who argues that in order to get a more realistic approximations for the properties of the TSLS when the number of instruments is large, it is important to model the number of instruments as increasing with the sample size.

Another explanation for the TSLS bias is over-fitting of the first stage when the number of regressors in the first stage regression is large. Suppose there are N instruments for a sample of size N, then the first stage regression would produce a perfect fit and we have $\widehat{X}_i = X_i$. In this case, the TSLS equals to the OLS, which is inconsistent and biased due to endogeneity of X. This is an extreme case, but it provides an intuition for why having many instruments may complicate estimation.

In order to avoid the over-fitting problem, one may consider using some alternative estimation strategies for the first stage. Donald and Newey (2001) proposed an instrument selection procedure based on a Mallows criteria. Suggestions to use LASSO selection on the first stage was put forward by Belloni et al. (2012) and Belloni et al. (2014). Okui (2011) proposed to use a shrinkage estimator, while Carrasco (2012) suggested several regularization procedures based on the spectral decomposition of the conditional expectation operator, such as the principal components approach and Tikhonov's regularization.

If one is willing to impose some assumptions about the form of the optimal instrument, then with a proper estimation technique on the first stage that allows for a consistent estimation of the optimal instrument, one may obtain a semi-parametric efficient estimator for

β. For example, Donald and Newey (2001) assumed a known ordering among instruments (or groups of instruments) by strength/informativeness. The LASSO procedure of Belloni et al. (2012) delivers a semi-parametric efficient estimator for β if the first-stage regression is approximately sparse, that is, a relatively small number of the instruments successfully approximates the optimal instrument. Another type of assumption often needed is a regularity condition placed on the conditional expectation operator. For example, Belloni et al. (2012) restricted eigenvalues of an empirical Gram matrix, while Carrasco (2012) assumed that the conditional expectation operator is a Hilbert-Schmidt operator.

When the assumptions about the form of the optimal instrument fail, the performance of these alternative estimation strategies are not always guaranteed. Hansen and Kozbur (2014) provided simulation evidence that the performance of IV estimators using LASSO in the first stage is less than stellar when the signal on the first stage is dense and weak. Angrist and Frandsen (2022) studied the performance of some Machine Learning (ML) techniques for instrument selection using simulations calibrated to two important empirical examples. They compared the performance of IV estimators using LASSO and random forest in the first stage with that of OLS, TSLS and several jack-knife and split-sample estimators that we will discuss below. In almost all cases the IV estimators using LASSO and random forest in the first stage resulted in biases whose magnitudes are comparable to those of OLS and TSLS without much improvement in variance. Moreover, the performance of these two ML methods depends heavily on the choice of the regularization parameter: the cross-validation or plug-in penalties for the LASSO, or the leaf-size for the random forest. None of the standard choices for the regularization parameter were totally satisfactory. One plausible explanation is that the sparsity of the first stage is a poor description of the data in these two empirical applications.

2.2 Jack-knifing or diagonal removing

The bias of the TSLS arises from using the same observation in both stages of estimation. Since the first stage estimate $\widehat{\pi}$ depends on X_i, the estimated optimal instrument $\widehat{X}_i$ used by the TSLS is endogenous and is correlated with the structural error e_i. In this subsection, we survey some influential ideas attempting to remove the bias by jack-knifing

and sample-splitting.

Angrist and Krueger (1995) proposed removing the bias by using separate samples in the two stages of the TSLS. They suggested splitting the original sample into two halves. If π is estimated using the first half of the sample, and the estimated optimal instrument is produced for the second half, then the constructed instrument will be exogenous and the IV estimator would avoid the over-fitting bias of the TSLS. This idea is called sample-splitting.

A refinement of sample-splitting that exploits the data in a more sophisticated way is jack-knifing (Angrist et al., 1999). The idea is to run a separate first stage for each observation. Namely, for observation i we run the OLS regression of X on Z on the sample excluding observation i, calling the resulting estimate $\widehat{\pi}_{(-i)}$. Define $Z_i^* = \widehat{\pi}'_{(-i)} Z_i$ and run the OLS regression of Y_i on X_i using Z_i^* as the single instrument. The resulting estimate of β is called a jack-knife IV (JIVE). JIVE breaks the dependency between two stages on the same observation and effectively satisfies the exogeneity condition $\mathbb{E}[Z_i^* e_i] = 0$ for the constructed instrument.

The idea of running a separate first stage OLS regression for each observation seems daunting but is ultimately unnecessary. There is a relatively easier formula for JIVE based on the Sherman-Morrison-Woodbury formula, which provides an explicit way to calculate the leave-one-out projection based on the full-sample orthogonal projection. Specifically, one can show that while the TSLS can be written as

$$\widehat{\beta}_{TSLS} = \frac{X' P_Z Y}{X' P_Z X} = \frac{\sum_{i,j} P_{ij} X_i Y_j}{\sum_{i,j} P_{ij} X_i X_j},$$

where $P_Z = Z(Z'Z)^{-1}Z'$ is a projection on Z, and P_{ij} are its elements, the JIVE has a similar form:

$$\widehat{\beta} = \frac{\sum_{i,j} P^*_{ij} X_i Y_j}{\sum_{i,j} P^*_{ij} X_i X_j}, \tag{1}$$

where elements P^*_{ij} are slightly re-weighted elements of P_{ij} with one important difference - all diagonal elements are zeros.

The diagonal elements of the projection matrix are tightly connected to the TSLS

bias, since the expectation of the numerator for $\widehat{\beta}_{TSLS} - \beta$ is

$$\mathbb{E}\left[\sum_{i,j} P_{ij} X_i e_j\right] = \mathbb{E}\left[\sum_{i,j} P_{ij} v_i e_j\right] = \sum_i P_{ii} \mathbb{E}\left[v_i e_i\right].$$

Therefore, a closely related estimator, which just removes the diagonal from the TSLS formula, is also often referred to as JIVE:

$$\widehat{\beta}_{JIVE} = \frac{\sum_{i \neq j} P_{ij} X_i Y_j}{\sum_{i \neq j} P_{ij} X_i X_j}. \tag{2}$$

This estimator was proposed in Angrist et al. (1999) and called JIV2 by the authors. It is numerically extremely close to the other JIVE described in (1). For simplicity we will call JIVE the estimator $\widehat{\beta}_{JIVE}$ defined in (2). The expectation of the numerator of the estimation error for JIVE:

$$\widehat{\beta}_{JIVE} - \beta = \frac{\sum_{i \neq j} P_{ij} X_i e_j}{\sum_{i \neq j} P_{ij} X_i X_j}.$$

is zero $\mathbb{E}\left[\sum_{i \neq j} P_{ij} X_i e_j\right] = 0$, and thus, JIVE avoids the over-fitting bias of the TSLS.

The JIVE can also be motivated as the optimizer of a slightly corrected objective function, specifically:

$$\widehat{\beta}_{JIVE} = \arg\min_{\beta} Q_{JIVE}(\beta) = \arg\min_{\beta} \sum_{i \neq j} P_{ij} (Y_i - \beta X_i)(Y_j - \beta X_j).$$

The JIVE objective function is only slightly different than the TSLS objective function:

$$Q_{TSLS}(\beta) = \sum_{i,j} P_{ij} (Y_i - \beta X_i)(Y_j - \beta X_j) = (Y - \beta X)' P_Z (Y - \beta X).$$

As was pointed out in Han and Phillips (2006) and Newey and Windmeijer (2009), the problem with the TSLS objective function is that its expectation at the true parameter value is not zero $\mathbb{E}Q_{TSLS}(\beta_0) \neq 0$, and therefore is not minimized at the true parameter value. The JIVE objective functions solves this issue by diagonal-removing. This reasoning can be generalized to other instrumental variable estimators: Hausman et al.

(2012) proposed JIVE-LIML and JIVE-Fuller, while Hansen and Kozbur (2014) proposed JIVE-ridge.

2.3 Consistency of estimators with many instruments

JIVE-type estimators have superior consistency properties when compared with the TSLS. In particular, Chao and Swanson (2005) established that under homoskedasticity the TSLS is consistent when $\frac{\pi' Z' Z \pi}{K} \to \infty$, while the JIVE is consistent when $\frac{\pi' Z' Z \pi}{\sqrt{K}} \to \infty$. The difference in these conditions is large when the number of instruments K is large. A similar statement under heteroskedasticity appeared in Hausman et al. (2012) along with consistency of JIVE-LIML and JIVE-Fuller when $\frac{\pi' Z' Z \pi}{\sqrt{K}} \to \infty$.

Notice that $\pi' Z' Z \pi$, the explained part of regressor X, measures the information contained in the optimal instrument. If one knew the optimal weights π, and used the TSLS with the single optimal instrument $Z\pi$, then this estimator is consistent as long as $\pi' Z' Z \pi \to \infty$. The TSLS bias is proportional to K and constitutes the leading term causing inconsistency of the TSLS when $\frac{\pi' Z' Z \pi}{K}$ is asymptotically bounded. Once the bias is removed, the consistency arises when the next asymptotic term is negligible.

The condition for consistency of JIVE ($\frac{\pi' Z' Z \pi}{\sqrt{K}} \to \infty$) emphasizes that additional instruments do not just increase the information extracted from the sample ($\pi' Z' Z \pi$) but also come with a cost. To justify adding new instruments, they should bring enough additional information so that $\frac{\pi' Z' Z \pi}{\sqrt{K}}$ increases. This happens since the optimal coefficients π are not known and have to be estimated. The estimation of the optimal coefficients always comes with mistakes, which accumulate with the dimensionality of instruments, K. The factor $\sqrt{K}$ is the price one pays for the need to search for an optimal combination in a very multi-dimensional setting.

