WORKING PAPERS

N° 1004

May 2021

“Identification-Robust Nonparametric Inference in a Linear IV Model”

Bertille Antoine and Pascal Lavergne

Toulouse School of Economics

Identification-Robust Nonparametric Inference in a Linear IV Model[*]

Bertille Antoine[†] and Pascal Lavergne[‡]

Updated May 17, 2021

Abstract

For a linear IV regression, we propose two new inference procedures on parameters of endogenous variables that are robust to any identification pattern, do not rely on a linear first-stage equation, and account for heteroskedasticity of unknown form. Building on Bierens (1982), we first propose an Integrated Conditional Moment (ICM) type statistic constructed by setting the parameters to the value under the null hypothesis. The ICM procedure tests at the same time the value of the coefficient and the specification of the model. We then adopt a conditionality principle to condition on a set of ICM statistics that informs on identification strength. Our two procedures uniformly control size irrespective of identification strength. They are powerful irrespective of the nonlinear form of the link between instruments and endogenous variables and are competitive with existing procedures in simulations and application.

Keywords: Weak Instruments, Hypothesis Testing, Semiparametric Model.
JEL Codes: C130, C120.

[*] Funding: Bertille Antoine acknowledges financial support from SSHRC. Pascal Lavergne acknowledges support from ERC-POEMH 33766 and ANR-17-EURE-0010 (Investissements d'Avenir program).

[†] Corresponding author. *Simon Fraser University. Email: bertille_antoine@sfu.ca.* Address correspondence: Department of Economics, 8888 University Drive, Burnaby, BC V5A1S6, CANADA.

[‡] *Toulouse School of Economics. Email: lavergnetse@gmail.com* Address correspondence: Toulouse School of Economics, 1 Esplanade de l'Université, 31080 Toulouse Cedex 06, FRANCE.

1 Introduction

We consider cross-section data observations and the linear model popular from microeconometrics

$$y_i = Y_{2i}'\beta + X_{1i}'\gamma + u_i \qquad \mathbb{E}\left(u_i|X_{1i}, X_{2i}\right) = 0 \quad i = 1, \ldots n\,, \tag{1.1}$$

where Y_2 are endogenous variables, X_1 are exogenous control variables, and X_2 are exogenous instrumental variables. We focus on inference on the parameter β of the endogenous variables. Over the last 30 years, it has become clear that standard asymptotic approximations may reflect poorly what is observed even for large samples when there is weak correlation between instrumental variables and endogenous explanatory variables. Alternative asymptotic frameworks have then been developed to account for potentially weak identification and tests have been proposed that deliver reliable inference about parameters of interest, see e.g. Staiger and Stock (1997), Stock and Wright (2000), Moreira (2003), Kleibergen (2002, 2005), Andrews and Cheng (2012), Andrews and Guggenberger (2019), Andrews (2016), and Andrews and Mikusheva (2016a,b). Surveys on weak identification issues include Stock et al. (2002), Dufour (2003), Hahn and Hausman (2003), and Andrews and Stock (2007). Existing inference procedures are robust to identification strength and uniformly control size, but rely on a parametric first-stage, and often on a linear projection of endogenous variables on instruments. We argue that this feature can artificially create a weak identification issue. If linear projection, or another parametric form, does not capture enough of the variation of the endogenous variable, tests have little power, and potentially no more than trivial one.

From an empirical perspective, Dieterle and Snell (2016) have documented significant nonlinearities in first-stage regression in several applied microeconomics papers. By comparing linear and quadratic first-stage specifications, they show that (second-stage) estimates of interest can be quite sensitive to the first-stage functional form. Since practitioners typically have little prior information on the form of the relation between endogenous variables and instruments, one may consider estimating the reduced form nonparametrically, e.g. using an increasing number of approximating series. However, nonparametrically estimated instruments cannot be relied upon under weak identification, see Jun and Pinkse (2012) and Mikusheva and Sun (2020). Indeed, if identification is not strong enough, the statistical variability of a nonparametric estimator will dominate the signal we aim to estimate. Hence, an inference procedure that leaves the functional form of the first stage equation unspecified, while being robust to identification strength should be extremely valuable for empirical analysis.

We propose two new inference procedures that are easy to implement, robust to any identification pattern, and do not rely on a linear projection in the first-stage equation. Our test statistics are constructed with practical convenience in mind, as well as their resemblance with standard statistics used in the presence of weak instruments. We build on the Integrated Conditional Moment (ICM) principle originally proposed by Bierens (1982). We first combine this principle with the Anderson and Rubin (1949) idea of setting the parameter value to the one under the null hypothesis $H_0 : \beta = \beta_0$. This yields a statistic that tests at the same time for the value of the parameter and the specification of the model. Second, we consider a quasi-likelihood ratio statistic and we adopt the conditionality principle used by Moreira (2003) to condition upon another ICM statistic (when Y_2 is univariate, or a set of ICM statistics when Y_2 is multivariate) that informs on the strength of (nonparametric) identification in the first-stage equation. The *Conditional* ICM (CICM) test does not test the whole specification of the model, but only whether β_0 is compatible with the data assuming the model is adequate. This is valuable in practice even if the linear IV model is misspecified but provides relevant information on average effects of endogenous variables. For both the ICM and CICM tests, asymptotic critical values can be simulated under heteroskedasticity of unknown form. We show that our tests control size uniformly and are thus robust to identification strength. Our tests are consistent in case of semi-strong identification, following the terminology of Andrews and Cheng (2012), and can have non-trivial power under weak identification. Since we remain agnostic on the first-stage functional relation between endogenous and instrumental variables, these properties are independent of its particular, potentially nonlinear, form.

Our conditional ICM test is related to Andrews and Mikusheva (2016a) since it is conditional upon a functional nuisance parameter. A key difference is that we consider conditional moment restrictions while they focus on unconditional ones. Other work that considers conditional moments or an increasing number of unconditional ones, thus yielding "many instruments", includes Han and Phillips (2006), Hansen et al. (2008), Newey and Windmeijer (2009), Jun and Pinkse (2012), Hausman et al. (2012), and Mikusheva and Sun (2020). Some procedures are optimal under strong identification, but only allow for some semi-strong identification. Unlike these authors, we cannot claim that our procedures are optimal when identification is strong. For this reason, we do not address the optimality of our procedures in terms of weighted average power, see Chernozhukov et al. (2009) and Montiel Olea (2020). There is thus a price to be paid for identification robustness using our procedures, but we believe that if a practitioner is actually worried about weak identification, optimality under strong identification would likely not be of primary concern. Our tests are valid irrespective of identification strength and do not necessitate the choice of the number of instruments.

We found that the level of our tests is well controlled in a series of simulations. Our tests have significant power advantage compared to existing tests when the reduced form equation is nonlinear. They also have good power for a linear reduced form, though they cannot be more powerful than the conditional likelihood ratio test, which is nearly optimal, see Andrews et al. (2006) and Andrews et al. (2019). In an empirical application on the effects of population decline in Mexico on land concentration in the sixteenth century, using the data and framework of Sellars and Alix-Garcia (2018), our procedures provide sensible and empirically valuable inference.

Our paper is organized as follows. In Section 2, we introduce our framework, we recall the main existing procedures for inference under possibly weak identification, and we motivate our new tests from a power perspective. In Section 3, we recall the ICM principle and we describe our two procedures, namely the ICM test and the conditional ICM test. Here and in what follows, we do not formally address subvector inference - though it is always possible to adopt a projection approach, see Dufour (1997) and Dufour and Taamouti (2005). In Section 4, we discuss critical values and the properties of our test in a Gaussian setup. In Section 5, we show that our procedures are generally asymptotically valid with heteroskedasticity of unknown form. We prove uniform asymptotic validity and study uniform power under strong and weak identification. In Section 6, we study the small sample performance of our tests through Monte-Carlo simulations and compare it to previous proposals. In Section 7, we present the results of our empirical application. Proofs are gathered in Section 8.

2 Review and Motivation

We are interested in inference on the parameter β of the l endogenous variables Y_2 in (1.1) and thus in testing null hypotheses of the form $H_0 : \beta = \beta_0$. The influence of exogenous control variables X_1 can be projected out through orthogonal projection in (1.1), which does not influence our reasoning, but simplifies exposition. Hence, in what follows, we consider a structural equation of the form

$$y_i = Y_{2i}'\beta + u_i \qquad \mathbb{E}\left(u_i|Z_i\right) = 0 \quad i = 1, \ldots n \,. \tag{2.2}$$

This is augmented by a first-stage reduced form equation for Y_2

$$Y_{2i} = \Pi(Z_i) + V_{2i} \qquad \mathbb{E}\left(V_{2i}|Z_i\right) = 0\,. \tag{2.3}$$

The exogenous Z, of dimension k, are the instrumental variables for Y_2. For simplicity of exposition, we assume in this section homoskedasticity of the error terms $(u_i, V_{2i}')'$.

The identification strength (of β) is introduced by letting the function $\Pi(\cdot)$ depend on the sample size n. For instance, one could assume that $\Pi_n(Z) = \frac{C(Z)}{r_n}$, where $C(\cdot)$ is a fixed function and r_n a diverging sequence. Extending the terminology of Andrews and Cheng (2012), we will talk about weak identification when $n\,\mathbb{E}\,\Pi_n(Z)\Pi_n'(Z)$ converges to a bounded positive definite matrix, and semi-strong identification when $r_n^2\,\mathbb{E}\,\Pi_n(Z)\Pi_n'(Z)$ does for some $r_n^2 = o(n)$.

2.1 Linear First Stage

In most work, the first-stage (2.3) is modelled through a linear projection of the form $Z\pi$, where Z is the $n \times k$ matrix of observations of the instrumental variables. The concentration parameter, defined as

$$\mu^2 = \frac{\pi' Z' Z \pi}{\sigma^2_{V_2}}$$

when Y_2 is scalar, is a unitless measure of the strength of the instruments. Under weak identification, i.e. when $\pi = n^{-1/2}C$, μ^2 converges to a finite limit and no test for β is consistent. If μ^2 diverges, then β can be consistently estimated, a situation that arises both under semi-strong identification (when $\pi = C/r_n$ with $r_n \to \infty$ and $n^{1/2}r_n^{-1} \to \infty$), and under strong identification (when $r_n = 1$).

Well-known inference procedures are constructed to be robust to identification strength. To test the null hypothesis $H_0 : \beta = \beta_0$, the statistic of Anderson and Rubin (1949) evaluates the orthogonality of $(y - Y_2'\beta_0)$ and Z and writes

$$\text{AR} = \frac{b_0' Y' P_Z Y b_0}{b_0' \widehat{\Omega} b_0}\,.$$

Here $b_0 = (1, -\beta_0')'$,

$$Y = \begin{bmatrix} y_1 & Y_{21}' \\ \vdots & \vdots \\ y_n & Y_{2n}' \end{bmatrix},$$

so that Yb_0 is the vector of generic components $y_i - Y_{2i}'\beta_0 = u_i$ under H_0, P_Z is the orthogonal projection on the space spanned by the columns of Z, and $\widehat{\Omega} = (n-k)^{-1} Y'(\mathbf{I} - P_Z)Y$ is an estimator of the errors' variance Ω under the assumption of homoskedasticity. Under linearity, one can rewrite the structural equation as

$$y_i - Y_{2i}'\beta_0 = Z_i'\Delta + \varepsilon_i, \qquad \text{where} \quad \Delta = \pi\left(\beta - \beta_0\right) \quad \text{and} \quad \varepsilon_i = u_i + V_{2i}\left(\beta - \beta_0\right).$$

The AR statistic is (up to a scale) the F statistic for the null hypothesis $\Delta = 0$. It tests at the same time H_0 and the correct specification of the model. The K test of Kleibergen (2005) is derived as a score test of H_0 under the assumptions of joint normality of u and V_2. The Conditional Likelihood Ratio (CLR) test is based on

$$\mathrm{LR} = \frac{b_0'Y'P_ZYb_0}{b_0'\widehat{\Omega}b_0} - \min_b \frac{b'Y'P_ZYb}{b'\widehat{\Omega}b},$$

which is derived as an approximate likelihood ratio test statistic for H_0 in the normal case by Moreira (2003). Unlike AR, it tests only whether $\beta = \beta_0$ irrespective of the validity of the linear IV model.

Under weak identification, the above test statistics can be used to obtain valid inference, and the tests have been shown to control size uniformly, see our references in the Introduction. Dufour and Taamouti (2007) further study the size robustness of such procedures to omitted relevant instruments and show that the AR procedure is particularly well behaved in this respect. Here we focus instead on the power of inference procedures with omitted instruments. Assuming a linear reduced-form for Y_2 is not restrictive as a linear approximation of the regression of Y_2 on the instruments. However, a linear approximation can yield little power for the tests.

As an example, assume Y_2 is scalar, $Z \sim N(0,1)$, and

$$\Pi(Z) = \frac{1}{r_n}(3Z - Z^3) + \frac{1}{\sqrt{n}}(Z^2 - 1), \quad r_n \geq 1.$$

If one approximates the unknown function $\Pi(\cdot)$ by a linear form, then $\min_{\pi_1} \mathbb{E}\left(\pi_1 Z - \Pi(Z)\right)^2$ yields the first-order condition

$$\mathbb{E}\left[Z\left(\pi_1 Z - \frac{1}{r_n}(3Z - Z^3) - \frac{1}{\sqrt{n}}(Z^2 - 1)\right)\right] = 0,$$

and the solution $\pi_1 = 0$.[1] Hence relying on a linear approximation may yield no more than trivial power for the above standard tests.

[1] If an intercept was included, it would be zero, so we dispense with it.

We may want to allow for a nonlinear form of the first-stage equation. The power of the tests, and then inference on parameters, will be affected by the accuracy of the chosen functional form. In our example, if one approximates the unknown function $\Pi(\cdot)$ by a quadratic form, then $\min_{\pi_1,\pi_2} \mathbb{E}\,\left(\pi_1 Z + \pi_2(Z^2-1) - \Pi(Z)\right)^2$ yields

$$\mathbb{E}\left[Z\left(\pi_1 Z + \pi_2(Z^2-1) - \frac{1}{r_n}(3Z - Z^3) - \frac{1}{\sqrt{n}}(Z^2-1)\right)\right] = 0$$
$$\mathbb{E}\left[(Z^2-1)\left(\pi_1 Z + \pi_2(Z^2-1) - \frac{1}{r_n}(3Z - Z^3) - \frac{1}{\sqrt{n}}(Z^2-1)\right)\right] = 0\,.$$

The solutions are $\pi_1 = 0$ and $\pi_2 = \frac{1}{\sqrt{n}}$. Thus, even if the relation between Y_2 and the instrument Z is not weak in the sense that $r_n << \sqrt{n}$, or is even strong with $r_n = 1$, the quadratic approximation will only pick up the weakest quadratic part. Hence, an inadequate functional form may artificially create a weak identification issue.[2]

2.2 Many Instruments

We may consider estimating the first-stage (2.3) nonparametrically by increasing the number of approximating polynomial or series terms with the sample size. We would then consider a linear approximation $\tilde{Z}\pi$, where $\tilde{Z}$ is a $n \times k_n$ matrix of series terms in the variables in Z. Work on "many weak" instruments, see our Introduction for references, relates the rate of increase on the number of instruments k_n to the unknown identification strength. There does not seem to exist any adaptive data-driven method to select k_n. One may consider a specification search to select k_n and then the best functional form of the reduced-form equation. However, specification tests may suffer from low power in case of weak identification, and, in addition, one would need to account for pre-testing in inference on parameters of interest.