In Mikusheva and Sun (2022) we showed that this price is unavoidable if one has no information about the direction of the optimal instrument. Specifically, we showed that if coefficients π are completely unknown and $\frac{\pi' Z' Z \pi}{\sqrt{K}}$ is asymptotically bounded, then for any two distinct values of parameter β there exists no asymptotically consistent test distinguishing them. This implies that $\frac{\pi' Z' Z \pi}{\sqrt{K}} \to \infty$ is a necessary condition for consistency.

However, if one is willing to impose assumptions on the form of the optimal instru-

ment, then the condition for consistency may be weakened. Specifically, if one is willing to assume that the optimal combination is sparse and uses a properly chosen LASSO procedure on the first stage, then the condition for consistency may be weakened to depend (up to logarithm multipliers) on the squared root of the sparsity parameter in place of $\sqrt{K}$.

Chao et al. (2012) and Hausman et al. (2012) established that under some minor assumptions, mainly additional moment restrictions and assumptions that the projection operator P_Z is well-balanced, the condition $\frac{\pi' Z' Z \pi}{\sqrt{K}} \to \infty$ implies that JIVE, JIVE-LIML and JIVE-Fuller are asymptotically gaussian. This implies that one can employ t-statistics and produce Wald-type confidence sets, so the inferences are somewhat standard in this case. The one caveat of this statement is that the usual formulas for standard errors are incorrect and understate the uncertainty. These papers also proposed the consistent asymptotic variance estimators.

3 Pre-testing for identification strength

The condition for consistency and gaussianity ($\frac{\pi' Z' Z \pi}{\sqrt{K}} \to \infty$) requires that the amount of information extracted by the instruments is large enough relative to the squared root of the number of instruments. If this condition does not hold, then the JIVE or other JIVE-type estimators and their associated Wald (t-statistics) inferences are unreliable. Specifically, the estimators tend to be biased, and the tests and confidence sets based on t-statistics have an asymptotically incorrect size. This corresponds to a phenomenon known as weak identification.

Weak identification has been recognized in IV estimation with small number of instruments by Staiger and Stock (1997) and Stock and Yogo (2005). Stock and Yogo proposed a pre-test for weak identification, which compares the first stage F-statistics for the hypothesis that $\pi = 0$ with a specially selected cut-off. The cut-off technically depends on the number of instruments and the goal of the pre-test (to bound the bias of the TSLS estimator or to bound the size distortions of the t - test), but in practice the universal and simple cut-off of 10 was suggested and has been widely adopted. In empirical research,

whenever the first stage F exceeds 10, it is considered reliable to use the standard TSLS inferences, while otherwise one should employ an identification robust inference.

As pointed out by Hansen et al. (2008), this pre-test does not work well in the case of many instruments. Specifically, the first stage F pre-test seems to indicate weak identification in a wide range of cases where a reliable estimator exists. This is mainly because the first stage F pre-test was built to assess the quality of inference procedures based on the TSLS estimator. As was shown in the previous section, the TSLS is a poor choice of estimator for a case with many instruments. Figure 4 in Stock et al. (2002) explicitly showed that the proper cut-off for the first stage F statistics should depend in a significant way on the estimator used when the number of instruments is larger than 5 and should decline fast with the number of instruments for the JIVE and some other bias-corrected estimators.

In Mikusheva and Sun (2022), we proposed a new pre-test for weak identification that is aimed at assessing the validity of the JIVE t-statistic inferences in cases of many instruments, while allowing for a general form of heteroskedasticity.

The logic behind such a pre-test is very similar to that behind the first stage F pre-test in settings with a small number of instruments. First, we derived the distribution of the JIVE t-statistics under some technical conditions but without imposing the assumption that $\frac{\pi' Z' Z \pi}{\sqrt{K}} \to \infty$. We found the parameter that directly quantifies the deviations of the asymptotic distribution of the JIVE t-statistics from the standard gaussian. This theoretical parameter is an analog of the concentration parameter in the IV setting with a small number of instruments introduced by Staiger and Stock (1997). We showed that this parameter is $\frac{\mu^2}{\Upsilon\sqrt{K}}$, where $\mu^2 = \sum_{i \neq j} P_{ij} (\pi' Z_i)(\pi' Z_j)$ is a diagonal-removed version of $\pi' Z' Z \pi$, while Υ is a measure of the first stage uncertainty analogous to the variance of the first stage error in the homoskedastic case. Notice that this parameter has $\sqrt{K}$ in the denominator, as we would expect for the case of many instruments. Then depending on how much distortion from the declared size a researcher agreed to tolerate, we derive a cut-off for the theoretical parameter. For example, for a 5%-test and a tolerance for 5% distortion, the cut-off is 2.5, which means that if $\frac{\mu^2}{\Upsilon\sqrt{K}} > 2.5$, then the JIVE t-test of the nominal 5% size cannot have an asymptotic size of above 10%. Finally, we proposed

	FF	$\widetilde{F}$	JIVE	(std. error)	JIVE-Wald
180 instruments	2.4	13.4	0.099	(0.017)	[0.066,0.132]
1530 instruments	1.3	6.2	0.072	(0.025)	[0.024,0.121]

Table 1: Pre-test results in an example based on Angrist and Krueger (1991)

Notes: Results on the first-stage F statistics (FF), the pre-tests for weak identification, the JIVE and its standard error and the JIVE-Wald confidence sets for IV specification underlying Table VII Column (6) of Angrist and Krueger (1991).

an estimator $\widetilde{F}$ for the theoretical parameter $\frac{\mu^2}{\Upsilon\sqrt{K}}$ and derived its accuracy to create the cut-off for statistics $\widetilde{F}$. For example, if $\widetilde{F} > 4.14$, then with 95% confidence $\frac{\mu^2}{\Upsilon\sqrt{K}} > 2.5$, and the JIVE t-test has less than 5 % size distortion.

This pre-test for weak identification provides the following empirical guidance in settings with many instruments: use the JIVE t-test when $\widetilde{F} > 4.14$, and employ the weak identification robust test otherwise. Such a decision guarantees that the total size of the two-step procedure is within 10% distortion from the nominal 5% level. Stata package `manyweakiv` implements the $\widetilde{F}$ pre-test for weak identification and several identification robust tests described below. The details of the package is described in Sun (2023).

Empirical example: Angrist and Krueger (1991). In the analysis of Angrist and Krueger (1991), Staiger and Stock (1997) pointed out that the first stage F statistic is low in the specification with many instruments and suspected weak identification. Hansen et al. (2008) argued that this is not the case of weak but rather many instruments.

Mikusheva and Sun (2022) formally assessed this question based on the $\widetilde{F}$ pre-test for weak identification created for a many instrument setting. We used the original data from Angrist and Krueger (1991), with a sample from the 1980 US census containing 329,509 men born 1930-39. We reproduced two specifications considered by Angrist and Krueger (1991). The first specification uses 180 instruments that include 30 interactions between quarter and year of birth dummies and 150 interactions between quarter and state of birth dummies. The second specification uses 1,530 instruments, the full set of three-way interactions among quarter of birth, year of birth and state of birth dummies. Table 1 reproduces part of Table 1 of Mikusheva and Sun (2022), reporting the first stage F, the $\widetilde{F}$ pre-test, the JIVE and the Wald confidence set based on JIVE.

Based on the $\widetilde{F}$ pre-test for weak identification, the JIVE t-tests are reliable in both specifications with 180 and 1,530 instruments. At the same time the value of the first stage F is low and suggests that the TSLS and the TSLS-based confidence sets should not be trusted. Another important observation is that having many uninformative instruments may be detrimental to the statistical accuracy, as we see the specification with 1,530 instruments produced wider confidence sets.

4 Identification robust tests

When the pre-test for identification strength indicates that the JIVE-inferences are not reliable, an important question is what valid statistical inferences can still be done. Low values of $\widetilde{F}$ suggest that information in the sample is low for the number of instruments used to the extent that the JIVE is inconsistent. This also implies that no other consistent estimator exists unless any additional information about the optimal instrument is available. However, even in the absence of a consistent estimator, we may still construct informative tests and confidence sets that are asymptotically valid, in the sense that their probability of incorrectly rejecting the null hypothesis and covering the true parameter value, respectively, remains well-controlled. A statistical inference procedure that remains valid no matter the identification strength is therefore called robust to weak identification. A large literature has developed a variety of statistical inference procedures robust to weak identification when the number of instruments is small. Here we discuss how one can refine the weak identification robust procedures to be robust under a large number of instruments.

Confidence sets. While we focus on robust tests, if one is interested in creating a robust confidence set for β, this can be done by numerically inverting any of the robust tests we discuss below. Specifically, one may conduct tests $H_0 : \beta = \beta_0$ for different values of β_0 and collect the values not rejected by the test to form a confidence set. The asymptotic coverage of such a set will be at the declared level no matter the strength of identification.

4.1 Robust AR

The Anderson-Rubin (AR) test was developed as an identification robust tests in IV settings with a small number of instruments. It uses the exogeneity assumption ($\mathbb{E}[e_i|Z_i] = 0$), but not relevance ($\pi \neq 0$). The idea behind the AR test, is that in order to test $H_0 : \beta = \beta_0$, one tests whether the implied errors $e(\beta_0) = Y - \beta_0 X$, which coincide with the true structural errors for the correct value of β_0, are correlated with Z in the sample. The AR test statistic is $e(\beta_0)'Z\Sigma^{-1}Z'e(\beta_0)$, where Σ is the covariance matrix of $e'Z$ or a good estimate of it. Under the null, this test statistic has an asymptotic χ^2_K distribution. Under the further assumption of homoskedasticity the statistic reduces to $\frac{1}{\hat{\sigma}^2}e(\beta_0)'Z(Z'Z)^{-1}Z'e(\beta_0) = \frac{1}{\hat{\sigma}^2}e(\beta_0)'P_Z e(\beta_0)$.

The AR test introduced in settings with a small number of instruments performs poorly in settings with a large number of instruments because the limit null distribution χ^2_K does not provide an accurate approximation. Notice that if a large number of instruments is modeled as $K \to \infty$, then the prescribed limit null distribution χ^2_K drifts to infinity. This issue is tightly related to the observation that under the null the test statistic has a non-zero mean

$$\mathbb{E}\left[e(\beta_0)'P_Z e(\beta_0)\right] = \sum_{i=1}^{N} P_{ii}\mathbb{E}e_i^2.$$

This coincides with the issue that the expected value of the TSLS objective function is not minimized at the true parameter value β_0, contributing to its inconsistency when the number of instruments is large. Therefore, the idea of jack-knifing or diagonal removing can be similarly applied to the original AR test statistic (Crudu et al., 2021; Mikusheva and Sun, 2022). The infeasible JIV (or leave-one-out) AR statistic is defined as

$$AR_0(\beta_0) = \frac{1}{\sqrt{K\Phi_0}}\sum_{i\neq j} e_i(\beta_0)P_{ij}e_j(\beta_0),$$

where the normalizing factor $\Phi_0 = \frac{2}{K}\sum_{i\neq j} P_{ij}^2\sigma_i^2\sigma_j^2$ is the variance of the quadratic form. Here and below we allow for a general form of heteroskedasticity where $\sigma_i^2 = \mathbb{E}e_i^2$. Under minor assumptions like finite fourth moments of the errors and a well-balanced design assumption, the central limit theorem for quadratic forms established in Chao et al. (2012)

guarantees that under $H_0 : \beta = \beta_0$ we have $AR_0(\beta_0) \Rightarrow N(0,1)$. It is worth pointing out that this theorem needs $K \to \infty$. The test rejects the null whenever we have a large positive value of the AR statistic.

Variance estimation. In order to obtain a feasible test of an asymptotically correct size, one needs to estimate variance Φ_0. A good estimator should be consistent under the null and should allow for a general form of heteroskedasticity. Ideally, it should also ensure a good power under alternatives.

Crudu et al. (2021) proposed using the squared implied errors $\widehat{\sigma}_i^2 = e_i^2(\beta_0)$ as an unbiased proxy for σ_i^2 and the corresponding variance estimator is defined as

$$\widehat{\Phi}_1 = \frac{2}{K} \sum_{i \neq j} P_{ij}^2 \widehat{\sigma}_i^2 \widehat{\sigma}_j^2.$$

It is very easy to show that under the null $\widehat{\Phi}_1$ is consistent for Φ_0, and thus the test using $\widehat{\Phi}_1$ has an asymptotically correct size under a general form of heteroscedasticity. However, such a test may have low power against distant alternatives. Namely, if the true β is very different from β_0, then $e(\beta_0)$ differs from structural errors e by a potentially large predictable component. Squaring the implied errors would drastically overestimate the variances of error terms and may produce unnecessarily large values for $\widehat{\Phi}_1$. This may lead to a significant power loss especially at large deviations of the postulated β_0 from the true β.

Since the difference between the implied errors $e(\beta_0)$ and structural errors e is predictable, one may residualize the implied error before squaring. Denote $M_Z = I - P_Z$ the projection matrix and let M_i be its ith row. Even under the null, the squared residualized error is biased $\mathbb{E}\left(M_i \mathbf{e}(\beta_0)\right)^2 \neq \sigma_i^2$. This is because the squared residual contains not only the squared error e_i but also the square of the regression estimation mistake. The latter can be large when the number of regressors K is large.

To construct an unbiased estimator for Φ_0 under the null, in Mikusheva and Sun (2022)

we suggested the following estimator:

$$\widehat{\Phi}_2 = \frac{2}{K} \sum_{i \neq j} \frac{P_{ij}^2}{M_{ii} M_{jj} + M_{ij}^2} \left[e_i(\beta_0) M_i e(\beta_0) \right] \left[e_j(\beta_0) M_j e(\beta_0) \right].$$

Our idea is based on the "cross-fit" variance proxies $\widehat{\sigma}_i^2 = \frac{1}{1-P_{ii}} e_i(\beta_0) M_i e(\beta_0)$, proposed by Newey and Robins (2018) and Kline et al. (2020). These proxies are unbiased under the null: $\mathbb{E}\widehat{\sigma}_i^2 = \sigma_i^2$. However, since the normalizing factor Φ_0 is quadratic in σ_i^2 and the proxies for variance of errors with different indexes (i.e. $\widehat{\sigma}_i^2$ and $\widehat{\sigma}_j^2$) depend on the same sample, one can show that

$$\mathbb{E}\left[(e_i M_i e)(e_j M_j e) \right] = (M_{ii} M_{jj} + M_{ij}^2) \sigma_i^2 \sigma_j^2,$$

and we can therefore obtain an unbiased estimator of Φ_0 by a simple re-weighting of summands.

Another alternative estimator for Φ_0 was proposed in Anatolyev and Sølvsten (2023). Their idea is to create an unbiased proxy for $\sigma_i^2 \sigma_j^2$ under the null by creating a product of four uncorrelated terms using "leave-three-out" estimator. They suggested using $\widehat{\Phi}_3 = \frac{2}{K} \sum_{i \neq j} P_{ij}^2 \widehat{\sigma_i^2 \sigma_j^2}$ with

$$\widehat{\sigma_i^2 \sigma_j^2} = e_i(\beta_0) e_j(\beta_0) \sum_k \tilde{M}_{ik,-(ij)} e_k(\beta_0) [e_j(\beta_0) - Z_j' \widehat{\delta}_{-(ijk)}],$$

where $\widehat{\delta}_{-(ijk)}$ are coefficient estimates from regressing $e(\beta_0)$ on Z leaving three observations (i, j, k) out, while $\tilde{M}_{ik,-(ij)}$ is an element of the projection matrix leaving out i and j. There are explicit formulas for leave-(one/two/three)-out projections available, however, the numerical complexity of implementing $\widehat{\Phi}_3$ is higher than the other two estimators.

As for their theoretical properties, all three estimators of Φ_0 are unbiased and consistent under the null. When the true value $\beta = \beta_0 + \Delta$ differs from the hypothesized value β_0, $\widehat{\Phi}_2$ and $\widehat{\Phi}_3$ are also consistent under local alternatives. In Mikusheva and Sun (2022) we derived the power function for the infeasible AR uniformly over a set of local

alternatives

$$AR_0(\beta_0) \Rightarrow \Delta^2 \frac{\mu^2}{\sqrt{K\Phi_0}} + \mathcal{N}(0,1).$$

This is also the theoretical description of the power functions for AR using $\widehat{\Phi}_2$ or $\widehat{\Phi}_3$. One can show numerically that using $\widehat{\Phi}_1$ leads to a power loss.

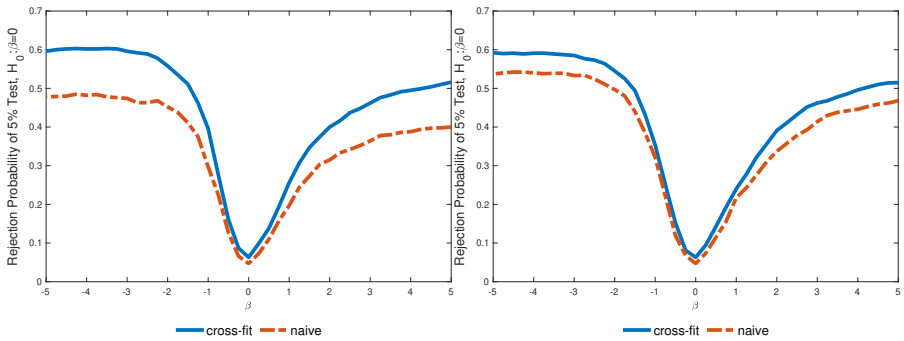
Small-scale simulations. We simulate the data according to a homoscedastic linear IV model (3) with a linear first stage $\Pi_i = \Pi' Z_i$ and true structural parameter $\beta = 0$. The sample size is $N = 200$. We divide the sample into $K = 40$ equal groups, and define the instruments to be the group indicators. We simulate two designs with varying levels of sparsity. In the sparse first stage we set one large coefficient $\pi_K = 2$ and $\pi_k = 0.001$ for all $k < K$. This is the setting where one instrument contains almost all information. The dense first stage has homogeneous first stage coefficients $\pi_k = 0.316$ for all $k = 1, \ldots, K$. Identification strength is held the same at $\frac{\mu^2}{\sqrt{K}} = 2.5$ for both designs. The error terms (e_i, v_i) are drawn i.i.d. from gaussian distribution with mean zero, unit variances and correlation $\rho = 0.2$. For each simulation draw we perform the leave-one-out AR tests using either $\widehat{\Phi}_1$ (red line) or $\widehat{\Phi}_2$ (blue line). The resulting power curves are reported on Figure 1.

One can make two observations based on Figure 1. First, that the settings considered are the cases of weak identification, as the power curves stabilize on the level well below 1. Second, there is a significant power loss due to using a naive estimator for the scale estimator $\widehat{\Phi}_1$. The usage of estimator $\widehat{\Phi}_2$ is preferred from a power perspective.

4.2 Robust LM

In over-identified settings with a small number of instruments, the AR test is known to be asymptotically inefficient if identification is strong. The main reason is that under strong identification the data contains a lot of information about the optimal instrument, which the AR test completely ignores. The solution has been an identification robust modification of the Lagrange Multiplier (LM) test, known as the KLM test.