The choice of k_n will affect power. To see this, consider the case of normal homoskedastic errors with known variance Ω and a scalar Y_2. When testing $H_0:\ \beta = \beta_0$, with β_0 the true value of the parameter,

$$\text{AR}(\beta_0) = \frac{b_0' Y' P_Z Y b_0}{b_0' \Omega b_0} \sim \chi^2_{k_n}\,,$$

[2] One can construct more involved examples where the same phenomenon shows up. For instance, if $\Pi(Z) = \frac{1}{r_n}(Z^5 - 10Z^3 + 15Z) + \frac{1}{\sqrt{n}}(Z^4 - 6Z^2 + 3)$, then the best cubic approximation is identically zero and the best quartic approximation only picks up a $\frac{1}{\sqrt{n}}$ component.

conditionally on the Z_i. When testing $H_0: \ \beta = \beta_1$, where $\beta_1 \neq \beta_0$,

$$\mathrm{AR}\left(\beta_1\right) = \frac{b_1' Y' P_Z Y b_1}{b_1' \Omega b_1} \sim \chi^2_{k_n}\left(\left(\beta_0 - \beta_1\right)^2 \frac{\Pi' P_Z \Pi}{b_1' \Omega b_1}\right) ,$$

conditionally on the Z_i, where Π is the vector with generic element $\Pi(Z_i)$. The test is consistent only if the non-centrality parameter dominates the variability of $\mathrm{AR}\left(\beta_0\right)$, as measured by the standard deviation of order $k_n^{1/2}$. Said differently, the test has power bounded away from one whenever $\Pi' P_Z \Pi / \sqrt{k_n}$ is bounded from above. Mikusheva and Sun (2020) show that the same is true for any test based on an increasing number of instruments.

One should then select k_n so that the above ratio is as large as possible. But the numerator depends on the unknown functional form $\Pi(\cdot)$ and on the unknown identification strength. While under strong identification (when $\Pi' P_Z \Pi$ diverges at rate n), nonparametric optimal instruments should be used for efficiency as they maximize the numerator, they cannot be relied upon under weak identification (when $\Pi' P_Z \Pi$ stays bounded) since the above ratio would then converge to zero and power would be trivial. The power of the AR test is also bounded away from one under some semi-strong identification. If we assume that $\Pi(Z) = \tilde{c}_n \frac{C(Z)}{\sqrt{n}}$, where $\mathbb{E}\, C^2(Z) < \infty$, then this happens whenever $\tilde{c}_n^2$ is of order equal to or smaller than $\sqrt{k_n}$. This does not depend on the method used to estimate the first-stage. For instance, Jun and Pinkse (2012) propose an AR-type test where optimal instruments are estimated using k_n nearest-neighbors. Their arguments can be used to show their test has power bounded away from one whenever $\tilde{c}_n$ is of order equal to or smaller than $\sqrt{n/k_n}$.[3]

By contrast, the tests we develop below do not necessitate the choice of the number of terms in a series expansion or of a smoothing parameter to estimate the first-stage equation. Consequently it is consistent under a fixed alternative for any diverging sequence $\tilde{c}_n$ and has more than trivial power for a bounded but large enough $\tilde{c}_n$, as shown in Theorem 5.2 below. As little prior information is typically available to appropriately parametrize the first-stage equation, a testing method that leaves the first-stage equation unspecified while being robust to weak identification appears extremely valuable from a practitioner's viewpoint.

[3] The conclusions reached for series approximation and nearest-neighbors can be intuitively compared by referring to degrees of freedom for each method. For series estimation, k_n is the degrees of freedom, while for a smoother such as the nearest-neighbors method, degrees of freedom are usually measured by the trace of the smoothing matrix, that is n/k_n.

3 ICM and Conditional ICM Tests Statistics

Without assuming linearity of $\Pi(\cdot)$ in (2.3), we can write

$$y - Y_2\beta_0 = \Pi(Z)\,(\beta - \beta_0) + \varepsilon, \qquad \text{where} \quad \varepsilon = u + V_2\,(\beta - \beta_0) \quad \text{and} \quad \mathbb{E}\,(\varepsilon|Z) = 0\,.$$

The variables Z include the instruments X_2 but also the exogenous X_1 to account for potential nonlinearities in X_1 in the function $\Pi(\cdot)$. We consider testing

$$\tilde{H}_0 : \mathbb{E}\,(y - Y_2'\beta_0|Z) = 0 \quad \text{a.s.}$$

which is implied by the model when $\beta = \beta_0$. That is, we consider at the same time H_0 and the correct specification of the model, in the same way the AR test does. We then apply a result of Bierens (1982) which states that $\tilde{H}_0$ holds if and only if

$$\mathbb{E}\,\left[(y - Y_2'\beta_0)\exp(is'Z)\right] = 0 \quad \forall s \in \mathbb{R}^k\,. \tag{3.4}$$

To test this hypothesis, Bierens' Integrated Conditional Moment (ICM) statistic is

$$\int_{\mathbb{R}^k} |n^{-1/2}\sum_{j=1}^{n}\left(y_j - Y_{2j}'\beta_0\right)\exp(is'Z_j)|^2\,d\mu(s)\,, \tag{3.5}$$

where μ is some symmetric probability measure with support $\mathbb{R}^k$ (except maybe a set of isolated points). Let us define

$$w(z) = \int_{\mathbb{R}^k}\exp(is'z)\,d\mu(s) = \int_{\mathbb{R}^k}\cos(s'z)\,d\mu(s)\,,$$

due to the symmetry of μ. We can then rewrite the statistic (3.5) as

$$\begin{aligned}
&\int_{\mathbb{R}^k} n^{-1}\sum_{j=1}^{n}\sum_{m=1}^{n}(Y_j'b_0)(Y_m'b_0)\exp(is'(Z_j - Z_m))\,d\mu(s)\\
&= n^{-1}\sum_{j=1}^{n}\sum_{m=1}^{n}(Y_j'b_0)(Y_m'b_0)\int_{\mathbb{R}^k}\exp(is'(Z_j - Z_m))\,d\mu(s)\\
&= b_0'Y'WYb_0\,,
\end{aligned}$$

where W is a matrix with generic element $n^{-1}w\,(Z_j - Z_m)$. The condition for μ to have support $\mathbb{R}^k$ translates into the restriction that $w(\cdot)$ should have a strictly positive Fourier transform almost everywhere. Examples include products of triangular, normal, logistic, see Johnson et al. (1995, Section 23.3), Student, including Cauchy, see Dreier and Kotz

(2002), or Laplace densities. To achieve scale invariance, we recommend, as in Bierens (1982), to scale the exogenous instruments by a measure of dispersion, such as their empirical standard deviation. The role of the function $w(\cdot)$ resembles the one of the kernel in nonparametric estimation, but, in contrast, it is a *fixed user-chosen function that does not vary with the sample size.* To make this explicit, we will impose that the squared integral of $w(\cdot)$ equals one.[4]

If Z has bounded support, results from Bierens (1982) yield that $\tilde{H}_0$ holds if and only if

$$\mathbb{E}\left[(y - Y_2'\beta_0)\exp(s'Z)\right] = 0$$

for all s in a (arbitrary) neighborhood of 0 in $\mathbb{R}^q$. Hence μ in (3.5) can be taken as any symmetric probability measure that contains 0 in the interior of its support. For instance, we can consider the product of uniform distributions on $[-\pi, \pi]$, so that $w(\cdot)$ is the product of sinc functions. As noted by Bierens (1982), there is no loss of generality to assume a bounded support, as his equivalence result equally applies to a one-to-one transformation of Z, which can be chosen with bounded image. Moreover, if it is known that

$$\mathbb{E}\left(y - Y_Z\beta_0 | Z\right) = \mathbb{E}\left(y - Y_Z\beta_0 | \Psi(Z)\right),$$

for some known dimension-reducing function $\Psi(\cdot)$, then W could be defined using this transformation instead.

The ICM principle replaces conditional moment restrictions by a continuum of unconditional moments such as (3.4). Other functions have been used beyond the complex exponential, see Bierens (1990) and Bierens and Ploberger (1997). Stinchcombe and White (1998) give a characterization of a large class of functions that could generate an equivalent set of unconditional moments. As detailed by Lavergne and Patilea (2013), this yields a full collection of potential estimators under strong (or semi-strong) identification, such as the ones developed by Dominguez and Lobato (2004), Antoine and Lavergne (2014), and Escanciano (2018) among others. This would also yield a collection of test statistics that could be used under weak identification, see Chen et al. (2021) for a recent instance. We here focus on a particular application of the ICM suitable for theoretical investigation and practical implementation, and we leave for future work the investigation of the relative merits of these different ICM-type tests.

[4] A more involved restriction would be to impose a similar condition on the Frobenius norm of W.

Let $\widehat{\Omega}$ be a (semiparametric) estimator of $\Omega = \mathbb{E}\left(\mathrm{Var}(Y|Z)\right)$. Our first test statistic is

$$\mathrm{ICM}(\beta_0) = \frac{b_0'Y'WYb_0}{b_0'\widehat{\Omega}b_0} \qquad \text{with} \qquad b_0 = (1, -\beta_0')' \,. \tag{3.6}$$

It is the ICM statistic with the value of the parameter set at β_0 and standardized by an estimator of the variance of $Y_i'b_0$. It resembles the AR statistic, with W replacing P_Z, the orthogonal projection on Z. The statistic is also related to Antoine and Lavergne (2014) Weighted Minimum Distance objective function, though they chose a different normalization and only consider semi-strong identification. Our normalization does not affect the main properties of the ICM test, but is convenient when computing critical values and studying theoretical properties. As apparent from its construction, ICM is designed to test the correct specification of the model together with the parameter value, as does the AR test under a linear reduced form. Since ICM equals (3.5) up to the positive term $b_0'\widehat{\Omega}b_0$, it is non-negative, and the test rejects the null hypothesis for large positive values of the statistic.

Our conditional ICM (CICM) test is based on the statistic

$$\mathrm{CICM}(\beta_0) = \frac{b_0'Y'WYb_0}{b_0'\widehat{\Omega}b_0} - \min_b \frac{b'Y'WYb}{b'\widehat{\Omega}b} \,. \tag{3.7}$$

The statistic has the form of a quasi likelihood-ratio statistic and is always non-negative. The test thus rejects the null hypothesis for large positive values of the statistic. It does not test the whole specification of the model, but only whether β_0 is compatible with the data assuming the model is adequate.

The CICM statistic resembles the LR one of Moreira (2003), with W replacing P_Z, the orthogonal projection on Z. We now follow his discussion and define

$$\widehat{S} \equiv \widehat{S}(\beta_0) = Yb_0\left(b_0'\widehat{\Omega}b_0\right)^{-1/2}, \; \widehat{T} \equiv \widehat{T}(\beta_0) = Y\widehat{\Omega}^{-1}A_0\left(A_0'\widehat{\Omega}^{-1}A_0\right)^{-1/2}, \; A_0 = [\beta_0 \; \mathbf{I}]' \,.$$

Then $\mathrm{ICM}(\beta_0) = \widehat{S}'W\widehat{S}$ and

$$\mathrm{CICM}(\beta_0) = \widehat{S}'W\widehat{S} - \lambda_{\min}\left(\begin{bmatrix} \widehat{S}' \\ \widehat{T}' \end{bmatrix} W \left[\widehat{S}, \widehat{T}\right]\right) , \tag{3.8}$$

where $\lambda_{\min}(A)$ is the smallest eigenvalue of the matrix A. When β_0 is scalar,

$$\mathrm{CICM}(\beta_0) = \frac{1}{2}\left[\widehat{S}'W\widehat{S} - \widehat{T}'W\widehat{T} + \sqrt{\left(\widehat{S}'W\widehat{S} - \widehat{T}'W\widehat{T}\right)^2 + 4\left(\widehat{S}'W\widehat{T}\right)^2}\right] . \tag{3.9}$$

To establish (3.8), note that

$$\min_b \frac{b'Y'WYb}{b'\widehat{\Omega}b} = \lambda_{\min}\left(\widehat{\Omega}^{-1/2}Y'WY\widehat{\Omega}^{-1/2}\right) .$$

where $\lambda_{\min}(M)$ is the minimum eigenvalue of M. Consider the orthogonal matrix

$$J = \left[\widehat{\Omega}^{1/2}b_0\left(b_0'\widehat{\Omega}b_0\right)^{-1/2}, \widehat{\Omega}^{-1/2}A_0\left(A_0'\widehat{\Omega}^{-1}A_0\right)^{-1/2}\right] ,$$

where $J'J = \mathbf{I}$ since $A_0'b_0 = \mathbf{0}$. The minimum eigenvalue of $\widehat{\Omega}^{-1/2}Y'WY\widehat{\Omega}^{-1/2}$ is thus the one of $J'\widehat{\Omega}^{-1/2}Y'WY\widehat{\Omega}^{-1/2}J$, and $Y\widehat{\Omega}^{-1/2}J = [\widehat{S}, \widehat{T}]$. We label our test as conditional because we will use conditional critical values. With homoskedastic errors, we will condition on Z and $\widehat{T}$. This allows to condition on the set of statistics $\widehat{T}'W\widehat{T}$ that convey information on identification strength. Consider for simplicity the scalar case. Then $\widehat{T}'W\widehat{T}$ is the ICM statistic for testing $\Pi(\cdot) = \mathbf{0}$ a.s. It can then be seen as the nonparametric ICM equivalent of the first-stage F statistic. In particular, its large sample mean can be viewed as some measure of identification strength similar to the concentration parameter.

4 Tests with Normal Errors and Known Covariance Structure

We now explain how to obtain critical values and P-values. We assume normal errors with a known covariance structure. We relax both assumptions in the next section, where we show that estimation of the covariance structure has no first-order asymptotic effect on the validity of our tests. Since Ω is considered known here, we replace $\widehat{S}$ and $\widehat{T}$ by $S = Yb_0\left(b_0'\Omega b_0\right)^{-1/2}$ and $T = Y\Omega^{-1}A_0\left(A_0'\Omega^{-1}A_0\right)^{-1/2}$.