An alternative modification to the LM test can make it robust to weak identification when there are many instruments. The LM test aims to construct the most powerful



(a) sparse, $\frac{\mu^2}{\sqrt{K}} = 2.5$ (b) dense, $\frac{\mu^2}{\sqrt{K}} = 2.5$

Figure 1: Power curves for the leave-one-out AR tests with $\widehat{\Phi}_1$ (red dash) and $\widehat{\Phi}_2$ (blue line) variance estimators under sparse vs. dense first stage. Instruments are $K = 40$ balanced group indicators, $N = 200$, based on 1,000 simulations.

combination of instruments and then conduct the AR test with the single instrument. The original LM statistic is based on the linear combination $e'(\beta_0)Z\widehat{\pi} = e'(\beta_0)P_Z X$ because under homoskedasticity the optimal instrument is $Z\pi$. However, similar to the bias of the TSLS, in the setting of many instruments, the original LM statistic is poorly centered due to the correlation between the first stage $\widehat{\pi}$ and the structural error. As before, the problem can be solved by removing the diagonal. The infeasible leave-one-out LM statistic is

$$LM^{1/2}(\beta_0) = \frac{1}{\sqrt{K\Psi}} \sum_{i \neq j} e_i(\beta_0) P_{ij} X_j,$$

where the normalization scalar is

$$\Psi = \frac{1}{K} \sum_{i=1}^{N} (\sum_{j \neq i} P_{ij} X_j)^2 \sigma_i^2 + \frac{1}{K} \sum_{i=1}^{N} \sum_{j \neq i} P_{ij}^2 \gamma_i \gamma_j,$$

with $\sigma_i^2 = \mathbb{E}e_i^2, \gamma_i = \mathbb{E}[X_i e_i]$. Matsushita and Otsu (2022) proposed this test statistic and showed that under minor technical conditions, including the assumption that $K \to \infty$ as $N \to \infty$, under $H_0 : \beta = \beta_0$ we have $LM^{1/2}(\beta_0) \Rightarrow N(0,1)$. The LM test rejects when $\left|LM^{1/2}(\beta_0)\right|$ is large (two-sided rejection).

As before, in order to implement this test one needs an estimator of Ψ. Similar to Crudu et al. (2021) in the case of the AR, Matsushita and Otsu (2022) suggested using the squared implied errors as proxies for variances under the null. Specifically, their proposed

estimator is

$$\widehat{\Psi}_1 = \frac{1}{K}\sum_i \widehat{\sigma}_i^2 (\sum_{j\neq i} P_{ij}X_j)^2 + \frac{1}{K}\sum_i \sum_{j\neq i} P_{ij}^2 \widehat{\gamma}_i \widehat{\gamma}_j,$$

where $\widehat{\sigma}_i^2 = e_i^2(\beta_0)$ and $\widehat{\gamma}_i = X_i e_i(\beta_0)$. Matsushita and Otsu (2022) showed that their estimator $\widehat{\Psi}_1$ is consistent for Ψ under the null ($H_0 : \beta = \beta_0$), and thus, the feasible LM with $\widehat{\Psi}_1$ has the correct asymptotic size. However, for similar reasons as in the case of the AR, the estimator $\widehat{\Psi}_1$ may lead to power losses under alternatives because $\widehat{\sigma}_i^2$ contains a large predictable part and overstates the variances σ_i^2.

In the current paper we propose a novel variance estimator $\widehat{\Psi}_2$ using ideas similar to Mikusheva and Sun (2022):

$$\widehat{\Psi}_2 = \frac{1}{K}\sum_i \frac{e_i M_i e}{M_{ii}} (\sum_{j\neq i} P_{ij}X_j)^2 + \frac{1}{K}\sum_i \sum_{j\neq i} \widetilde{P}_{ij}^2 X_i M_i e X_j M_j e.$$

Here we use $\widehat{\sigma}_i^2 = \frac{e_i M_i e}{M_{ii}}$, an unbiased proxy for σ_i^2, $\widehat{\gamma}_i = X_i M_i e$, a proxy for γ_i, and re-weighting $\widetilde{P}_{ij}^2 = \frac{P_{ij}^2}{M_{ii}M_{jj}+M_{ij}^2}$ to correct for correlation in proxies.

We also establish a new theoretical result showing that under both the null and local alternatives this estimator $\widehat{\Psi}_2$ is consistent under a general form of heteroskedasticity. For that we need the following assumptions.

Assumption 1

(i) P_Z *is an* $N \times N$ *projection matrix of rank* K, $K \to \infty$ *as* $N \to \infty$ *and there exists a constant* δ *such that* $\max_i P_{ii} \leq \delta < 1$;

(ii) Errors $\varepsilon_i = (e_i, v_i)', i = 1, ..., N$ *are independent with* $\mathbb{E}\varepsilon_i = 0, \max_i \mathbb{E}\|\varepsilon_i\|^6 < \infty$, *and for some positive constants* c^* *and* C^* *that do not depend on* N

$$c^* \leq \min_i \min_x \frac{x' Var(\varepsilon_i) x}{x'x} \leq \max_i \max_x \frac{x' Var(\varepsilon_i) x}{x'x} \leq C^*.$$

Theorem 1 *Let Assumption 1 hold, and* $\frac{\pi' Z' Z \pi}{K^{2/3}} \to 0$ *as* $N \to \infty$, *then*

(1) if $\beta = \beta_0$, *we have* $\frac{\widehat{\Psi}_2}{\Psi} \to^p 1$ *as* $N \to \infty$;

(2) if $\beta = \beta_0 + \Delta$ *such that* $\Delta \cdot \frac{\pi' Z' Z \pi}{K} \to 0$, *we have* $\frac{\widehat{\Psi}_2}{\Psi} \to^p 1$ *as* $N \to \infty$.

	FF	$\widetilde{F}$	JIVE-t-test	Robust AR	Robust LM
180 instruments	2.4	13.4	[0.066,0.132]	[0.008,0.201]	[0.067,0.135]
1530 instruments	1.3	6.2	[0.024,0.121]	[-0.047, 0.202]	[0.022,0.127]

Table 2: Robust and non-robust confidence sets in Angrist and Krueger (1991)

Notes: Results on pre-tests for weak identification and the confidence sets based on the JIVE t-test, the leave-one-out AR and the leave-one out LM for IV specification underlying Table VII Column (6) of Angrist and Krueger (1991). The confidence sets are constructed via analytical test inversion.

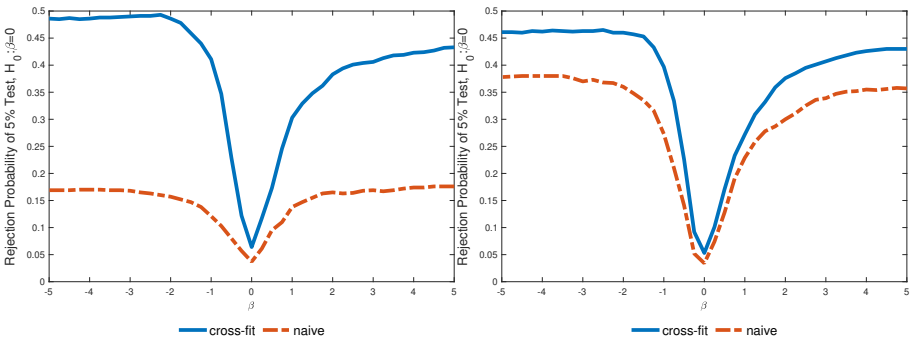
Part (1) of Theorem 1 gives consistency under the null hypothesis, and thus implies that the leave-one-out LM test using $\widehat{\Psi}_2$ has an asymptotically correct size. Part (2) addresses the consistency under local alternatives and guarantees that the power curves of the LM test using our proposed estimator $\widehat{\Psi}_2$ are the same as those of the infeasible LM test. Specifically, the power function of the infeasible leave-one-out LM test is described by the following convergence uniformly over the alternatives $\beta = \beta_0 + \Delta$:

$$LM^{1/2} \Rightarrow \Delta \frac{\mu^2}{\sqrt{K\Psi}} + \mathcal{N}(0,1).$$

Notice that Δ can be both positive or negative, thus, if one uses $LM^{1/2}(\beta_0)$ statistics then they can employ the standard gaussian critical values with two-sided rejection. Alternatively, one may use the squared statistic and the 95-percentile of χ^2_1 distribution. Another observation is that the proposed LM test is consistent for fixed alternatives as soon as $\mu^2/\sqrt{K} \to \infty$.

Small-scale simulation (continued). We repeat the same simulation design as before, but calculate the power curves for the leave-one-out LM test using $\widehat{\Psi}_1$ (red) and $\widehat{\Psi}_2$ (blue) estimates of the scale. They are reported in Figure 2. Here the power loss due to usage of the naive estimate of the scale $\widehat{\Psi}_1$ is extremely pronounced, especially in the sparse design. Another observation is that in both designs, contrary to the conjecture that LM is more efficient than AR, the leave-one-out AR test has higher power than the leave-one-out LM. This is not a universal observation and we discuss their trade-offs in the next sub-section.

Empirical example (Angrist and Krueger, 1991). We return to the empirical example of Angrist and Krueger (1991) and report the robust confidence sets obtained by



(a) sparse, $\frac{\mu^2}{\sqrt{K}} = 2.5$ (b) dense, $\frac{\mu^2}{\sqrt{K}} = 2.5$

Figure 2: Power curves for the leave-one-out LM with $\widehat{\Psi}_1$ (red dash) and $\widehat{\Psi}_2$ (blue line) variance estimators under sparse vs. dense first stage. Instruments are $K = 40$ balanced group indicators, $N = 200$.

inverting the leave-one-out AR and LM tests robust to many weak instruments in Table 2. We note that all confidence sets are finite intervals and are somewhat informative, though the leave-one-out AR confidence is significantly wider than the other two. This is mostly due to the fact we established before that the identification seems to be strong in Angrist and Krueger (1991). We may also notice that the leave-one-out LM confidence set is nearly identical to the JIVE-Wald confidence set. This is a reflection of the fact that the leave-one-out LM test is asymptotically equivalent to the JIVE t-test under strong identification. Unlike the JIVE t-test, which fails to controls size under weak identification, the leave-one-out LM test is fully robust to weak identification. Thus, we recommend using the leave-one-out LM test as a default option without a pre-test for weak identification. Another observation is that the case with 1,530 instruments is less informative and produces wider confidence sets for all three test statistics than when only 180 instruments are used.

4.3 Combination tests

In the empirical example of Angrist and Krueger (1991) we have seen that the leave-one-out LM test produces much shorter confidence sets than the robust leave-one-out AR test. At the same time our simulation exercise showed the opposite ordering of power. This raises the question of power comparison between these two tests. The answer is that neither of the two tests dominates the other. As mentioned before, the power curves for the infeasible tests (or for the feasible tests with our proposed estimators $\widehat{\Phi}_2$ or $\widehat{\Psi}_2$

for the normalizing factors) under the alternative $\beta = \beta_0 + \Delta$ and when $\frac{\mu^2}{K} \to 0$ can be characterized by:

$$LM^{1/2} \Rightarrow \Delta \frac{\mu^2}{\sqrt{K\Psi}} + \mathcal{N}(0,1),$$
$$AR \Rightarrow \Delta^2 \frac{\mu^2}{\sqrt{K\Phi}} + \mathcal{N}(0,1).$$

These power curves imply that when a setting is strongly identified in the sense of Mikusheva and Sun (2022), that is, when $\frac{\mu^2}{\sqrt{K}} \to \infty$, then both the leave-one-out AR and the leave-one-out LM are asymptotically consistent for fixed alternatives β. Under strong identification the two tests have different sets of local alternatives, namely, alternatives with asymptotically non-trivial probability of detection. Specifically, for the leave-one-out AR test the set of local alternatives is $\{\Delta : \frac{\Delta^2\mu^2}{\sqrt{K}} = C\}$ i.e., $|\Delta| \propto \sqrt{\frac{\sqrt{K}}{\mu^2}}$, while for the leave-one-out LM test it is $\{\Delta : \frac{|\Delta|\mu^2}{\sqrt{K}} = C\}$ i.e., $|\Delta| \propto \frac{\sqrt{K}}{\mu^2}$. So, we observe that under strong identification the leave-one-out AR test has a slower speed of detection than the leave-one-out LM. This suggests that when identification is strong the leave-one-out LM will tend to produce shorter confidence sets.

However, when identification is weak ($\frac{\mu^2}{\sqrt{K}}$ is bounded), there is no consistency for fixed alternatives, and the confidence sets tend to be non-shrinking. In such settings, the leave-one-out AR will tend to have higher power for distant alternatives and would be preferable to use.

These considerations suggest that one may want to combine the robust test statistics in some optimal way based on the strength of identification. For example, one may do a switch or "soft switch" between the leave-one-out AR and the leave-one-out LM statistics depending on the size of $\widetilde{F}$. In cases with small number of instruments, Andrews (2016) found the conditional likelihood ratio (CLR) test of Moreira (2003) has very good power properties because it can be considered as a robust test that implements a soft switch. A recent paper by Ayyar et al. (2022) suggests how to construct a weak instrument robust analog of CLR test when the number of instruments is large. A recent paper by Lim et al. (2022) searches for an optimal linear combination test that optimizes a minimax criterion that was first proposed in Andrews (2016). The resulting test statistic is a weighted

average of the leave-one-out AR and the leave-one-out LM.

5 Allowing for many exogenous regressors

To illustrate the issue arising from many instruments, previously we simplified the exposition by assuming no exogenous regressors (controls) in the structural equation. However, this assumption is unrealistic in practice. In models with many instruments, it is typical to also have many exogenous regressors. For example, in settings like Angrist and Krueger (1991) once we use interactions of baseline instruments with covariates as instruments, it is common to include those covariates as exogenous regressors as well. More broadly, in practice many instrumental variables are only valid after conditioning on additional covariates. "Saturated" specifications that control for covariates nonparametrically can ensure proper interpretation of the IV estimate under the LATE framework of Imbens and Angrist (1994). A "saturated" specification includes rich interactions between the instrument and the covariates, giving rise to many instruments and many exogenous regressors, as recently discussed in Słoczyński (2020) and Blandhol et al. (2022).

To analyze the impact of many exogenous regressors, we consider the following model:

$$\begin{cases} Y_i = \beta X_i + \gamma' W_i + e_i, \\ X_i = \pi' Z_i + \delta' W_i + v_i, \end{cases} \tag{3}$$

for $i = 1, ..., N$. This is a linear IV regression with a scalar outcome Y_i, an endogenous scalar regressor X_i, a $K_Z \times 1$ vector of instrumental variables Z_i and a $K_W \times 1$ vector of exogenous regressors W_i. The main assumption is $\mathbb{E}[e_i|Z_i, W_i] = \mathbb{E}[v_i|Z_i, W_i] = 0$. We are interested in estimation of and statistical inference about β, and treat parameters π, γ, δ as nuisance parameters.

New challenges of estimation with many instruments and many exogenous regressors. The TSLS estimator of β in this case is equivalent to the TSLS in a model that first partials out W. Let us introduce a projection matrix $M_W = I - W(W'W)^{-1}W'$. Denote $Y^\perp = M_W Y$, $X^\perp = M_W X$ and $Z^\perp = M_W Z$ to be the outcome variable, en-

dogenous regressor and instruments with the exogenous regressors partialled out. Finally denote $P^\perp = P_{Z^\perp} = Z^\perp \left((Z^\perp)'Z^\perp\right)^{-1}(Z^\perp)'$ to be the projection matrix based on the residualized instruments. Then the TSLS estimator of β is

$$\widehat{\beta}_{TSLS} = \frac{(X^\perp)'P^\perp Y^\perp}{(X^\perp)'P^\perp X^\perp}.$$

As we have seen in the case without exogenous regressors, the TSLS estimator is very biased when K_Z is large. We should expect a similar issue in the case with many instruments and many exogenous regressors. One solution in the case without exogenous regressors is to remove a diagonal from the projection matrix, so, a natural though a naive approach is the following estimator

$$\widehat{\beta}_1 = \frac{\sum_{i\neq j} X_i^\perp P_{ij}^\perp Y_j^\perp}{\sum_{i\neq j} X_i^\perp P_{ij}^\perp X_j^\perp}.$$

Here we first partial out exogenous regressors, and then use a JIVE- type estimator without exogenous regressors that removes a diagonal from the projection on the instruments. One can show that this approach does not quite work and $\widehat{\beta}_1$ tends to have large biases. Indeed,

$$\widehat{\beta}_1 - \beta = \frac{\sum_{i\neq j} X_i^\perp P_{ij}^\perp e_j^\perp}{\sum_{i\neq j} X_i^\perp P_{ij}^\perp X_j^\perp}, \text{ where } e^\perp = M_W e.$$

We can show that the numerator has a non-trivial mean. Let us denote $M_{W,ij}$ as elements of M_W and use an observation that $M_W P^\perp = P^\perp M_W = P^\perp$:

$$\mathbb{E}\sum_{i\neq j} X_i^\perp P_{ij}^\perp e_j^\perp = \sum_k \sum_{i\neq j} \mathbb{E}[v_k e_k] M_{W,ki} P_{ij}^\perp M_{W,kj} = \sum_k \mathbb{E}[v_k e_k] P_{kk}^\perp (1 - M_{W,kk}).$$

The last expression is non-trivial, since on average $1 - M_{W,kk}$ is $\frac{K_W}{N}$, while the trace of $P^\perp$ is K_Z. Thus, under homoskedasticity if the values of $P_{kk}^\perp$ are uncorrelated with $M_{W,kk}$, we should expect the last expression to be $\sigma^2 \frac{K_W K_Z}{N}$. Removing the diagonal from $P^\perp$ has not solved the many instruments issue here, because the procedure of partialling out introduces the dependence across residuals so that $e_i^\perp$ are not independent across i.

Another suggestion is to write the numerator of the TSLS as $(X^\perp)'P^\perp Y^\perp = X'P^\perp Y$,

which is due to $M_W P^{\perp} M_W = P^{\perp}$, and to remove the diagonal from $P^{\perp}$. That is, one may propose the following estimator:

$$\widehat{\beta}_2 = \frac{\sum_{i \neq j} X_i P^{\perp}_{ij} Y_j}{\sum_{i \neq j} X_i P^{\perp}_{ij} X_j}.$$

This suggestion does not work either but for a different reason. The operator $P^{\perp}$ has a property that it projects out the exogenous regressors, namely $P^{\perp} W = 0$, but the same operator without the diagonal does not have this property: $\sum_{j \neq i} P^{\perp}_{ij} W_j' \neq 0$. Thus

$$\widehat{\beta}_2 - \beta = \frac{\sum_{i \neq j} X_i P^{\perp}_{ij} W_j' \gamma + \sum_{i \neq j} X_i P^{\perp}_{ij} e_j}{\sum_{i \neq j} X_i P^{\perp}_{ij} X_j}.$$

The term in the numerator $\sum_{i \neq j} X_i P^{\perp}_{ij} W_j' \gamma = -\sum_i X_i P^{\perp}_{ii} W_i' \gamma$ corresponds to a bias. Since the average value of $P^{\perp}_{ii}$ is $\frac{K_Z}{N}$, this term is (approximately) the fraction of the bias that would arise if one does not include exogenous regressors in the regression (omitted variable bias).