4.1 Homoskedastic Case

Under H_0, $S \sim N(\mathbf{0}, \mathbf{I})$ conditionally on Z. Then ICM $= S'WS$ follows a weighted sum of independent chi-squares, specifically ICM $\sim \sum_{k=1}^{n}\lambda_k G_k^2$ conditionally on Z, where $G_1, \ldots, G_n$ are standard independent normal random variables and $\lambda = (\lambda_1, \ldots, \lambda_n)$ are the positive eigenvalues of W, see e.g. de Wet and Venter (1973). The distribution of ICM under H_0 can thus easily be simulated by drawing many times $G \sim N(\mathbf{0}, \mathbf{I})$, and computing the associated quadratic form $G'WG$. Critical values are then obtained as the quantiles of the empirical distribution of the simulated statistic. Equivalently, one can compute the P-value of the test as the empirical probability that the original test statistic is lower than the simulated statistic.

Consider now the joint behavior of $S = Yb_0\left(b_0'\Omega b_0\right)^{-1/2}$ and the columns of $T = Y\Omega^{-1}A_0\left(A_0'\Omega^{-1}A_0\right)^{-1/2}$. Under H_0, they are jointly normally distributed. Each column of T is uncorrelated with S, and thus independent of S, conditionally on Z. This entails that the distribution of CICM(β_0) under H_0 can be simulated *keeping Z and T fixed* by replacing S by $G \sim N(\mathbf{0}, \mathbf{I})$ in the formula of the statistic. The resulting quantiles now depend on β_0 via $T = T(\beta_0)$. This conditional method of obtaining critical values allows in particular to condition on the matrix $T'WT$ that contains the set of ICM statistics that evaluates the strength of the link of endogenous regressors to instruments.

4.2 Heteroskedastic Case

Heteroskedasticity is often encountered in microeconometric applications. The usual way to account for potential unknown heteroskedasticity is to modify the test statistic at the outset. For instance, Stock and Wright (2000) and Chernozhukov and Hansen (2008) adapt the Anderson-Rubin statistic using a heteroskedasticity-robust estimator of the covariance matrix. Conditional tests that are robust to heteroskedasticity have been proposed by Andrews et al. (2006) (in the working paper version of their article), Kleibergen (2007), Andrews (2016), Moreira and Ridder (2017), and Moreira and Moreira (2019). Andrews and Mikusheva (2016a) note that standard CLR could be used in heteroskedastic contexts by conditioning on the statistic of Kleibergen (2005), and more generally that a wide class of QLR tests are valid when conditioning on a nuisance process. Here we work with the statistics ICM and CICM and we adapt their critical values to heteroskedasticity. There may well be modifications of our statistics that could account for heteroskedasticity, but they would be of a different form and thus would not have the same intuitive interpretation. We leave this topic for future investigation.

Let us assume for now that the conditional variance function

$$\Omega_i \equiv \Omega(Z_i) = \operatorname{Var}\left(Y_i|Z_i\right) = \begin{pmatrix} \operatorname{Var}(y_i|Z_i) & \operatorname{Cov}(y_i, Y_{2i}|Z_i) \\ \operatorname{Cov}'(Y_{2i}, y_i|Z_i) & \operatorname{Var}(Y_{2i}|Z_i) \end{pmatrix}, \tag{4.10}$$

is known, so that we can compute $\Sigma = \operatorname{Var}(S|Z)$. Then

$$\mathrm{ICM} = S\Sigma^{-1/2}\Sigma^{1/2}W\Sigma^{1/2}\Sigma^{-1/2}S\,,$$

and, under H_0, ICM follows the same distribution as $G'\Sigma^{1/2}W\Sigma^{1/2}G$, where $G \sim N(\mathbf{0}, \mathbf{I})$. We can then again simulate the distribution of ICM under H_0 and recover critical values.

The null distribution of CICM only depends on the covariance structure of S and T conditional on Z under Lindeberg-type conditions, see Rotar' (1979). Under H_0, $(S_i, T_i')'$ has conditional mean

$$\begin{pmatrix} 0 \\ \Pi(Z_i)(A_0'\Omega^{-1}A_0)^{1/2} \end{pmatrix} \tag{4.11}$$

and conditional variance matrix

$$\begin{pmatrix} \left(b_0'\Omega b_0\right)^{-1} b_0'\Omega_i b_0 & \left(b_0'\Omega b_0\right)^{-1/2} b_0'\Omega_i\Omega^{-1}A_0\left(A_0'\Omega^{-1}A_0\right)^{-1/2} \\ \cdot & \left(A_0'\Omega^{-1}A_0\right)^{-1/2} A_0'\Omega^{-1}\Omega_i\Omega^{-1}A_0\left(A_0'\Omega^{-1}A_0\right)^{-1/2} \end{pmatrix},$$

so that S and T are not conditionally independent with normal errors. We can however condition on the part of T that is uncorrelated with S. Specifically, let

$$R = [R_1 \ldots R_n] \qquad R_i = T_i - \frac{\operatorname{Cov}(T_i, S_i|Z_i)}{\operatorname{Var}(S_i|Z_i)} S_i \,.$$

Then, under H_0, $(S_i, R_i')'$ has the same conditional mean (4.11) as $(S_i, T_i')'$ and conditional variance matrix

$$\begin{pmatrix} \left(b_0'\Omega b_0\right)^{-1} b_0'\Omega_i b_0 & \mathbf{0} \\ \cdot & \left(A_0'\Omega^{-1}A_0\right)^{-1/2}\left(A_0'\Omega^{-1}\Omega_i\Omega^{-1}A_0 - \frac{A_0'\Omega^{-1}\Omega_i b_0 b_0'\Omega_i\Omega^{-1}A_0}{b_0'\Omega_i b_0}\right)\left(A_0'\Omega^{-1}A_0\right)^{-1/2} \end{pmatrix}.$$

Hence, with Gaussian errors, S and R are conditionally jointly Gaussian and independent, S is pivotal under H_0, while R is sufficient for Π. To simulate the distribution of CICM keeping R and Z fixed, we generate G_i, $i = 1, \ldots n$, as independent normal with mean 0 and variance $\operatorname{Var}(S_i|Z_i)$ for each i, and we compute CICM with drawings of G_i in place of S_i and

$$R_i + \frac{\operatorname{Cov}(T_i, S_i|Z_i)}{\operatorname{Var}(S_i|Z_i)} G_i$$

in place of T_i.

The above orthogonalization method is related to the one proposed by Andrews and Mikusheva (2016a). In a linear IV model, they consider testing $\mathbb{E}\left[Z(y - Y_2'\beta_0)\right] = 0$. They suggest to view the mean function $\mathbb{E}\left[Z(y - Y_2'\beta)\right]$ for all other values of β as a nuisance parameter, and they propose to condition a test of the null hypothesis on the process of sample moments evaluated at any other value β. To do so, the sample process $n^{-1}\sum_{i=1}^n Z_i\left(y_i - Y_{2i}'\beta\right)$ needs to be orthogonalized with respect to the sample mean $n^{-1}\sum_{i=1}^n Z_i\left(y_i - Y_{2i}'\beta_0\right)$ through their estimated covariance function. The issue with CICM is similar but more intricate, as we are interested in the mean process $\mathbb{E}\left[(y - Y_2'\beta_0)\exp(is'Z)\right]$ for all s, and we consider $\mathbb{E}\left[(y - Y_2'\beta)\exp(it'Z)\right]$ for all other values of β and all t as a nuisance parameter. To orthogonalize the process $n^{-1}\sum_{i=1}^{n}(y_i - Y_{2i}'\beta)\exp(it'Z_i)$ with respect to $n^{-1}\sum_{i=1}^{n}(y_i - Y_{2i}'\beta_0)\exp(is'Z_i)$, we use a transformation that removes correlation at the level of individual observations.

4.3 Similarity of the Tests

Similar tests have been shown to perform well in weakly identified linear IV models, see Andrews et al. (2006). The ideal normal setup may seem unrealistic, but retains however the main ingredients of the problem. Indeed, the test statistics ultimately depend on empirical processes that are jointly asymptotically Gaussian whatever the particular error distribution, see Section 8. Hence the ideal setup allows to study the properties of our test abstracting from finite-sample considerations.

Define the conditional critical values as

$$\begin{aligned} c_{1-\alpha}(Z) &= \inf\left\{c \,:\, \Pr\left[\mathrm{ICM}\left(\beta_0\right) \leq c|Z\right] \geq 1-\alpha\right\} \\ c_{1-\alpha}(Z, R(\beta_0)) &= \inf\left\{c \,:\, \Pr\left[\mathrm{CICM}\left(\beta_0\right) \leq c|Z, R(\beta_0)\right] \geq 1-\alpha\right\}\,. \end{aligned}$$

Then, in the normal case with known $\Omega(\cdot)$,

$$\begin{aligned} \Pr\left[\mathrm{ICM}(\beta_0) > c_{1-\alpha}(Z)|Z\right] &= \Pr\left[\mathrm{ICM}(\beta_0) > c_{1-\alpha}(Z)\right] = \alpha\,. \\ \Pr\left[\mathrm{CICM}(\beta_0) > c_{1-\alpha}(Z, R(\beta_0))|Z, R(\beta_0)\right] &= \Pr\left[\mathrm{CICM}(\beta_0) > c_{1-\alpha}(Z, R(\beta_0))\right] = \alpha\,. \end{aligned}$$

The ICM test is similar because $\Sigma^{-1/2}S \sim N(\mathbf{0}, \mathbf{I})$ conditionally on Z. For CICM, the result follows because, in addition, (i) the components of $\left[\Sigma^{-1/2}S, R\right]$ are jointly conditionally normal, and (ii) $\Sigma^{-1/2}S$ is conditionally uncorrelated with the components of R, and thus conditionally independent of R.

5 Asymptotic Tests

The setup of normal errors with known conditional covariance structure is ideal but not realistic. However, our method for simulating critical values remains asymptotically valid when errors are not Gaussian, and conditional variances are estimated instead of known.

5.1 Homoskedastic Case

If we first drop the normality assumption, ICM asymptotically follows the conditional distribution described in the last section. This is mainly based on the invariance principle developed by Rotar' (1979). Specifically, ICM $= S'WS$ is a quadratic form in S, and its asymptotic distribution only depends on the first two conditional moments of S. Under homoskedasticity, $S \sim N(\mathbf{0}, \mathbf{I})$ conditionally on Z, so replacing S by a standard Gaussian vector G results in the same asymptotic distribution. The procedure explained in the last section thus provides asymptotically valid critical value $c_{1-\alpha}(Z, \widehat{\Omega})$, depending upon a consistent estimator $\widehat{\Omega}$, as the $1-\alpha$ quantile of the statistic obtained by simulations. Under homoskedasticity, this critical value is independent of the particular value of β_0. The confidence set obtained by inverting the ICM test is $\left\{\beta_0 : ICM(\beta_0) < c_{1-\alpha}(Z, \widehat{\Omega})\right\}$.

When β_0 is scalar, ICM (β_0) is a ratio of two quadratic forms in β_0, and the confidence interval is obtained by solving a quadratic inequality, as is the AR confidence interval, see Dufour and Taamouti (2005) and Mikusheva (2010). We thus obtain that our confidence interval has four possible forms: (i) a finite interval (β_1, β_2); (ii) the union of two infinite intervals $(-\infty, \beta_2) \cup (\beta_1, +\infty)$; (iii) the whole real line $(-\infty, +\infty)$; (iv) the empty set. The last possibility arises as our null hypothesis $\tilde{H}_0$ states the validity of the model given β_0. Indeed, ICM is designed to test the correct specification of the model together with the parameter value.

The conditional ICM statistic depends on $S'WS$, $S'WT$, and $T'WT$ as seen from (3.8), which are linear and quadratic forms in S. Under homoskedasticity, S is uncorrelated with the columns of T (conditional on Z), and the method exposed previously in the Gaussian case provides asymptotically correct critical values. As any quasi-likelihood ratio test, the CICM test is one-sided and rejects the null hypothesis when the statistic is large. A confidence set for β is defined as $\left\{\beta_0 : ICM(\beta_0) < c_{1-\alpha}(Z, \widehat{\Omega}, \widehat{R}(\beta_0))\right\}$, where $c_{1-\alpha}(Z, \widehat{\Omega}, \widehat{R}(\beta_0))$ is the $1-\alpha$ quantile of the statistic obtained by simulations. However, it does not seem possible to obtain a simple characterization of CICM-based confidence intervals as done by Mikusheva (2010) for CLR.

5.2 Heteroskedastic Case

Accounting for unknown heteroskedasticity requires estimating conditional variances of Y. In order to state our uniform asymptotic validity result, see Theorem 5.1 below, one of our main tasks will be to establish asymptotic results accounting for estimation of $\Omega = \mathbb{E}\,\mathrm{Var}(Y|Z)$ and $\Omega(\cdot) = \mathrm{Var}(Y|Z = \cdot)$. One should note that weak identification does not preclude consistent estimation of these objects. If Ω is unknown, there are many existing estimators in the literature, for instance the difference-based estimator of Rice (1984) and generalizations by Seifert et al. (1993) among others. The conditional variance $\Omega(\cdot)$ can be estimated parametrically if one is ready to make an assumption on its functional form. Otherwise, we can resort to nonparametric conditional variance estimation. Several consistent ones have been developed for a univariate Y, and generalize easily. To make things concrete, we focus on kernel smoothing, which is used in our simulations and application. Let

$$\overline{Y}(z) = (nb_n)^{-1}\sum_{i=1}^{n} Y_i K\left((Z_i - z)/b_n\right)$$

based on the n iid observations (Y_i, Z_i), a kernel $K(\cdot)$, and a bandwidth b_n. With $e = (1, \ldots 1)'$, let $\widehat{f}(z) = \overline{e}(z)$ and $\widehat{Y}(z) = \overline{Y}(z)/\widehat{f}(z)$. The conditional variance estimator of Y is defined as

$$\widehat{\Omega}(z) = (nb_n)^{-1}\,\frac{\sum_{i=1}^{n}\left(Y_i - \widehat{Y}(Z_i)\right)\left(Y_i - \widehat{Y}(Z_i)\right)' K\left((Z_i - z)/b_n\right)}{\widehat{f}(z)}\,.$$

This estimator, studied by Yin et al. (2010), is a generalization of the kernel conditional variance, and is positive definite whenever $K(\cdot)$ is positive. It provides a consistent estimator of the variance matrix function $\Omega(\cdot)$, and a consistent estimator of Ω using $\widehat{\Omega} = n^{-1}\sum_{i=1}^{n}\widehat{\Omega}(Z_i)$. Note that we could equivalently consider an estimator of the uncentered moment $\mathbb{E}\,(Y'Y)$ and then avoid preliminary estimation of $\mathbb{E}\,(Y|Z)$. Indeed $\mathbb{E}\,(S|Z) = 0$ a.s. under H_0 so that $\text{Var}\,(S|Z) = \mathbb{E}\,(S^2|Z)$ and $\text{Cov}\,(T, S|Z) = \mathbb{E}\,(T'S|Z)$.

With at hand a parametric or nonparametric estimator of $\Omega(\cdot)$, one can estimate the conditional variance of S_i by $\widehat{\text{Var}}(S_i|Z_i) = b_0'\widehat{\Omega}_i b_0\left(b_0\widehat{\Omega}b_0\right)^{-1}$, where $\widehat{\Omega}_i \equiv \widehat{\Omega}(Z_i)$. To approximate the asymptotic distribution of ICM $= S'WS$, we generate independent Gaussian $\widehat{G}_i$, $i =, \ldots n$, with mean 0 and variance $\widehat{\text{Var}}(S_i|Z_i)$ for each i, and proceeds similarly as above. The intuition carries over for CICM, provided we condition on the part of $\widehat{T}$ which is asymptotically uncorrelated with $\widehat{S}$ conditional on Z. The conditional covariance of $\widehat{T}_i$ and $\widehat{S}_i$ can be estimated as $\left(A_0'\widehat{\Omega}^{-1}A_0\right)^{-1/2} A_0'\widehat{\Omega}^{-1}\widehat{\Omega}_i b_0\left(b_0'\widehat{\Omega}b_0\right)^{-1/2}$. Then the asymptotic distribution of CICM will be approximated by first computing $\widehat{R} = \left[\widehat{R}_1 \ldots \widehat{R}_n\right]$, with

$$\widehat{R}_i = \widehat{T}_i - \frac{\widehat{\text{Cov}}(T_i, S_i|Z_i)}{\widehat{\text{Var}}(S_i|Z_i)}\widehat{S}_i = \left(A_0'\widehat{\Omega}^{-1}A_0\right)^{-1/2}\left[Y_i'\widehat{\Omega}^{-1}A_0 - \frac{A_0'\widehat{\Omega}^{-1}\widehat{\Omega}_i b_0}{b_0'\widehat{\Omega}_i b_0}Y_i'b_0\right],$$

then recomputing CICM with drawings of G_i in place of $\widehat{S}_i$ and $\widehat{R}_i + \frac{\widehat{\text{Cov}}(T_i,S_i|Z_i)}{\widehat{\text{Var}}(S_i|Z_i)}G_i$ in place of $\widehat{T}_i$.

5.3 Uniform Asymptotic Validity

We consider the following assumptions.

Assumption A *(i) The observations (y_i, Y_{2i}, Z_i) form a rowwise independent triangular array that follows (2.2) and (2.3), where the marginal distribution of Z remains unchanged.*
(ii) For some $\delta > 0$ and $M' < \infty$, $\sup_z \mathbb{E}\left(\|Y\|^{2+\delta} | Z = z\right) \leq M'$ uniformly in n.

The assumption of a constant distribution for Z could be weakened, but is made to formalize that identification strength is related to the conditional distribution of Y given Z only. For the sake of simplicity, we will not use a double index for observations and will denote by $\{Y_1, \ldots, Y_n\}$ the independent copies from Y for a sample size n. We denote by $\mathcal{P}$ the class of distributions on which our observations lie.

Let $\mathcal{E}$ be a class of vector-valued functions $\Pi(\cdot)$ and let $N(\varepsilon, \mathcal{E}, L_2(Q))$ be the covering number of $\mathcal{E}$, that is the minimum number of $L_2(Q)$ ε-balls needed to cover $\mathcal{E}$, where an $L_2(Q)$ ε-ball around $\Pi(\cdot)$ is the set of vector functions $\left\{h \in L_2(Q) \ : \ \int \|h - \Pi\|^2 \, dQ < \varepsilon\right\}$.

Assumption B *The conditional expectation vector $\mathbb{E}(Y_2 | Z = \cdot)$ belongs to a class of vector functions $\mathcal{E}$ such that $\forall\, \Pi(\cdot) \in \mathcal{E}$, $\|\Pi(\cdot)\| \leq F(\cdot)$ with*

$$\lim_{M \to \infty} \sup_{\mathcal{P}} \mathbb{E}\left[F^2(Z) \mathbb{I}\left(F(Z) > M\right)\right] = 0$$

and

$$\log N\left(\varepsilon \mathbb{E}^{1/2}\left(F^2(Z)\right), \mathcal{E}, L^2(P)\right) \leq K\varepsilon^{-V} \quad \text{for some } V < 2\,,$$

for all $P \in \mathcal{P}$ and some K, V independent of P.

Andrews (1994) and van der Vaart (1994), among others, exhibit classes of smooth functions that fulfill the above conditions.

Let $\mathcal{O}$ be a class of matrix-valued functions and let $N(\varepsilon, \mathcal{O}, L_2(Q))$ be the covering number of $\mathcal{O}$, defined similarly as above.

Assumption C *(i)* $\sup_{P \in \mathcal{P}} \Pr\left[\|\widehat{\Omega} - \Omega\| > \varepsilon\right] \to 0 \quad \forall \varepsilon > 0$.

(ii) $\Omega(\cdot)$ belongs to a class of matrix functions $\mathcal{O}$ such that $O < \underline{\lambda} \leq \inf_z \lambda_{\min}(\Omega(z)) \leq \sup_z \lambda_{\max}(\Omega(z)) \leq \overline{\lambda} < \infty$ for all $\Omega(\cdot) \in \mathcal{O}$ and

$$\log N\left(\varepsilon, \mathcal{O}, L^2(P)\right) \leq K\varepsilon^{-V} \quad \text{for some } V < 2\,,$$

for all $P \in \mathcal{P}$ and some K, V independent of P.

(iii) $\sup_{P\in\mathcal{P}} \Pr\left(\widehat{\Omega}(\cdot)\in\mathcal{O}\right)\to 1$ *as* $n\to\infty$

(iv) $\sup_{P\in\mathcal{P}} \int \|\widehat{\Omega}(Z)-\Omega(Z)\|^2\,dP(Z)\xrightarrow{p} 0.$

This assumption entails in particular that conditional variance estimation does not affect the asymptotic behavior of our statistics. There is a tension between the generality of the class of functions $\mathcal{O}$ and the class of possible distributions $\mathcal{P}$. When $\Omega(\cdot)$ is of a parametric form, Assumption C will be satisfied for a large class of distributions. When $\Omega(\cdot)$ is considered nonparametric and estimated accordingly, one typically assumes that its components are smooth functions, and to prove (iii) one has to show that $\widehat{\Omega}(\cdot)$ also satisfies the same smoothness conditions with probability converging to 1. Such results have been derived, see e.g. Andrews (1995) for kernel estimators or Cattaneo and Farrell (2013) for partitioning estimators. Uniform convergence of nonparametric regression estimators (and their derivatives) generally requires the domain of the functions to be bounded and the absolutely continuous components of the distributions of the conditioning variables to have densities bounded away from zero on their support. When they are not, Andrews (1995) discusses the use of a vanishing trimming that is compatible with the stochastic equicontinuity results of Andrews (1994). Condition (iv) is dealt with in the literature on honest confidence intervals using L^2 norm, see e.g. Robins and van der Vaart (2006) and the references therein.

Assumption D *$w(\cdot)$ is a symmetric, bounded density with $\int w^2(x)\,dx=1$. Its Fourier transform is a density, which is positive almost everywhere, or whose support contains a neighborhood of the origin if Z is bounded.*

We respectively denote by $c_{1-\alpha}(\beta_0,Z,\widehat{\Omega}(\cdot))$ and $c_{1-\alpha}(\beta_0,Z,\widehat{\Omega}(\cdot),\widehat{R}(\beta_0))$ the conditional critical values of ICM and CICM obtained by the simulation-based method detailed above.[5] Let $\mathcal{P}_{\beta_0}$ be the subset of distributions in $\mathcal{P}$ such that $\beta=\beta_0$. The following result establishes that our tests control size uniformly over a large class of probability distributions.

Theorem 5.1 *Under Assumptions A, B, C, and D,*

$$\limsup_{n\to\infty}\sup_{\beta_0}\sup_{P\in\mathcal{P}_{\beta_0}}\Pr\left[\mathrm{ICM}(\beta_0)>c_{1-\alpha}(\beta_0,Z,\widehat{\Omega}(\cdot))\right]\le\alpha$$

$$\limsup_{n\to\infty}\sup_{\beta_0}\sup_{P\in\mathcal{P}_{\beta_0}}\Pr\left[\mathrm{CICM}(\beta_0)>c_{1-\alpha}(\beta_0,Z,\widehat{\Omega}(\cdot),\widehat{R}(\beta_0))\right]\le\alpha\,.$$

[5] We neglect the approximation error due to a finite number of simulations by assuming the number of simulations is infinite so that the critical values are exact.

Our theorem readily implies that our tests are asymptotically valid whatever identification strength. Indeed, for any sequence $\Pi_n(\cdot), n\ geq 1$ of functions in $\mathcal{E}$, that can decrease in norm to zero arbitrarily fast, our result yields asymptotic validity under this sequence, see e.g. van der Vaart and Wellner (2000, Chap. 2.8).

5.4 Asymptotic Power

We adopt here a large local alternatives setup similar to Bierens and Ploberger (1997).

Assumption E *There exists a fixed matrix $C(\cdot)$ such that $\mathbb{E}\, C(Z)C'(Z)$ is bounded and positive definite, and either (i) $\Pi(Z) = \tilde{c}_n\, \frac{C(Z_i)}{\sqrt{n}}$ or (ii) $\Pi(Z) = C(Z_i)$.*

Condition (i) allows to study the power of our tests against weak and semi-strong identification, when considering a test of $H_0 :\ \beta = \beta_1$ where $\beta_1 \neq \beta_0$, the true parameter value. Condition (ii) is the strong identification case and we consider local alternatives of the type $H_{1n} :\ \beta_{1n} = \beta_0 + \tilde{c}_n \frac{\delta}{\sqrt{n}}$, where $\delta \neq 0$ is fixed. In both cases, the object of interest is the asymptotic power of our two tests when $\tilde{c}_n$ becomes large.

Theorem 5.2 *Under Assumptions A, C, and D,*

(i) under Assumption E-(i), for any fixed $\beta_1 \neq \beta_0$,

$$\liminf_{\tilde{c}_n\to\infty}\ \inf_{P\in\mathcal{P}_{\beta_0}} \Pr\left[\mathrm{ICM}(\beta_1) > c_{1-\alpha}(\beta_1, Z, \widehat{\Omega}(\cdot))\right] = 1$$

$$\liminf_{\tilde{c}_n\to\infty}\ \inf_{P\in\mathcal{P}_{\beta_0}} \Pr\left[\mathrm{CICM}(\beta_1) > c_{1-\alpha}(\beta_1, Z, \widehat{\Omega}(\cdot), \widehat{R}(\beta_1))\right] = 1\,.$$

(ii) under Assumption E-(ii), for $\beta_{1n} = \beta_0 + \tilde{c}_n \frac{\delta}{\sqrt{n}}$ and a fixed $\delta \neq 0$,

$$\liminf_{\tilde{c}_n\to\infty}\ \inf_{P\in\mathcal{P}_{\beta_0}} \Pr\left[\mathrm{ICM}(\beta_{1n}) > c_{1-\alpha}(\beta_{1n}, Z, \widehat{\Omega}(\cdot))\right] = 1$$

$$\liminf_{\tilde{c}_n\to\infty}\ \inf_{P\in\mathcal{P}_{\beta_0}} \Pr\left[\mathrm{CICM}(\beta_{1n}) > c_{1-\alpha}(\beta_{1n}, Z, \widehat{\Omega}(\cdot), \widehat{R}(\beta_{1n}))\right] = 1\,.$$

Result (i) shows that under weak identification power is non trivial for a large enough $\tilde{c}_n$. For ICM, one can understand the result from the following arguments due to Bierens and Ploberger (1997). The asymptotic distribution of $\mathrm{ICM}(\beta_1)$ is given by $\sum_{i=1}^{n} \lambda_i \left(G_i + c_i\right)^2$, where $\lambda_i, i = 1, \ldots n$, are strictly positive real numbers, $G_i, i = 1, \ldots n$, are independent standard normals, and $c_i, i = 1, \ldots n$, are non-zero real numbers. This distribution stochastically dominates at first order the asymptotic distribution of ICM(β_0), which is similar but with $c_i = 0$ for all i. The behavior of CICM is more involved because it depends on the behavior of the whole process ICM(β) for any β. Result (ii) implies that, under strong identification, power is non trivial under a sequence of Pitman local alternatives for $\tilde{c}_n$ large enough.

6 Small Sample Behavior

We investigate the small sample properties of our tests in the structural model

$$\begin{aligned} y_i &= \alpha_0 + Y_{2i}\beta_0 + \sigma(Z_i)u_i \,, \\ Y_{2i} &= \gamma_0 + \frac{c}{\sqrt{n}} f(Z_i) + \sigma(Z_i)v_{2i} \,. \end{aligned} \tag{6.12}$$

where c is a constant that controls the strength of the identification and Y_{2i} is univariate. The joint distribution of (u_i, v_{2i}) is a bivariate normal with mean $\mathbf{0}$, unit unconditional variances, and unconditional correlation ρ. We set $\alpha_0 = \beta_0 = \gamma_0 = 0$ and $\rho = 0.8$. We consider three different specifications for the function $f(\cdot)$: (i) a polynomial function of degree 3 proportional to $z - 2z^3/5$ (ii) a linear function, and (iii) a function compatible with first-stage group heterogeneity, see Abadie et al. (2016), proportional to $(2z_2 - 1)\,(z_1 - 2z_1^3/5)$. Here Z (or Z_1) is deterministic with values evenly spread between -2 and 2, and Z_2 follows a Bernoulli with probability 1/2. Also $f(Z)$ is centered and scaled to have variance one to make the different cases comparable. We consider heteroskedasticity depending on the first component of Z of the form

$$\sigma(z) = \sqrt{\frac{3(1+z^2)}{7}} \,.$$

We focus on the 10% asymptotic level tests for the slope parameter β_0. In all our experiments, $w(\cdot)$ is a triangle density, and conditional covariances are estimated through kernel smoothing with Gaussian kernel and rule-of-thumb bandwidth. We compare the performance of our two tests, ICM and the conditional ICM (CICM), to four inference procedures: the heteroskedasticity-robust S test proposed by Stock and Wright (2000), another heteroskedasticity-robust version of AR (CH) proposed by Chernozhukov and Hansen (2008), the heteroskedasticity-robust conditional LR (RCLR) proposed by Andrews et al. (2006), and the test by Jun and Pinkse (2012), see below for details on implementation. In homoskedastic models, we also considered the CLR test, which is known to have excellent power. We consider 5000 replications for each value under test, and 299 simulations for each replication to compute our tests' p-values.[6]

Polynomial Model (i). Our benchmark is the heteroskedastic version of the polynomial model, a degree of weakness $c = 3$, and a sample size $n = 101$, where the competitors of our tests use a linear reduced form. We consider in turn the following variations of our benchmark model: a homoskedastic version with $\sigma(x) = 1$; a sample size of 401; increasing the number of instruments to 3 and 7; finally, 3 IV with a sample size of 401. This represents a total of 6 versions of Model (i). In Table 1, we report the empirical sizes associated with the different inference procedures for these six versions of the model. In Figure 1, we display the power curves for different values β_1 when testing $H_0 : \beta = \beta_1$. Starting with the benchmark model, CH and RCLR are oversized, while ICM is undersized. Only ICM and CICM have good power, while all the other methods have trivial power. For the homoskedastic case, patterns in size and power are similar to the benchmark case. When increasing the sample size, the over-rejection of CH and RCLR disappears, but ICM and CICM are slightly undersized. Doubling the sample size does not improve the power properties of our competitors.
We now consider increasing the number of instruments to 3 and 7. We do this by fitting piecewise linear functions on intervals defined by the quartiles of Z. E.g. the three considered instruments are $\mathbb{I}(z \leq 0)$, $z \times \mathbb{I}(z \leq 0)$, and z. For JP, there is no automatic method to choose the number of neighbors k, so we set it such that the number of degrees of freedom, as measured by the trace of the smoothing matrix, equals the number of instruments used in other procedures. E.g. for three instruments, $k \approx n/3$.[7]

All tests now have good power, but size control deteriorates for CH, RCLR, and JP. For instance, the size of RCLR is 0.144 and 0.266 with 3 and 7 IV, respectively, whereas it is 0.107 for CICM. By contrast, size is well controlled for S, but at the cost of a smaller power. Depending on the number of instruments, S has comparable or lower power than ICM. Among the most powerful tests, only CICM controls size well. When increasing the sample size with 3 IV, we observe that CH and RCLR now have the right size, while JP is still oversized. The best power is obtained with RCLR and CICM.