To summarize, the challenges from both many instruments and many exogenous regressors are a counterplay of two needs: the need to partial out exogenous regressors and the need to remove the diagonal. An ideal estimator takes the form of

$$\widehat{\beta}_3 = \frac{X' A Y}{X' A X},$$

where a $N \times N$ matrix A has the following properties: (i) $AW = 0$ (partialing out property) and (ii) $A_{ii} = 0$ for all i (zero diagonal property). Matrix A can be constructed using Z and W, with some preferences for a matrix close to $P^{\perp}$.

A recent paper by Chao et al. (2023) suggests a matrix A of the following form $A = M_W(P^{\perp} - D_\theta)M_W$ where D_θ is a diagonal matrix with diagonal elements $\theta_1, ..., \theta_N$ selected in such a way that A has zero diagonal. Chao et al. (2023) showed that such θ_i's can be found in a well-balanced design (when $\min_i M_{W,ii} > 1/2$) and provided a proof of consistency and asymptotic gaussianity of estimator $\widehat{\beta}_3$ under some assumptions.

Robust AR. Following the ideas stated in the previous sections, we create a test for $H_0 : \beta = \beta_0$ robust to many instruments/exogenous regressors using an AR-type statistic

$$AR_W(\beta_0) = \frac{1}{\sqrt{K_Z \widehat{\Phi}_W}}(Y - \beta_0 X)' A (Y - \beta_0 X),$$

where we propose a novel estimator for the normalization factor

$$\widehat{\Phi}_W = \frac{2}{K} \sum_{i,j} \frac{A_{ij}^2}{M_{ZW,ii} M_{ZW,jj} + M_{ZW,ij}^2} \widehat{\sigma}_i^2 \widehat{\sigma}_j^2. \tag{4}$$

Here $\widehat{\sigma}_i^2 = \sum_k M_{ZW,ik}(Y_i - \beta_0 X_i)(Y_k - \beta_0 X_k)$,where $M_{ZW,ij}$ are elements of M_{ZW}, a projection matrix orthogonal to both Z and W. This proposal is similar to the $\widehat{\Phi}_2$ estimator in the case with no exogenous regressors. There is no consistent "naive" variance estimator (like $\widehat{\Phi}_1$) in the presence of exogenous regressors since under the null $Y_i - \beta_0 X_i$ is not equal to the error but rather contains the predictable non-zero part $(\gamma - \beta_0 \delta)' W_i$. This term squared overstates the true variance σ_i^2 drastically. The ideas of cross-fit variance estimation are very useful here. We propose a test that rejects the null when $AR_W(\beta_0)$ exceeds the right α- quantile of the standard normal distribution. We show this test has the correct size.

Assumption 2

(i) Projection matrices M_W and $P^\perp$ are such that $\min_i M_{W,ii} > 1/2$, all components of vector $\theta = (M_W \circ M_W)^{-1} diag(P^\perp)$ are non-negative; there exists a constant δ such that $\frac{P_{ii}^\perp}{M_{W,ii}^2} \leq \delta < 1$, and the rank K_Z of projection matrix $P^\perp$ grows to infinity when $N \to \infty$;

(ii) Errors $e_i, i = 1, ..., N$ are independent with $\mathbb{E} e_i = 0, \max_i \mathbb{E}\|e_i\|^6 < \infty$, and for some positive constants c^* and C^* that do not depend on N

$$c^* \leq \min_i \mathbb{E}\|e_i\|^6 \leq \max_i \mathbb{E}\|e_i\|^6 \leq C^*.$$

Assumption 2 (i) can be characterized as an assumption about a balanced design. In a case with no exogenous regressors ($M_W = I$) this assumption is equivalent to Assumption 1 (i).

Theorem 2 *Let Assumption 2 hold in model (3), then under the true null hypothesis* $H_0 : \beta = \beta_0$ *as* $N \to \infty$ *we have*

$$AR_W(\beta_0) \Rightarrow N(0,1).$$

Simulation study. In order to assess the size property of the newly proposed robust test and to compare it with naive approaches to deal with both many instruments and many exogenous regressors in a realistic setting we calibrate the simulation to Gilchrist and Sands (2016).

Gilchrist and Sands (2016) are interested in estimating social spillovers from movie viewership, namely, the effect of viewership from a movie's opening weekend on subsequent viewership. To identify the causal effect they use weather during the opening weekend as set of exogenous instruments. The setting contains both a large number of instruments and a large number of exogenous regressors. The set of instruments includes 52 different measures of weather conditions around a movie theater, including temperature, indicators for snow/rain, precipitation etc. The set of exogenous regressors includes indicators for calendar year, day of the week, week of the year, holidays, as well as weather conditions in periods for which subsequent viewership is measured. The number of exogenous regressors is relatively high in comparison to the dimension of the instruments, which provides an empirically relevant setting for showing that the original leave-one-out AR test might not be robust to many exogenous regressors, and the adjustment proposed in this paper is able to restore the correct size.

In order to calibrate the simulation to Gilchrist and Sands (2016), we follow the simulation design proposed in Angrist and Frandsen (2022). Specifically, we take the LIML model as the ground truth. Let $\widehat{y}(W_i)$ be the linear function in W_i equal to the dependent variable fitted value, after subtracting $\widehat{\beta}_{LIML} X_i$. We set π to be the first stage coefficients. We simulate the data from the model:

$$\widetilde{Y}_i = \widehat{y}(W_i) + \beta \widetilde{X}_i + \omega_i(\epsilon_i - 1.5 v_i),$$
$$\widetilde{X}_i = \pi' Z_i^{\perp} + v_i.$$

Here $\beta = 0.6$, weights ω_i are the absolute values of the LIML residuals to mimic the heteroskedasticity of the data. We perform 1,000 simulations, drawing (v_i, ϵ_i) independently from the standard normal distribution. After removing multi-collinearities, we have $K_Z = 48$ and $K_W = 119$ for a sample size $N = 1,669$.

We calculate the simulated size for our proposed AR test robust to many instruments and many exogenous regressors by comparing statistics $AR_W(\beta_0)$ with the upper 95% quantile of the standard normal distribution. We also check two naive approaches to testing using the AR test by calculating statistics

$$AR_1(\beta_0) = \frac{1}{\sqrt{K_Z \widehat{\Phi}_1}} \sum_{i \neq j} P_{ij}^{\perp} (Y_i^{\perp} - \beta_0 X_i^{\perp})'(Y_j^{\perp} - \beta_0 X_j^{\perp}),$$
$$AR_2(\beta_0) = \frac{1}{\sqrt{K_Z \widehat{\Phi}_2}} \sum_{i \neq j} P_{ij}^{\perp} (Y_i - \beta_0 X_i)'(Y_j - \beta_0 X_j),$$

where $\widehat{\Phi}_i$ are properly constructed estimators of the normalization factor using the cross-fit ideas stated in this paper. The results are reported in Table 3.

N	K_Z	K_W	size of AR_1	size of AR_2	size of AR_W
1,669	48	119	11%	1.3%	5%

Table 3: Simulation results for size of different modifications of the AR tests with many instruments and many exogenous regressors. Simulation design mimics data from Gilchrist and Sands (2016).

The results show that if the test statistics fail at any of the two tasks, either at removing the diagonal (as AR_1 does) or at partialling out the exogenous regressors (as AR_2) then the size may differ from the declared level. At the same time a properly constructed statistics that by construction performs both tasks, paired with the proper estimator of the normalization factor, controls size effectively even with a large number of both instruments and exogenous regressors.

6 Conclusion and open questions

The goal of this paper is to provide an overview for the statistical challenges surrounding estimation and inferences in a linear IV model with many instruments. We show that aside from the obvious benefits of bringing additional identifying information, many instruments come at a cost as one typically needs to estimate the optimal way to combine many instruments. If one has many not very informative instruments, then the uncertainty surrounding the first stage estimation may produce significant biases of the TSLS, and even lead to an inconsistency.

We showcased one set of methods and ideas that allowed reliable estimation and inferences. One of the central ideas, jack-knifing or deleting a diagonal, produces both new estimators with superior convergence properties and new identification robust tests.

We presented results established in the econometric literature that inform a coherent empirical strategy. Specifically, one may use a pre-test for weak identification robust to heteroscedasticity presented in Section 3, and depending on its results either use the JIV estimator based on the idea of removing the diagonal paired with its standard errors, or use any of the identification robust tests presented in Section 4. This paper also establishes some new results including a version of the LM test robust to many weak instruments with a new variance estimator and a modification of the AR test robust to both many instruments and many exogenous regressors.

As a final word we wish to mention several open questions in this research area that have a chance to be solved within the next few years and we hope this encourages some researchers looking for next project to tackle them.

The first open question is related to the observation that a pre-test for weak identification is tightly related to an estimator one hopes to use and is formulated as whether one can trust a specific estimator with a confidence set or a test based on it. The current test for weak identification is created for the JIVE. However, there are results pointing out that other estimators like the JIVE-LIML (or its heteroskedasticity-robust version) are more efficient than the JIVE under strong identification (Hausman et al. (2012)). It is still an open question to establish a pre-test for reliability of the heteroskedasticity-robust JIVE-LIML. Along the same line of thoughts, currently there is no pre-test for

any estimator that accommodates not just many instruments but also many exogenous regressors, though empirically there is an important need for such a pre-test.