[6] A supplementary appendix provides additional simulation results, where we vary endogeneity, sample size, and the design.

[7] With one linear instrument, we do not implement JP.

	ICM	CICM	CH	RCLR	S	JP
Polynomial Model (i)						
Homoskedastic	0.0644	0.1024	0.1180	0.1152	0.1086	n.a
Benchmark (Heter. 1 IV)	0.0844	0.1068	0.1168	0.1148	0.1068	n.a
Heter. 3 IV			0.1484	0.1442	0.1034	0.1840
Heter. 7 IV			0.2966	0.2658	0.0834	0.4260
Heter. 1 IV $n = 401$	0.0624	0.0888	0.0998	0.0986	0.0968	n.a
Heter. 3 IV $n = 401$			0.0982	0.1078	0.0872	0.1516
Linear Model (ii)						
Homoskedastic	0.0644	0.1120	0.1180	0.1152	0.1086	n.a
Benchmark (Heter. 1 IV)	0.0844	0.1302	0.1168	0.1148	0.1068	n.a
Heter. 3 IV			0.1484	0.1522	0.1034	0.1376
Heter. 7 IV			0.2966	0.2370	0.0834	0.4326
Stronger identif. 1 IV	0.0844	0.1334	0.1168	0.1148	0.1068	n.a
No identif. 1 IV	0.0844	0.1002	0.1168	0.1148	0.1068	n.a
Group Heterogeneity Model (iii)						
Benchmark (Heter. 3 IV)	0.1004	0.1050	0.1188	0.2806	0.0890	0.1138
Heter. 7 IV			0.1606	0.1866	0.0848	0.2654
Heter. 15 IV			0.3684	0.3260	0.0694	0.7106

Table 1: Empirical sizes associated with the different inference procedures for the three models and their different variations considered in Section 6 for a theoretical 10% level. Note: the sizes for ICM and CICM do not depend on the number of instruments and are only reported once for each case; JP is not implemented with one instrument.

Linear Model (ii). For a linear reduced form, the standard tests are known to possess good properties, so it is of interest to know how our tests comparatively behave in this context. Our benchmark version of this model is heteroskedastic, a degree of weakness $c = 3$, and a sample size $n = 101$, where the competitors of our test use the correct linear reduced form. We then consider the following variations of our benchmark model: the homoskedastic model; increasing the number of instruments to 3 and 7; increasing the value of c to get stronger identification; setting c to 0 to get no identification at all. This represents a total of 6 versions of Model (ii). Empirical sizes are reported in Table 1, and power curves are gathered in Figure 2.

Starting with the benchmark model, CH, RCLR, and CICM are somewhat oversized, S has a correct size, while ICM is undersized. ICM has least power, while all others have similar power. In the homoskedastic case, CICM has only slightly lower power than CLR. When increasing the number of instruments to 3 and 7, size control deteriorates for

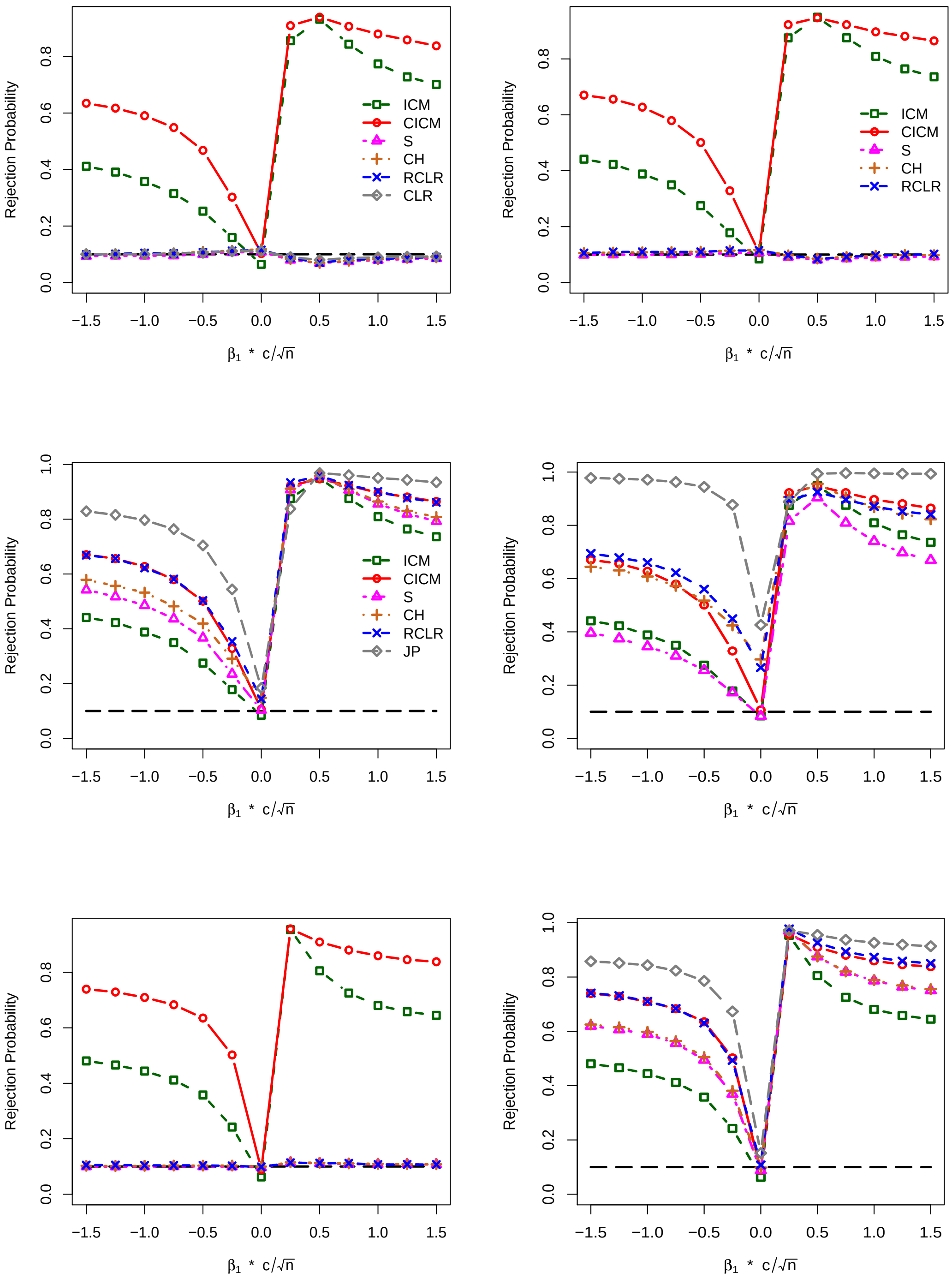

Figure 1: Power curves for Polynomial Model (i): homoskedastic case (top left), benchmark heter. 1 IV (top right), heter. 3 IV (middle left), heter. 7 IV (middle right), heter. 1 IV with sample size 401 (bottom left), and heter. 3 IV with sample size 401 (bottom right).

all our competitors but S. When increasing identification strength, all methods display similar power curves. In the case of no identification, the percentage rejection is constant whatever the value under test for all procedures. Classical tests are oversized, and ICM is undersized, while S and CICM maintain a 10% level across the board.

Group Heterogeneity Model (iii). This model is considered to investigate the behavior of the tests when we increase the number of instrumental variables. It also shows how the tests behave when one of the instrumental variables is discrete, which is quite common in applications. Abadie et al. (2016) consider this setup as empirical applications of instrumental variable estimators often involve settings where the reduced form varies depending on subpopulations. Our benchmark is the heteroskedastic version, a degree of weakness $c = 3$, and a sample size $n = 201$, where the competitors of our test use a reduced form with 3 instruments, namely the continuous Z_1, the discrete Z_2, and an interaction term. We then consider increasing the number of instruments to 7 and 15. We construct these instruments as piecewise linear and interaction terms on intervals defined by the quartiles of z_1. E.g. the seven considered instruments are $\mathbb{I}(z_1 \leq 0)$, $z_1 \times \mathbb{I}(z_1 \leq 0)$, z_1, $z_2 \times \mathbb{I}(z_1 \leq 0)$, z_2, $z_2 z_1 \times \mathbb{I}(z_1 \leq 0)$, $z_2 z_1$. Empirical sizes are reported in Table 1, and power curves are gathered in Figure 3. Starting with the benchmark model, the most powerful inference procedures are ICM and CICM, while the other methods have trivial power. In addition, both control size very well, while all other tests are oversized. When we increase the number of instruments to 7 and to 15, the size distortions mentioned for the competitors worsen, while CICM controls size well and is powerful.

Our results show that our tests are more powerful than competitors when the functional form of the link between instrumental variables and endogenous regressors is nonlinear. When trying to account for nonlinearities, the standard tests get size-distorted as more instruments are used under weak identification, as already noted by Jun and Pinkse (2012). This phenomenon may be linked to the heteroskedasticity-robust versions of the tests. Although these distortions disappear asymptotically, they are a concern in moderate-size samples. By contrast, our tests perform well with heteroskedasticity of unknown form. Overall, our two inference procedures have good power together with correct size control.

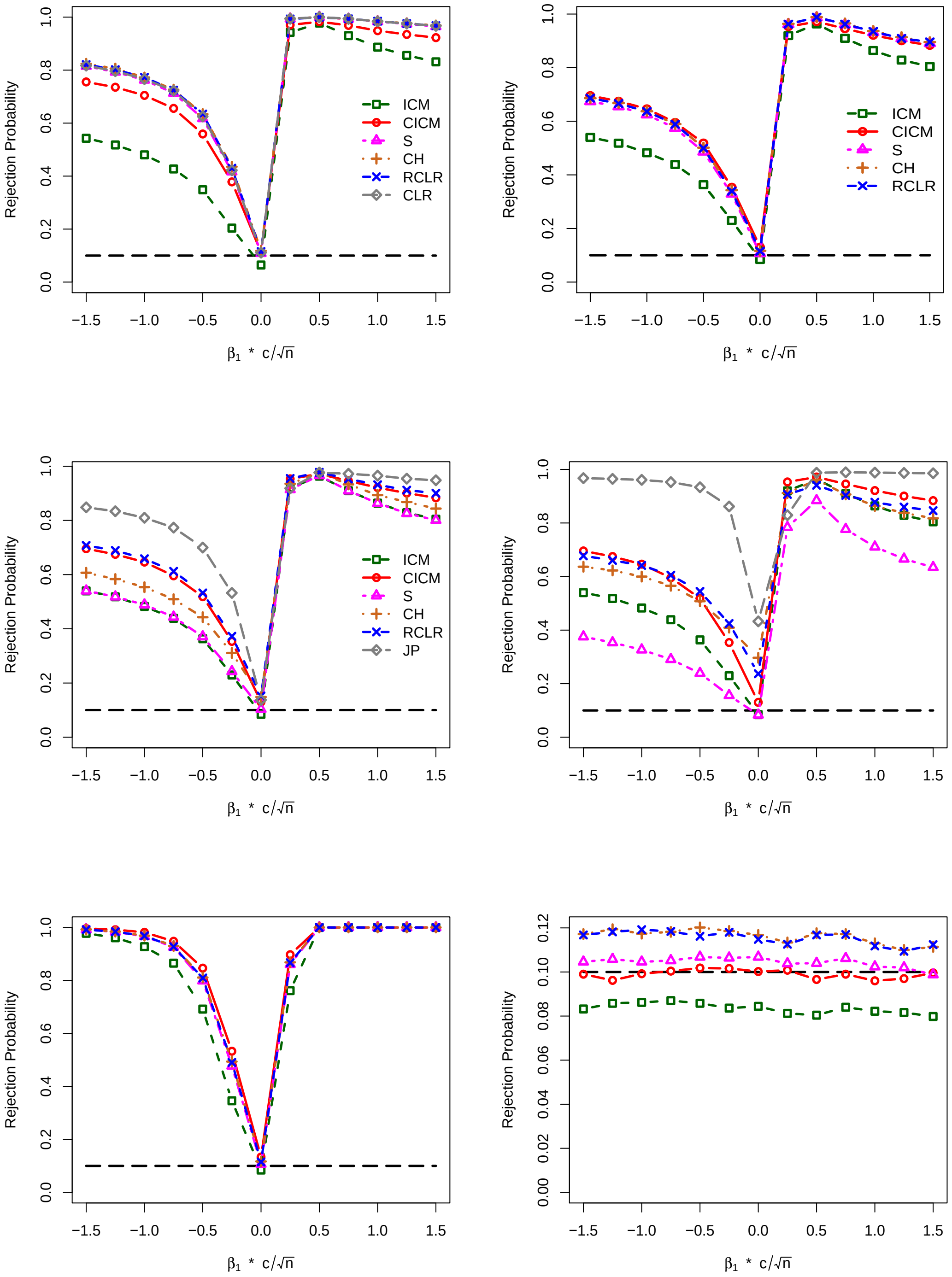

Figure 2: Power curves for Linear Model (ii): homoskedastic case (top left), benchmark heter. 1 IV (top right), heter. 3 IV (middle left), heter. 7 IV (middle right), stronger identification (bottom left) and no identification (bottom right).

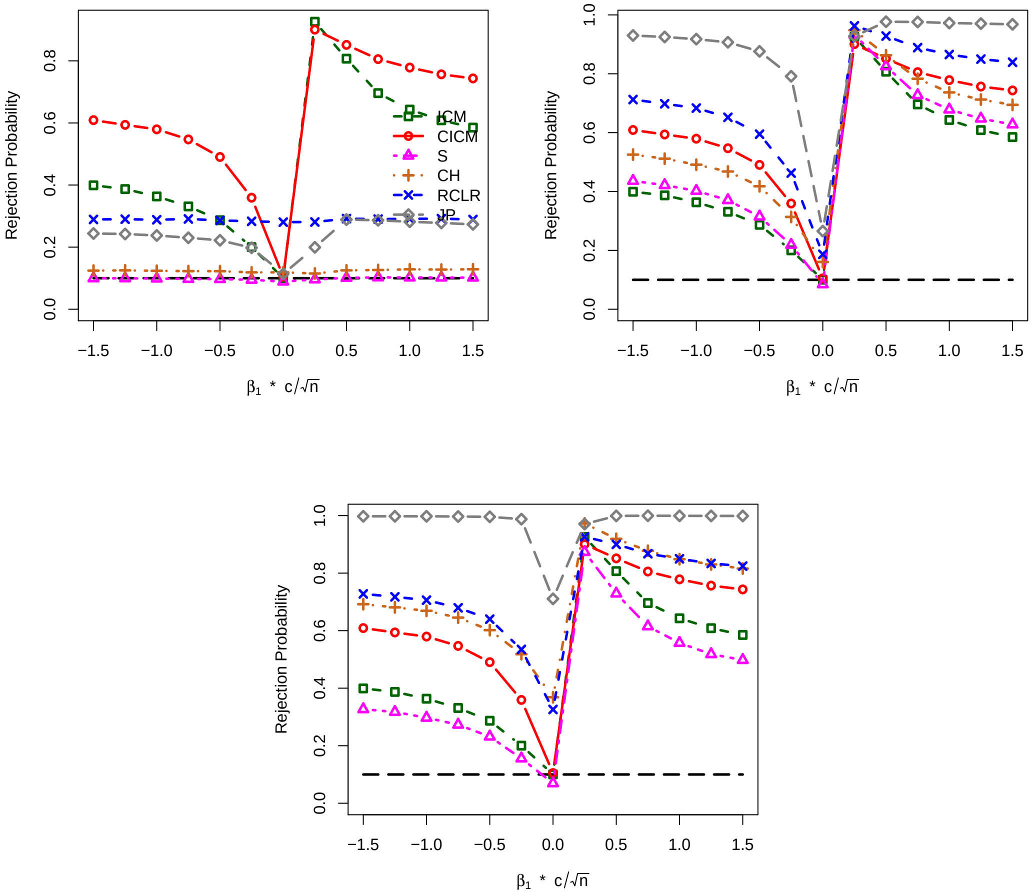

Figure 3: Power curves for Group Heterogeneity Model (iii): benchmark 1 IV (top left), 7 IV (top right), and 15 IV (bottom).

7 Empirical illustration: Mexico's 16th-century demographic collapse and the Hacienda

We extend some of the results presented in Sellars and Alix-Garcia (2018) who trace the impact of a large population collapse in 16th-century Mexico on land institutions through the present day. Such demographic collapse - which reduced the indigenous population by between 70 and 90 percent - is shown to have had a significant and persistent impact on Mexican land tenure and political economy by facilitating land concentration and the rise of a landowner class that dominated Mexican political economy for centuries. The authors adopt an instrumental-variables empirical strategy based on the characteristics of a massive epidemic in the mid-1570s which is believed to have been caused by a rodent-transmitted pathogen that emerged after several years of drought were followed by a period of above-average rainfall. Accordingly, proxies for these climate conditions are used as instrumental variables. Sellars and Alix-Garcia (2018) rely on the Palmer Drought Severity Index (PDSI), a normalized measure of soil moisture that captures deviations from typical conditions at a given location: their excluded instruments are, (i) *drought*, the sum of the 2 lowest consecutive PDSI values between 1570 and 1575 (more negative numbers indicate severe and prolonged drought), (ii) *rainfall*, the maximum PDSI between 1576 and 1580 (as a measure of excess rainfall), and (iii) *gap*, the difference between the minimum PDSI between 1570 and 1575 and the maximum between 1576 and 1580.

We focus here on the short-term effects of the population collapse: more specifically, the sharp decline in population lowered the costs and increased the benefits of acquiring land from indigenous villages in many areas. We used the data constructed in Sellars and Alix-Garcia (2018) to estimate the model

$$y_i = \beta_0 + \beta_1 Y_{2i} + \gamma' X_{1i} + u_i \,, \qquad \mathbb{E}\left(u_i | X_{1i}, X_{2i}\right) = 0$$

where y_i is the inverse hyperbolic sine of the percent rural population living in hacienda communities in 1900, Y_{2i} is the population decline in municipality i measured as the log ratio of 1650 and 1570 density, X_{2i} represents the vector of the 3 climate instruments, and X_{1i} is a vector of control variables of geographic features related to population and agriculture.[8]

[8] This specification corresponds to Column 6 in Table 2 in Sellars and Alix-Garcia (2018). It includes their full set of 12 control variables (the standard deviation of PDSI, a measure of maize productivity, various measures of elevation and slope) as well as the log of tributary density in 1570 and governorship-

Next, we present our main empirical results followed by a counterfactual analysis. They reveal a significant, negative, and economically relevant causal impact of the collapse of the population between 1570 and 1650 on the hacienda population. We also document first-stage heterogeneity and nonlinearities.

Main results. Our results are presented in Table 2, where we report the 95% confidence intervals for the parameter of population decline constructed from the 2 tests proposed in this paper, ICM and CICM. We also present confidence regions computed from two-stage-least squares (TSLS) and standard weak-identification robust inference procedures relying on a linear first-stage

$$Y_{2i} = \Pi X_{2i} + \delta' X_{1i} + v_i\,, \qquad \mathbb{E}\left(v_i | X_{1i}, X_{2i}\right) = 0\,. \tag{7.13}$$

With three instruments, the associated F-test statistic is moderate. Confidence intervals from TSLS, CLR, RCLR, JP, and CICM indicate a significant and negative impact of the log-ratio of 1650 to 1570 density on the dependent variable.[9] Hence a decrease in the ratio of 1650 to 1570 density increases the likelihood of having more large estates per area in 1900, in line with the results of Sellars and Alix-Garcia (2018). Confidence intervals obtained from ICM, AR, CH, and S tests are all empty. Since these are specification tests, this implies rejection of the model.

To address concerns about the validity of the instruments, we re-estimate the model using only the two most reliable of the three climate instruments, *drought* and *gap*.[10] Our results are reported in the second column of Table 2. The model is not rejected anymore by either ICM, AR, CH, or S, which all indicate a significant and negative impact of the log-ratio of 1650 to 1570 density on the dependent variable. Similar results are obtained with CLR, RCLR and CICM. ICM and CICM confidence intervals are wider: this is inline with the simulation results we obtained where both our tests control size while others are oversized.

level fixed effects, see their Sections 3 and 4 for a detailed description of the data and their identification strategy. The inverse hyperbolic sine transformation can be interpreted similarly to a log transformation and is preferable to it for a variety of reasons, see Burbage et al. (1988).

[9] Similarly to our simulation study, the number of neighbors k is chosen such that the degrees of freedom are equal to the number of instruments used in other procedures. E.g. for three instruments, $k \sim n/3$.

[10] The remaining climate instrument is added to the set of control variables. The first-stage equation (7.13) is updated accordingly.

	3 climate IV	**2 climate IV**	**2 climate IV and additional controls**
ICM	∅	[-2.16, -0.52]	∅
CICM	[-2.26, -0.99]	[-2.16, -0.58]	[-0.87, -0.50]
TSLS	[-1.77, -0.59]	[-1.30, -0.69]	
AR	∅	[-1.37, -0.65]	
CLR	[-1.52, -0.79]	[-1.32, -0.70]	
CH	∅	[-1.37, -0.68]	
RCLR	[-1.43, -0.68]	[-1.31, -0.72]	
S	∅	[-1.38, -0.68]	
JP	[-1.45, -0.51]	[-1.49, -0.62]	
F-stat	19.22	4.25	
Adj. R^2	0.21	0.05	

Table 2: 95% Confidence Intervals for the population collapse, using either the 3 climate instruments (column 1), 2 climate instruments (column 2), or 2 climate instruments with additional controls (column 3) over the full sample of size equal to 1030.

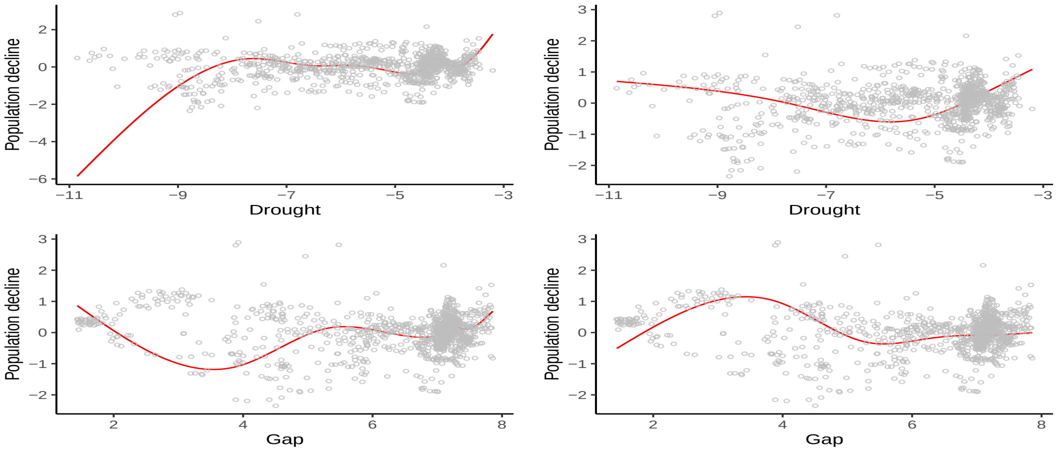

Figure 4: Generalized additive model explaining population collapse by a univariate function of the two reliable climate instruments, drought (top panel) and gap (bottom panel) by region for the 2 largest regions, NE (left panel) and NG (right panel).

We now document heterogeneity and nonlinearities in the relationship between the population decline and the two reliable climate instruments. We estimate a nonparametric additive model for each geographic region using the mgcv package in R, see Wood (2017). More specifically, after projecting out all the control variables, we estimate a generalized additive model explaining population collapse by a univariate function of each climate instrument by region; smoothing parameters are automatically selected by generalized cross-validation. Figure 4 plots the population collapse as a function of each instrument, drought (top panel) and gap (bottom panel), for the 2 largest regions, NE (left panel) and NG (right panel). Analysis of variance tests on models that replace in turn each function by a linear term reveal that relationships to drought and gap are indeed nonlinear, with p-values smaller than 2.10^{-4}.

Allowing for such regional heterogeneity and nonlinearities can easily be handled with our ICM and CICM procedures by varying the information set of the conditional moments. Focusing on the model estimated with the two reliable instruments (gap and drought), we report in the third column of Table 2 confidence intervals for ICM and CICM obtained using a larger information set: specifically, we augment the original conditioning information set (see also column 2 in Table 2) by adding regional dummies and all control variables to allow for nonlinearities in the first-stage, not only with respect to the instruments but also different first-stage specifications by region, as well as other nonlinearities with respect to control variables. Allowing for these extensions with other inference procedures (e.g. RCLR) is rather cumbersome as it would entail using a much larger number of moments, and, as documented in our simulation study, this may also be associated with size distortions. CICM still indicates a significant and negative impact of the ratio of 1650 to 1570 density on the dependent variable; the confidence interval is much narrower and is a subset of the confidence intervals obtained by all procedures reported in the second column of Table 2. However, the model is rejected by ICM.

Counterfactual analysis. To conclude, we revisit part of the counterfactual analysis in Sellars and Alix-Garcia (2018): we subtract off the predicted marginal effect of the population change in each municipality from the actual 1900 outcome to obtain what landholdings would be in the absence of a population collapse. We found that the distribution of hacienda population changes substantially under our counterfactual. The median percentage of 1900 population living in haciendas in our data is 16.7%. When we remove the effect of population collapse given by CICM in the third column of Table 2, it drops between 2.8% and 4.6%. The change is even larger at the 3rd quartile. In actual

1900 levels, 44.5% of the population lives on haciendas, but this drops between 9.0% and 13.6% with our counterfactual estimate. Such an impact is economically meaningful, practically relevant, and on par with the ones obtained from TSLS or RCLR.

Panel A: Changes for the median			**Panel B: Changes for the 3rd quartile**		
	Counterfactual	Difference		Counterfactual	Difference
CICM-low	2.8	13.9	CICM-low	9.0	35.5
CICM-up	4.6	12.1	CICM-up	13.6	30.9
IV	3.4	13.4	IV	10.3	34.2
CLR-low	1.8	14.9	CLR-low	6.9	37.5
CLR-up	5.4	11.3	CLR-up	15.9	28.6
RCLR-low	1.9	14.8	RCLR-low	7.0	37.4
RCLR-up	5.3	11.4	RCLR-up	15.4	29.1

Table 3: Counterfactual analysis of the causal impact of the demographic collapse: we report the predicted marginal effect of the population change in each municipality under ”Counterfactual” as well as the difference from the actual collapse under ”Difference”; the median and the 3rd quartile of the percentage of 1900 population living in haciendas are respectively 16.7% and 44.5% in our data.

Overall, our empirical study emphasizes the advantage of using an inference procedure such as CICM, that is robust to the presence of heteroskedasticity of unknown form and and does not necessitate to pin down the (potentially nonlinear) relationship between endogenous variable and instruments.

8 Proofs

8.1 Proof of Theorem 5.1

To simplify exposition, we consider the case where Ω is known and the statistic is based on $S = Yb_0\left(b_0'\Omega b_0\right)^{-1/2}$. It is easy to adapt our reasoning to account for a consistent estimator of Ω using Assumption C-(iv). However, we do not assume that the conditional variance $\Omega(\cdot)$ is known.

8.1.1 Uniform Convergence of Processes

The class of functions $\left\{s'Z, s \in \mathbb{R}^k\right\}$ has Vapnik-Červonenkis dimension $k+2$ and thus has bounded uniform entropy integral (BUEI). Since the functions $t \to \cos(t)$ and $t \to \sin(t)$ are bounded Lipschitz with derivatives bounded by 1, the class $\left\{\cos(s'Z), \sin(s'Z), s \in \mathbb{R}^k\right\}$ is BUEI, see Kosorok (2008, Lemma 9.13).