This paper discussed in detail one approach based on jack-knifing or diagonal removal. There are other influential ideas mentioned in Section 2.1 related to instrument selection or construction of the optimal instrument using Machine Learning (ML) approaches. Those ideas seem very powerful and produce quite efficient estimators if the first stage can be well described by a model in which the selected ML technique produces a consistent estimator of the optimal instrument. Unfortunately, the performance of an ML first stage is generally unknown if the first stage does not satisfy the assumptions of a model needed for ML consistency. Based on simulation studies from Angrist and Frandsen (2022) we are pessimistic that the aforementioned techniques work well under many weak instruments. As has been shown in Mikusheva (2022) the conditions needed for consistency of an IV estimator using LASSO selection on the first stage depends in a significant way on the true sparsity of the true first stage model. It has also been suggested in Mikusheva (2022) that using sample-split and cross-fit may be a powerful idea for breaking the dependence between two stages when ML techniques are used on the first stage. Unfortunately, currently there is no good technique to assess whether an IV estimator with some ML algorithm used in the first stage is reliable in any given data set. One technical challenge for developing such methods is understanding the asymptotic behavior of ML estimators in settings where modeling assumptions needed for consistency of an ML algorithm (like sparsity) do not hold.

It is worth pointing out that this paper as well as a vast majority of research papers devoted to many and/or weak instruments are written for cross-sectional settings, while there is a large number of empirical settings using macroeconomic or financial data that can be labeled as many weak instruments. Mikusheva (2022) showcased a very clear need to develop inferences robust to many weak instruments in time series settings as well as the associated challenges.

References

Anatolyev, S. and M. Sølvsten (2023). Testing Many Restrictions Under Heteroskedasticity. *Journal of Econometrics 236*(1), 105473.

Andrews, I. (2016). Conditional linear combination tests for weakly identified models. *Econometrica 84*, 2155–2182.

Angrist, J. and A. Krueger (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics 106*, 979–1014.

Angrist, J. and A. Krueger (1995). Split-sample instrumental variables estimates of the return to schooling. *Journal of Business & Economic Statistics 13*(2), 225–35.

Angrist, J. D. and B. Frandsen (2022). Machine labor. *Journal of Labor Economics 40*(S1), S97–S140.

Angrist, J. D., G. W. Imbens, and A. B. Krueger (1999, January). Jackknife instrumental variables estimation. *Journal of Applied Econometrics 14*(1), 57–67.

Ayyar, S., Y. Matsushita, and T. Otsu (2022). Conditional likelihood ratio test with many weak instruments.

Bekker, P. A. (1994). Alternative Approximations to the Distributions of Instrumental Variable Estimators. *Econometrica 62*(3), 657–681.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica 80*(6), 2369–2429.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies 81*(2 (287)), 608–650.

Bhuller, M., G. B. Dahl, K. V. LÃžken, and M. Mogstad (2020). Incarceration, Recidivism, and Employment. *Journal of Political Economy 128*(4), 1269–1324.

Blandhol, C., J. Bonney, M. Mogstad, and A. Torgovitsky (2022, January). When is tsls actually late? Working Paper 29709, National Bureau of Economic Research.

Carrasco, M. (2012). A regularization approach to the many instruments problem. *Journal of Econometrics 170*(2), 383–398.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics 34*(3), 305–334.

Chao, J. C. and N. R. Swanson (2005). Consistent Estimation with a Large Number of Weak Instruments. *Econometrica 73*(5), 1673–1692.

Chao, J. C., N. R. Swanson, J. A. Hausman, W. K. Newey, and T. Woutersen (2012). Asymptotic Distribution of JIVE in a Heteroskedastic IV Regression with Many Instruments. *Econometric Theory 28*(1), 42–86.

Chao, J. C., N. R. Swanson, and T. Woutersen (2023). Jackknife Estimation of a Cluster-Sample IV Regression Model with Many Weak Instruments.

Crudu, F., G. Mellace, and Z. Sándor (2021). Inference in Instrumental Variable Models with Heteroskedasticity and Many Instruments. *Econometric Theory 37*(2), 281–310. Publisher: Cambridge University Press.

Dobbie, W., J. Goldin, and C. S. Yang (2018, February). The effects of pretrial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *American Economic Review 108*(2), 201–40.

Donald, S. G. and W. K. Newey (2001). Choosing the Number of Instruments. *Econometrica 69*(5), 1161–1191.

Gilchrist, D. S. and E. G. Sands (2016). Something to talk about: Social spillovers in movie consumption. *Journal of Political Economy 124*(5), 1339–1382.

Han, C. and P. C. B. Phillips (2006). Gmm with many moment conditions. *Econometrica 74*(1), 147–192.

Hansen, C., J. Hausman, and W. Newey (2008). Estimation With Many Instrumental Variables. *Journal of Business & Economic Statistics 26*(4), 398–422.

Hansen, C. and D. Kozbur (2014). Instrumental variables estimation with many weak instruments using regularized jive. *Journal of Econometrics 182*(2), 290–308.

Hausman, J. A., W. K. Newey, T. Woutersen, J. C. Chao, and N. R. Swanson (2012). Instrumental variable estimation with heteroskedasticity and many instruments. *Quantitative Economics 3*(2), 211–255.

Imbens, G. and J. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*, 467–475.

Kline, P., R. Saggio, and M. Sølvsten (2020). Leave-out estimation of variance components. *Econometrica 88*(5), 1859–1898. arXiv: 1806.01494.

Lim, D., W. Wang, and Y. Zhang (2022). A conditional linear combination test with many weak instruments.

Maestas, N., K. J. Mullen, and A. Strand (2013, August). Does disability insurance receipt discourage work? using examiner assignment to estimate causal effects of ssdi receipt. *American Economic Review 103*(5), 1797–1829.

Matsushita, Y. and T. Otsu (2022). A jackknife lagrange multiplier test with many weak instruments. *Econometric Theory*, 1–24.

Mikusheva, A. (2022). Many weak instruments in time series econometrics. *Proceedings of the World Congress of Econometric Society*.

Mikusheva, A. and L. Sun (2022). Inference with Many Weak Instruments. *The Review of Economic Studies 89*(5), 2663–2686.

Moreira, M. (2003). A conditional likelihood ratio test for structural models. *Econometrica 71*, 1027–1048.

Newey, W. K. and J. R. Robins (2018). Cross-Fitting and Fast Remainder Rates for Semiparametric Estimation. *arXiv:1801.09138 [math, stat]*. arXiv: 1801.09138.

Newey, W. K. and F. Windmeijer (2009). Generalized method of moments with many weak moment conditions. *Econometrica 77*(3), 687–719.

Okui, R. (2011). Instrumental variable estimation in the presence of many moment conditions. *Journal of Econometrics 165*(1), 70–86. Moment Restriction-Based Econometric Methods.

Sampat, B. and H. L. Williams (2019, January). How do patents affect follow-on innovation? evidence from the human genome. *American Economic Review 109*(1), 203–36.

Słoczyński, T. (2020). When should we (not) interpret linear iv estimands as late?

Sølvsten, M. (2020). Robust estimation with many instruments. *Journal of Econometrics 214*(2), 495–512.

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

Stock, J. and M. Yogo (2005). *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, Chapter Testing for Weak Instruments in Linear IV Regression, pp. 80–108. Cambridge University Press.

Stock, J., M. Yogo, and J. Wright (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics 20*, 518–529.

Sun, L. (2023, June). manyweakiv.

A Appendix with Proofs

Let C be a universal constant (that may be different in different lines but does not depend on N or K).

Proof of Theorem 1. The assumptions laid out in this theorem are exactly the ones stated in Theorem 5 of Mikusheva and Sun (2022). The only difference is that since we focus on linear first stage that the residualization is complete and satisfies $\Pi' M \Pi = 0$. Part (1) of Theorem 1 follows from Lemma S3.1 (a) from the Supplementary Appendix to Mikusheva and Sun (2022), which establishes the consistency of the scale parameter $\widehat{\Psi}_2$ under the null hypothesis.

Part (2) addresses the consistency of the scale parameter for local alternatives

$$e(\beta_0) = e + \Delta v + \Delta \cdot Z\pi = \eta + \Delta \cdot Z\pi.$$

Notice that $Me(\beta_0) = M\eta$ as the predicted part partials out completely. Define

$$\tilde{\Psi}_2 = \frac{1}{K}\sum_i \frac{\eta_i M_i \eta}{M_{ii}} (\sum_{j \neq i} P_{ij} X_j)^2 + \frac{1}{K}\sum_i \sum_{j \neq i} \widetilde{P}_{ij}^2 X_i M_i \eta X_j M_j \eta.$$

Part(1) of Theorem 1 shows that $\tilde{\Psi}_2 / \Psi_2 \to^p 1$ as long as $\Delta \to 0$. Given the partialling out property we have:

$$\widehat{\Psi}_2 - \tilde{\Psi}_2 = \frac{1}{K}\sum_i \frac{\Delta Z_i \pi M_i \eta}{M_{ii}} (\sum_{j \neq i} P_{ij} X_j)^2.$$

We now apply part (d) of Lemma S3.2 from the Supplementary Appendix to Mikusheva and Sun (2022) noticing that $w_i = M_{ii} Z_i \pi$ in notations of that lemma:

$$\frac{1}{K}\sum_i \frac{\Delta Z_i \pi M_i \eta}{M_{ii}} (\sum_{j \neq i} P_{ij} X_j)^2 - \frac{2}{K}\sum_i \sum_{j \neq i} P_{ij}^2 \frac{\Delta (Z_i \pi)^2}{M_{ii}^2} \mathbb{E}[v_j \eta_j] \to^p 0.$$

Finally, given that $\mathbb{E}[v_j \eta_j]$ is bounded by the constant we get that the last expression is bounded by $\Delta \frac{\pi' Z' Z \pi}{K} \to 0$. $\square$

Lemma 1 *Let $A = P^\perp - M_W D_\theta M_W$ where $\theta_1, ..., \theta_N$ are selected in such a way that A has all zero elements on the diagonal. Specifically $\theta = (M_W \circ M_W)^{-1} diag(P^\perp)$. Then under Assumption 2 we have*

(i) $c < \frac{1}{K_Z} \sum_{i,j} A_{ij}^2 < C$;

(ii) $\frac{1}{K_Z^2}\sum_{i,j,k} A_{ij}^2 A_{ik}^2 \to 0$;

(iii) $\frac{1}{K_Z^2}\sum_{i,j} A_{ij}^4 \to 0$.