By Assumption B, the class $\mathcal{E}$ is BUEI. From Kosorok (2008, Theorem 9.15), the class $\left\{\Pi(Z)\cos(s'Z), \Pi(Z)\sin(s'Z), \Pi(\cdot) \in \mathcal{E}, s \in \mathbb{R}^k\right\}$ is BUEI, and from van der Vaart and Wellner (2000, Lemma 2.8.3)

$$\begin{pmatrix} n^{-1/2}\sum_{i=1}^n \left[\mathbb{E}\left(Y_i|Z_i\right)\cos(s'Z_i) - \mathbb{E}\left(Y\cos(s'Z)\right)\right] \\ n^{-1/2}\sum_{i=1}^n \left[\mathbb{E}\left(Y_i|Z_i\right)\sin(s'Z_i) - \mathbb{E}\left(Y\sin(s'Z)\right)\right] \end{pmatrix} \rightsquigarrow \begin{pmatrix} \mathbb{G}_1(s) \\ \mathbb{G}_2(s) \end{pmatrix},$$

uniformly in $P \in \mathcal{P}$ where $(\mathbb{G}_1'(\cdot), \mathbb{G}_2'(\cdot))$ is a vector Gaussian process with mean $\mathbf{0}$. Formally weak convergence uniform in P means that

$$\sup_{P\in\mathcal{P}} d_{BL}(\mathbb{G}_n, \mathbb{G}) \to 0 \quad \text{where} \quad d_{BL}(\mathbb{G}_n, \mathbb{G}) = \sup_{f\in BL_1}\left|\mathbb{E}\, f\left(\mathbb{G}_n\right) - \mathbb{E}\, f\left(\mathbb{G}\right)\right|$$

is the bounded Lipschitz metric, that is BL_1 is the set of real functions bounded by 1 and whose Lipschitz constant is bounded by 1. This implies that

$$n^{-1/2}\sum_{i=1}^n \left[\mathbb{E}\left(Y_i|Z_i\right)\exp(is'Z_i) - \mathbb{E}\left(Y\exp(is'Z)\right)\right] \rightsquigarrow \mathbb{G}_1(s) + \mathbb{G}_2(s) \tag{8.14}$$

Since $\mathbb{E}\left\|Y\right\|^{2+\delta} < \infty$, and because $\mathcal{E}$ is BUEI,

$$n^{-1/2}\sum_{i=1}^n \left(Y_i - \mathbb{E}\left(Y_i|Z_i\right)\right)\exp(is'Z_i) \rightsquigarrow \mathbb{G}_3(s) + \mathbb{G}_4(s) \tag{8.15}$$

Since $\Omega(\cdot)$ is a variance matrix with uniformly bounded elements, the functions $a'\Omega(\cdot)b$ for $\|a\|, \|b\| \le M$, and $\Omega \in \mathcal{O}$ satisfies

$$\left|a'\Omega_1(\cdot)b - a'\Omega_2(\cdot)b\right| \le \|a\|\|b\|\|\Omega_1 - \Omega_2\| \le M^2\|\Omega_1 - \Omega_2\|.$$

From Assumption C and Kosorok (2008, Lemma 9.13), these functions forms a BUEI class. Consider now the class of functions $\mathcal{B} = \{a'\Omega(\cdot)b/b'\Omega(\cdot)b, \|a\|, \|b\| \le M, \Omega \in \mathcal{O}\}$. Since the function $\phi(f,g) = f/g$ is Lipschitz for f, g uniformly bounded and g uniformly bounded away from zero, $\mathcal{B}$ is a BUEI class. Gathering results, for $B \in \mathcal{B}$

$$\mathbb{G}_n(B,s) = n^{-1/2}\sum_{i=1}^n B(Z_i)\left(Y_i - \mathbb{E}\left(Y_i|Z_i\right)\right)\exp(is'Z_i) \rightsquigarrow \mathbb{G}(B,s), \tag{8.16}$$

converges uniformly in $P \in \mathcal{P}$ to a centered Gaussian vector process. The joint uniform convergence of the processes in (8.14)–(8.16) follows.

Now let us show that replacing $\Omega(\cdot)$ by its estimator, or replacing $B(\cdot) = a'\Omega(\cdot)b/b'\Omega(\cdot)b$ by $\widehat{B}(\cdot) = a'\widehat{\Omega}(\cdot)b/b'\widehat{\Omega}(\cdot)b$, does not change the uniform weak limit of the process. From Assumption C-(iii) and (iv), it is sufficient to show that

$$\sup_{P\in\mathcal{P}} \Pr\left[\sup_{m\ge n}\sup_{s}\|\mathbb{G}_m(\widehat{B}_m, s) - \mathbb{G}_m(B,s)\|_{\mathcal{B}} > \varepsilon\right] \to 0 \qquad \forall\varepsilon > 0.$$

This follows as $\mathbb{G}_n(B,s)$ is asymptotically equicontinuous uniformly in P, see van der Vaart and Wellner (2000, Theorem 2.8.2).

8.1.2 Notations and Preliminary Results

For vector complex-valued functions $h_1(s)$ and $h_2(s)$, define the scalar product

$$\langle h_1, h_2\rangle = \frac{1}{2}\left(\int \left(\overline{h}_1'(s)h_2(s) + h_1'(s)\overline{h}_2(s)\right) d\mu(s)\right)$$

and the norm $\|h_1\| = \langle h_1, h_1\rangle^{1/2}$. Denote

$$h_{\beta_0,S}(s) \equiv n^{-1/2}\sum_{i=1}^{n} S_i \exp(is'Z_i)\,,$$

and note that $\|h_{\beta_0,S}\|^2 = S'WS$, so that we can write $\mathrm{ICM}(\beta_0) = \mathrm{ICM}(h_{\beta_0,S}) = \|h_{\beta_0,S}\|^2$. Let

$$h_{\beta_0,T}(s) \equiv n^{-1/2}\sum_{i=1}^{n} T_i \exp(is'Z_i)\,.$$

From (3.8), write $\mathrm{CICM}(\beta_0)$ as of a function of $h_{\beta_0,S}$ and $h_{\beta_0,T}$

$$\mathrm{CICM}(h_{\beta_0,S}, h_{\beta_0,T}) = \|h_{\beta_0,S}\|^2 - \min_{\|a\|=1} \|a_S h_{\beta_0,S} + a_T' h_{\beta_0,T}\|^2\,, \tag{8.17}$$

where $a = (a_S, a_T')'$.

Lemma 8.1 *Over the set $\{h : \|h\| \leq C\}$, (a)* $\mathrm{ICM}(h)$ *is bounded and Lipschitz continuous in h. (b)* $\mathrm{CICM}(h,g)$ *is bounded and Lipschitz continuous in (h,g).*

Proof. (a) Boundedness is trivial. For Lipschitz continuity,

$$\begin{aligned}|\mathrm{ICM}(h_1) - \mathrm{ICM}(h_2)| &= \left|\|h_1\|^2 - \|h_2\|^2\right| = |\langle h_1 - h_2, h_1 + h_2\rangle| \\ &\leq \|h_1 - h_2\|\|h_1 + h_2\| \leq \|h_1 - h_2\|(\|h_1\| + \|h_2\|) \leq 2\,C\,\|h_1 - h_2\|\,.\end{aligned}$$

(b) Since $0 \leq \mathrm{CICM}(h,g) \leq \mathrm{ICM}(h)$, boundedness follows. Let $a^* = (a_S^*, a_T^{*'})'$ be the value of a that optimizes (8.17). Let $a_i^*, i = 1,2$ be the value that optimizes $\mathrm{CICM}(h, g_i)$. Then

$$\begin{aligned}|\mathrm{CICM}(h,g_1) - \mathrm{CICM}(h,g_2)| &= \left|\min_{\|a\|=1}\|a_S h + a_T' g_1\|^2 - \min_{\|a\|=1}\|a_S h + a_T' g_2\|^2\right| \\ &\leq \max_{a\in\{a_1^*,a_2^*\}} \left|\|a_S h + a_T' g_1\|^2 - \|a_S h + a_T' g_2\|^2\right| \\ &= \max_{a\in\{a_1^*,a_2^*\}} \left|\langle a_T'\left(g_1 - g_2\right), \left(g_1 + g_2\right)' a_T + 2ha_S\rangle\right| \\ &\leq \max_{a\in\{a_1^*,a_2^*\}} \|a_T'\left(g_1 - g_2\right)\|\|\left(g_1 + g_2\right)' a_T + 2ha_S\| \\ &\leq \|g_1 - g_2\| \max_{a\in\{a_1^*,a_2^*\}} \|\left(g_1 + g_2\right)' a_T + 2ha_S\|\,.\end{aligned}$$

By definition, $\|ha^*_{1,S}+g_1'a^*_{1,T}\|^2 \leq \|h\|^2 \leq C^2$, and

$$\begin{aligned}\|\left(g_1+g_2\right)'a^*_{1,T}+2ha^*_{1,S}\| &\leq 2\|g_1a^*_{1,T}+ha^*_{1,S}\|+\|\left(g_1-g_2\right)'a^*_{1,T}\| \\ &\leq 2C+\|g_1-g_2\|\,,\end{aligned}$$

A similar inequality holds true for $a=a^*_2$. Hence

$$|\text{CICM}(h,g_1)-\text{CICM}(h,g_2)| \leq \|g_1-g_2\|\left(2C+\|g_1-g_2\|\right)\,.$$

If $\|g_1-g_2\| \leq 2C$, this yields the upper bound $4C\|g_1-g_2\|$, while if $\|g_1-g_2\| \geq 2C$,

$$|\text{CICM}(h,g_1)-\text{CICM}(h,g_2)| \leq 2C^2 \leq C\|g_1-g_2\|\,.$$

These results show that $\text{CICM}(h,g)$ is Lipschitz in g when $\{h:\|h\|\leq C\}$. Similarly, define now $a^*_i, i=1,2$ as the value that optimizes $\text{CICM}(h_i,g)$, then

$$\begin{aligned}&|\text{CICM}(h_1,g)-\text{CICM}(h_2,g)| \\ &\qquad=\left|\|h_1\|^2-\min_{\|a\|=1}\|a_Sh_1+a_T'g\|^2-\|h_2\|^2+\min_{\|a\|=1}\|a_Sh_2+a_T'g\|^2\right| \\ &\qquad\leq \left|\|h_1\|^2-\|h_2\|^2\right|+\max_{a\in\left\{a^*_1,a^*_2\right\}}\left|\langle a_S\left(h_1-h_2\right),a_S\left(h_1+h_2\right)+2g'a_T\rangle\right| \\ &\qquad\leq \langle h_1-h_2,h_1+h_2\rangle+2\max_{a\in\left\{a^*_1,a^*_2\right\}}\|a_S\left(h_1-h_2\right)\|\|a_S\left(h_1+h_2\right)+2g'a_T\| \\ &\qquad\leq 2\|h_1-h_2\|\left(C+\max_{a\in\left\{a^*_1,a^*_2\right\}}\|a_S\left(h_1+h_2\right)+2g'a_T\|\right)\,.\end{aligned}$$

Now

$$\begin{aligned}\|a^*_{1,S}\left(h_1+h_2\right)+2g'a^*_{1,T}\| &\leq 2\|a^*_{1,S}h_1+g'a^*_{1,T}\|+\|a^*_{1,S}\left(h_1-h_2\right)\| \\ &\leq 2C+\|h_1-h_2\|\,,\end{aligned}$$

and a similar inequality obtains for $a=a^*_2$. Hence

$$|\text{CICM}(h_1,g)-\text{CICM}(h_2,g)| \leq 2\|h_1-h_2\|\left(3C+\|h_1-h_2\|\right)\,.$$

Reason as above to conclude that $\text{CICM}(h,g)$ is Lipschitz in h over $\{h:\|h\|\leq C\}$. ■

Lemma 8.2 *Under Assumption A and D,*

$$\lim_{M\to\infty}\sup_{\beta_0}\sup_{P\in\mathcal{P}_{\beta_0}}\Pr\left[\text{ICM}(\beta_0)>M\right]\to 0\,.$$

Proof. By definition

$$\mathrm{ICM}(\beta_0) = S'WS = n^{-1}\sum_{i=1}^{n} S_i^2 w(0) + n^{-1}\sum_{i=1}^{n}\sum_{j\neq i} S_i S_j w(Z_i - Z_j)\,.$$

Hence, for some constants $C, C', C'' > 0$ independent of $P \in \mathcal{P}_{\beta_0}$ and of β_0,

$$\Pr\left[n^{-1}\sum_{i=1}^{n} S_i^2 w(0) > M/2\right] \leq 2w(0)\frac{\mathbb{E}\,S_1^2}{M} \leq \frac{C}{M}$$
$$\Pr\left[n^{-1}\sum_{i=1}^{n}\sum_{j\neq i} S_i S_j w(Z_i - Z_j) > M/2\right] \leq 4C'\frac{\mathbb{E}^{\,2}(S_1^2)}{M^2} \leq \frac{C''}{M}\,,$$

using the boundedness of $w(\cdot)$ and Markov's inequality. ■

8.1.3 ICM

Let $\mathcal{P}_{\beta_0} = \{P \in \mathcal{P} : \beta = \beta_0\}$. From (8.15),

$$h_{\beta_0,S}(s) \rightsquigarrow \mathbb{G}_S(s)\,, \tag{8.18}$$

uniformly in $P \in \mathcal{P}_{\beta_0}$ and in β_0, where $\mathbb{G}_S(s)$ is a centered complex Gaussian process. Let $\widehat{\Omega}_i = \widehat{\Omega}(Z_i)$ and

$$\widehat{G}_i = \left(b_0'\Omega b_0\right)^{-1/2}\left(b_0'\widehat{\Omega}_i b_0\right)^{1/2}\varepsilon_i\,,$$

where the ε_i are independent $N(0,1)$. From our results in Section 8.1.1,

$$h_{\widehat{G}}(s) = n^{-1/2}\sum_{i=1}^{n}\widehat{G}_i \exp(is'Z_i) \rightsquigarrow \mathbb{G}_S(s)\,,$$

uniformly in $P \in \mathcal{P}$. We say that $h_{\beta_0,S}$ *uniformly weakly converges* to $h_{\widehat{G}}$ in $P \in \mathcal{P}$, i.e.

$$\sup_{\beta_0}\ \sup_{P\in\mathcal{P}_{\beta_0}} d_{BL}(h_{\beta_0,S}, h_{\widehat{G}}) \to 0\,,$$

see Kasy (2018) for a similar terminology. Let $F(x) = \mathbb{I}\left[x < C_1\right] + \frac{C_2 - x}{C_2 - C_1}\mathbb{I}\left[C_1 \leq x \leq C_2\right]$ for some $0 < C_1 < C_2$ and consider the continuous truncation of $\mathrm{ICM}(h_S)$ defined by $\mathrm{ICM}_F(h_S) = \mathrm{ICM}(h_S)F(\|h_S\|)$. Consider the conditional quantile of $\mathrm{ICM}_F(h)$

$$c_{F,1-\alpha}(h) = \inf\left\{c : \Pr\left[\mathrm{ICM}_F(h) \leq c\right] \geq 1-\alpha\right\}\,.$$

Lemma 8.1 ensures that $\mathrm{ICM}_F(h)$ is Lipschitz, and it follows that $c_{F,1-\alpha}(h)$ is also Lipschitz. Indeed,

$$\begin{aligned}1-\alpha &\leq \Pr\left[\mathrm{ICM}_F(h_1) \leq c_{F,1-\alpha}(h_1)\right]\\ &\leq Pr\left[\mathrm{ICM}_F(h_2) \leq c_{F,1-\alpha}(h_1) + K\|h_1 - h_2\|\right]\,,\end{aligned}$$

so that $c_{F,1-\alpha}(h_2) \leq c_{F,1-\alpha}(h_1) + K\|h_1 - h_2\|$ for some constant $K > 0$. Interverting the role of h_1 and h_2 we get $c_{F,1-\alpha}(h_1) \leq c_{F,1-\alpha}(h_2) + K\|h_1 - h_2\|$, so $c_{F,1-\alpha}(h)$ is Lipschitz in h.