Proof of Lemma 1. First we notice that Assumption 2 implies that $\theta_i \leq \delta$. Indeed θ_i's are the solution to a system of linear equations $\sum_j M_{W,ij}^2 \theta_j = P_{ii}^{\perp}$ and specifically $\theta = (M_W \circ M_W)^{-1} diag(P^{\perp})$. Here matrix $M_W \circ M_W$ with elements $M_{W,ij}^2$ is diagonal-dominant as $\sum_{j\neq i} M_{W,ij}^2 = M_{W,ii} - M_{W,ii}^2 \leq M_{W,ii}^2$ under Assumption 2 and thus is invertible. Furthermore, the system of equations can be rewritten as $\theta_i = \frac{1}{M_{W,ii}^2}(P_{ii}^{\perp} - \sum_{j\neq i} \theta_j M_{W,ij}^2)$. Thus, if $\theta_i \geq 0$ for all i as prescribed by Assumption 2 then $\theta_i \leq \frac{P_{ii}^{\perp}}{M_{W,ii}^2} \leq \delta$.

Below we use projection property $M_W P^{\perp} = P^{\perp} M_W = P^{\perp}$ and $M_W^2 = M_W$ as well as the definition of matrix A: $A_{ij} = P_{ij}^{\perp} - \sum_k M_{W,ik}\theta_k M_{W,kj}$. For part (i) notice that

$$\begin{aligned}
\sum_{i,j} A_{ij}^2 =& \sum_i (P_{ij}^{\perp})^2 - 2\sum_{i,j,k} P_{ij}^{\perp} M_{W,ik}\theta_k M_{W,kj} + \sum_{i,j,k,n} M_{W,ik}\theta_k M_{W,kj} M_{W,in}\theta_n M_{W,nj} = \\
=& \sum_j P_{jj}^{\perp} - 2\sum_{j,k} P_{jk}^{\perp}\theta_k M_{W,kj} + \sum_{j,k,n} M_{W,jk} M_{W,kn} M_{W,nj}\theta_k\theta_n \\
=& K_Z - 2\sum_k P_{kk}^{\perp}\theta_k + \sum_{k,n} M_{W,kn}^2 \theta_k\theta_n \\
=& K_Z - 2\sum_k P_{kk}^{\perp}\theta_k + \sum_k P_{kk}^{\perp}\theta_k = K_Z - \sum_k P_{kk}^{\perp}\theta_k.
\end{aligned}$$

Given $0 \leq \theta_i \leq \delta$ we have $(1-\delta)K_Z \leq \sum_{i,j} A_{ij}^2 \leq K_Z$.

For (ii) notice that:

$$\begin{aligned}
\sum_i A_{ij}^2 =& P_{jj}^{\perp} - 2\sum_k P_{jk}^{\perp}\theta_k M_{W,kj} + \sum_{k,n} M_{W,jk} M_{W,kn} M_{W,nj}\theta_k\theta_n = \\
=& P_{jj}^{\perp} - 2\sum_k P_{jk}^{\perp}\theta_k M_{W,kj} + \sum_k M_{W,jk}\theta_k (P_{jk}^{\perp} - A_{jk}); \\
\sum_{i,j,k} A_{ij}^2 A_{ik}^2 =& \sum_i (P_{ii}^{\perp} - \sum_k P_{ik}^{\perp}\theta_k M_{W,ki} - \sum_k A_{ik}\theta_k M_{W,ki})^2.
\end{aligned}$$

The sum above has terms:

$$\sum_i (P_{ii}^{\perp})^2 \leq P_{ii}^{\perp} \leq K_Z;$$

$$\sum_i \left(\sum_k P_{ik}^{\perp} \theta_k M_{W,ki} \right)^2 \leq \sum_i \left(\sum_k (P_{ik}^{\perp})^2 \sum_k M_{W,ki}^2 \right) \max_k \theta_k^2 \leq C \sum_i P_{ii}^{\perp} \leq C K_Z;$$

$$\sum_{i,k} P_{ii}^{\perp} P_{ik}^{\perp} \theta_k M_{W,ki} \leq \sum_i P_{ii}^{\perp} \sqrt{\sum_k (P_{ik}^{\perp})^2 \sum_k M_{W,ki}^2} \max_k |\theta_k| \leq C K_Z;$$

$$\sum_i \left(\sum_k A_{ik} \theta_k M_{W,ki} \right)^2 \leq \sum_i \left(\sum_k A_{ik}^2 \sum_k M_{W,ki}^2 \right) \max_k \theta_k^2 \leq C \sum_{i,k} A_{ik}^2 \leq C K_Z;$$

$$\sum_{i,k} P_{ii}^{\perp} A_{ik} \theta_k M_{W,ki} \leq \sqrt{\sum_i (P_{ii}^{\perp})^2} \sqrt{\sum_i \left(\sum_k A_{ik} \theta_k M_{W,ki} \right)^2} \leq C K_Z;$$

$$\sum_i \left(\sum_k P_{ik}^{\perp} \theta_k M_{W,ki} \right) \left(\sum_k A_{ik} \theta_k M_{ki} \right) \leq \sqrt{\sum_i \left(\sum_k P_{ik}^{\perp} \theta_k M_{W,ki} \right)^2} \sqrt{\sum_i \left(\sum_k A_{ik} \theta_k M_{W,ki} \right)^2} \leq C K_Z.$$

Putting all terms together we get $\sum_{i,j,k} A_{ij}^2 A_{ik}^2 \leq C K_Z$.

For (iii) notice that

$$\sum_k |M_{W,ik} M_{W,jk}| \leq \sqrt{\sum_k M_{W,ik}^2 \sum_k M_{W,jk}^2} \leq \sqrt{M_{W,ii} M_{W,jj}} \leq 1.$$

Thus $|A_{ij}| = |P_{ij}^{\perp} - \sum_k M_{W,ik} M_{W,jk} \theta_k| \leq 1 + \max_k |\theta_k| \leq C$. This implies

$$\sum_{i,j} A_{ij}^4 \leq C^2 \sum_{i,j} A_{ij}^2 \leq C K_Z,$$

and (iii) holds. □

Proof of Theorem 2. Statements proved in Lemma 1 alone with Assumption 2 lead to the validity of all conditions of the Central Limit Theorem for quadratic forms stated in Corollary A2.8 in Sølvsten (2020). This implies that under the null ($H_0 : \beta = \beta_0$) we have:

$$\frac{1}{\sqrt{K_Z \Phi_W}} (Y - \beta_0 X)' A (Y - \beta_0 X) = \frac{1}{\sqrt{K_Z \Phi_W}} e' A e \Rightarrow N(0, 1),$$

when $N, K_Z \to \infty$ with $\Phi_W = \frac{2}{K}\sum_{i,j} A_{ij}^2 \sigma_i^2 \sigma_j^2$, and $\sigma_i = \mathbb{E}e_i^2$. What is left to prove is the consistency of $\widehat{\Phi}_W$, the estimator for Φ_W defined in equation (4). For this proof we follow closely the structure of the proof of Lemma 2 in Mikusheva and Sun (2022). Specifically, define $M_{ij} = M_{ZW,ij}$, $\widetilde{A}_{ij}^2 = \frac{A_{ij}^2}{M_{ii}M_{jj}+M_{ij}^2}$, and we want to show that

$$\frac{2}{K}\sum_{i,j} \widetilde{A}_{ij}^2 e_i M_i e e_j M_j e - \frac{2}{K}\sum_{i,j} A_{ij}^2 \mathbb{E}e_i^2 \mathbb{E}e_j^2 \to^p 0.$$

For this we first notice that

$$\mathbb{E}\frac{2}{K}\sum_{i,j} \widetilde{A}_{ij}^2 e_i M_i e e_j M_j e = \frac{2}{K}\sum_{i,j} A_{ij}^2 \mathbb{E}e_i^2 \mathbb{E}e_j^2,$$

define $\xi_{ij} = e_i M_i e e_j M_j e - \mathbb{E}[e_i M_i e e_j M_j e]$. Our goal is to show that $\frac{1}{K}\sum_{i,j} \widetilde{A}_{ij}^2 \xi_{ij} \to^p 0$. The covariance structure of ξ_{ij} including statements that $\max_{i,j} \mathbb{E}\xi_{ij}^2 < C$ and $\max_{i,j,k} |\mathbb{E}\xi_{ij}\xi_{ik}| < C$ are proven in Lemma 2 of Mikusheva and Sun (2022). Following the proof of Lemma 2 of Mikusheva and Sun (2022) the only conditions needed are $\sum_{i,j} A_{ij}^4 \leq CK_Z$ and $\sum_{i,k,j} A_{ij}^2 A_{ik}^2 \leq CK_Z$, both of which are proven in Lemma 1. $\square$