Assume now that the conclusion of Theorem 5.1 does not hold. Then there exists some $\delta > 0$, an infinitely increasing subsequence of sample sizes n_j and a sequence of probability measures $P_{n_j} \in \mathcal{P}_{\beta_{0,n_j}}$, with corresponding sequences of β_{0,n_j} and $\Pi_{n_j}(\cdot)$, such that

$$\Pr_{n_j}\left[\text{ICM}(h_{\beta_{0,n_j},S}) > c_{1-\alpha}(h_{\widehat{G}})\right] > \alpha + 3\delta \qquad \forall n_j\,.$$

Choose C_1 such that $\Pr_{n_j}\left[\text{ICM}(h_{\beta_{0,n_j},S}) \geq C_1\right] < \delta$, which is possible from Lemma 8.2. Now

$$\Pr\left[\text{ICM}(h_{\beta_0,S}) > x\right] \leq \Pr\left[\text{ICM}_F(h_{\beta_0,S}) > x\right] + \Pr\left[\text{ICM}(h_{\beta_0,S}) \geq C_1\right]$$

for any β_0 and any P_{β_0}, and $c_{F,1-\alpha}(h) \leq c_{1-\alpha}(h)$, so that

$$\Pr_{n_j}\left[\text{ICM}_F(h_{\beta_{0,n_j},S}) > c_{F,1-\alpha}(h_{\widehat{G}})\right] > \alpha + 2\delta \qquad \forall n_j\,.$$

As $\text{ICM}_F(h)$ is bounded and Lipschitz in h, by the uniform convergence of $h_{\beta_0,S}$ to $h_{\widehat{G}}$,

$$\sup_{\beta_0}\sup_{P\in\mathcal{P}_{\beta_0}}\sup_{x}\left|\Pr\left[\text{ICM}_F(h_{\beta_0,S}) > x\right] - \Pr\left[\text{ICM}_F(h_{\widehat{G}}) > x\right]\right| \to 0\,.$$

Therefore for n_j large enough

$$\Pr_{n_j}\left[\text{ICM}_F(h_{\widehat{G}}) > c_{F,1-\alpha}(h_{\widehat{G}})\right] \geq \alpha + \delta\,,$$

which contradicts the definition of $c_{F,1-\alpha}(h_{\widehat{G}})$.

8.1.4 CICM

Write now $h_{\beta_0,T} = h_{\beta_0,\tilde{S}} + h_{\beta_0,R} = h_{\beta_0,\tilde{S}} + h_{\beta_0,U} + h_{\beta_0,E}$, where

$$\tilde{S}_i = \left(A_0'\widehat{\Omega}^{-1}A_0\right)^{-1/2}\frac{A_0'\Omega^{-1}\widehat{\Omega}_i b_0}{b_0'\widehat{\Omega}_i b_0}Y_i'b_0, \quad R_i = T_i - \tilde{S}_i, \quad E_i = \mathbb{E}\left(T_i|Z_i\right), \quad U_i = R_i - E_i\,.$$

Denote by $E_{\beta_0}(s)$ the non-random function $n^{1/2}\mathbb{E}\left(Y\exp(is'Z)\right)$. Results in Section 8.1.1 show joint uniform weak convergence of $h_{\beta_0} = \left(h_{\beta_0,S}, h_{\beta_0,\tilde{S}}, h_{\beta_0,U}, h_{\beta_0,E} - E_{\beta_0}\right)$ to a Gaussian complex process with zero asymptotic covariance between the elements of $\left(h_{\beta_0,S}, h_{\beta_0,\tilde{S}}\right)$ and those of $(h_{\beta_0,U}, h_{\beta_0,E} - E_{\beta_0})$. Let

$$\tilde{G}_i = \left(A_0'\widehat{\Omega}^{-1}A_0\right)^{-1/2}\frac{A_0'\Omega^{-1}\widehat{\Omega}_i b_0}{b_0'\widehat{\Omega}_i b_0}\varepsilon_j,$$

where the ε_j are independent $N(0,1)$. Then $\widehat{h}_{\beta_0} = \left(h_{\widehat{G}}, h_{\tilde{G}}, h_{\beta_0,U}, h_{\beta_0,E} - E_{\beta_0}\right)$ uniformly weakly converges to h_{β_0}, i.e.

$$\sup_{\beta_0}\sup_{P\in\mathcal{P}_{\beta_0}} d_{BL}(h_{\beta_0}, \widehat{h}_{\beta_0}) \to 0\,.$$

Therefore $\left(h_{\beta_0,S}, h_{\beta_0,\tilde{S}}, h_{\beta_0,U}, h_{\beta_0,E}\right)$ uniformly weakly converges to $\left(h_{\widehat{G}}, h_{\tilde{G}}, h_{\beta_0,U}, h_{\beta_0,E}\right)$, because sequences of bounded Lipschitz functionals of h_{β_0} can be expressed as sequences of bounded Lipschitz functionals $h_{\beta_0} + (0,0,0,E_{\beta_0})$.

Consider the continuous truncation of $\mathrm{CICM}(h_S, h_T)$ defined by

$$\mathrm{CICM}_F(h_S, h_T) = \mathrm{CICM}(h_S, h_T) F(\|h_S\|) \,,$$

and the conditional quantile of CICM_F

$$c_{F,1-\alpha}(h,g) = \inf\left\{c : \Pr\left[\mathrm{ICM}_F(h,g) \leq c\right] \geq 1-\alpha\right\} \,.$$

Lemma 8.1 ensures that $\mathrm{CICM}_F(h,g)$ is bounded and Lipschitz in h and g, and it follows that $c_{F,1-\alpha}(h,g)$ is also Lipschitz.

Assume now that the conclusion of Theorem 5.1 does not hold. Then there exists some $\delta > 0$, an infinitely increasing subsequence of sample sizes n_j and a sequence of probability measures $P_{n_j} \in \mathcal{P}_{\beta_{0,n_j}}$, with corresponding sequences β_{0,n_j} and $\Pi_{n_j}(\cdot)$, such that

$$\Pr_{n_j}\left[\mathrm{CICM}(h_{\beta_{0,n_j},S}, h_{\beta_{0,n_j},\tilde{S}} + h_{\beta_{0,n_j},R}) > c_{1-\alpha}(h_{\widehat{G}}, h_{\tilde{G}} + h_{\beta_{0,n_j},R})\right] > \alpha + 3\delta \qquad \forall n_j \,.$$

Choose C_1 such that $\Pr_{n_j}\left[\mathrm{ICM}(h_{\beta_{0,n_j},S}) \geq C_1\right] < \delta$. Since for any β_0 and any P_{β_0}

$$\Pr\left[\mathrm{CICM}(h_{\beta_0,S}, h_{\beta_0,T}) > x\right] \leq \Pr\left[\mathrm{CICM}_F(h_{\beta_0,S}, h_{\beta_0,T}) > x\right] + \Pr\left[\mathrm{ICM}(h_{\beta_0,S}) \geq C_1\right]$$

and $c_{F,1-\alpha}(h_{\beta_0,S}, h_{\beta_0,T}) \leq c_{1-\alpha}(h_{\beta_0,S}, h_{\beta_0,T})$ for all h, g and β_0,

$$\Pr_{n_j}\left[\mathrm{CICM}_F(h_{\beta_{0,n_j},S}, h_{\beta_{0,n_j},\tilde{S}} + h_{\beta_{0,n_j},R}) > c_{F,1-\alpha}(h_{\widehat{G}}, h_{\tilde{G}} + h_{\beta_{0,n_j},R})\right] > \alpha + 2\delta \qquad \forall n_j \,.$$

Because $\mathrm{CICM}_F(h, g + h_R)$ is bounded and Lipschitz in (h,g) from Lemma 8.1,

$$\sup_{\beta_0} \sup_{P \in \mathcal{P}_{\beta_0}} \sup_{x} \left| \Pr\left[\mathrm{CICM}_F(h_{\beta_0,S}, h_{\beta_0,\tilde{S}} + h_{\beta_0,R}) > x\right] - \Pr\left[\mathrm{CICM}_F(h_{\beta_0,\widehat{G}}, h_{\beta_0,\tilde{G}} + h_{\beta_0,R}) > x\right] \right| \to 0 \,.$$

Therefore for n_j large enough

$$\Pr_{n_j}\left[\mathrm{CICM}_F(h_{\widehat{G}}, h_{\tilde{G}} + h_{\beta_{0,n_j},R}) > c_{F,1-\alpha}(h_{\widehat{G}}, h_{\tilde{G}} + h_{\beta_{0,n_j},R})\right] \geq \alpha + \delta \,,$$

which contradicts the definition of the quantile.

8.2 Proof of Theorem 5.2

Write

$$\mathrm{ICM}(\beta_1) = a' \begin{bmatrix} S' \\ T' \end{bmatrix} W \left[S, T\right] a \,,$$

with $a = (a_1,\ a_2')' = Qb_1\,(b_1'\Omega b_1)^{-1/2}$ and $Q = \left[(b_0'\Omega b_0)^{-1/2}\,b_0'\Omega \quad \left(A_0'\Omega^{-1}A_0\right)^{-1/2}A_0'\right]$. Since $\beta_1 \neq \beta_0$, $a_2 \neq 0$ and

$$\begin{aligned}\text{ICM}(\beta_1) - \text{ICM}(\beta_0) &= (a_1^2-1)S'WS + a_2'T'WT\,a_2 + 2a_1a_2'T'WS \\ &= (a_1^2-1)\|h_{\beta_0,S}\|^2 + 2\langle a_1h_{\beta_0,S}, a_2'h_{\beta_0,T}\rangle + \|a_2'h_{\beta_0,T}\|^2\,.\end{aligned}$$

(i) From our previous results, $\|h_{\beta_0,S}\|^2$ is uniformly bounded in probability. Moreover, under Assumption E-(i), $\|\tilde{c}_n^{-1}h_{\beta_0,T}(s) - \tilde{c}_n^{-1}\left(h_{\beta_0,E}(s) - E_{\beta_0}(s)\right)\|_\infty \xrightarrow{as} 0$ as $\tilde{c}_n \to \infty$ and

$$\|\tilde{c}_n^{-1}h_{\beta_0,E}(s) - \left(A_0'\Omega^{-1}A_0\right)^{-1/2}\mathbb{E}\,\left(A_0\Omega^{-1}C(Z)\exp(is'Z)\right)\|_\infty \xrightarrow{as} 0\,,$$

uniformly in $P \in \mathcal{P}_{\beta_0}$. Hence

$$\begin{aligned}\tilde{c}_n^{-2}\left(\text{ICM}(\beta_1) - \text{ICM}(\beta_0)\right) \xrightarrow{as} a_2'\left(A_0'\Omega^{-1}A_0\right)^{-1/2}A_0\Omega^{-1}\mathbb{E}\,&\left[C(Z_1)C(Z_2)\,w(Z_1-Z_2)\right] \\ &\Omega^{-1}A_0\left(A_0'\Omega^{-1}A_0\right)^{-1/2}a_2\,.\end{aligned}$$

By the arguments of Bierens (1982, Theorem 1), this is a positive definite matrix since

$$a'\mathbb{E}\,\left(C(Z_1)C(Z_2)\,w(Z_1-Z_2)\right)a \Rightarrow a = \mathbf{0} \quad \text{or} \quad C(Z) = \mathbf{0}\,,$$

but the last statement would contradict Assumption E-(i). Then

$$\lim_{\tilde{c}_n\to\infty}\ \sup_{P\in\mathcal{P}_{\beta_0}}\ \Pr\left[\text{ICM}(\beta_1) - \text{ICM}(\beta_0) > M\right] \to 1 \qquad \forall M > 0\,. \tag{8.19}$$

Assume now that the conclusion of Theorem 5.2 does not hold. Then there exists some $\delta > 0$, an infinitely increasing subsequence of sample sizes n_j and a sequence of probability measures $P_{n_j} \in \mathcal{P}_{\beta_0}$, with corresponding sequences $\Pi_{n_j}(\cdot)$ and $\tilde{c}_{n_j}$, such that

$$\Pr_{n_j}\left[\text{ICM}(\beta_1) < c_{1-\alpha}(h_{\widehat{G}})\right] > \delta \qquad \forall n_j\,.$$

Then

$$\Pr_{n_j}\left[\text{ICM}(\beta_1) - \text{ICM}(\beta_0) < c_{1-\alpha}(h_{\widehat{G}}) - \text{ICM}(\beta_0)\right] > \delta \qquad \forall n_j\,.$$

But ICM$(h_{\beta_0,S})$ is uniformly bounded in probability by Lemma 8.2 and so is the critical value $c_{1-\alpha}(h_{\widehat{G}})$. This contradicts (8.19).

For CICM, we can apply a similar reasonning because ICM(β_1) − ICM(β_0) = CICM(β_1) − CICM(β_0), $0 \leq$ CICM$(\beta_0) \leq$ ICM(β_0) is uniformly bounded, and thus its critical value is uniformly bounded as well.

(ii) Under Assumption E-(ii), we have a similar decomposition as above for ICM(β_{1n}) − ICM(β_0), with

$$\left(a_{1n},\ a_{2n}'\right) = \left[\left(b_0'\Omega b_0\right)^{-1/2}\left(b_0'\Omega b_0 + b_0'\Omega d\frac{\tilde{c}_n}{\sqrt{n}}\right) \quad \left(A_0'\Omega^{-1}A_0\right)^{-1/2}A_0'd\frac{\tilde{c}_n}{\sqrt{n}}\right]\left(b_{1n}'\Omega b_{1n}\right)^{-1/2}\,,$$

where $d = (0,\ \delta)$. Note that $A_0'd = \delta \neq 0$. We then proceed as above to obtain that $\tilde{c}_n^{-2}\left(\text{ICM}(\beta_1) - \text{ICM}(\beta_0)\right)$ converges to a positive definite limit. The rest of the proof follows similarly.

Acknowledgements We thank seminar participants at Amsterdam, Brown, Columbia, Georgia State, NYU, Ohio State, Penn State, Stonybrook, UC3M, ULB, Université de Montreal, University of Victoria, and at conferences (AMES, 2018; Econometrics Study Group, 2018; French Econometrics Conference, 2017; IAAE, 2018; CIREQ, 2017; Meeting in Econometrics, Toulouse, 2017; Tinbergen Institute, 2017) for helpful comments and discussions.

References

Abadie, A., J. Gu, and S. Shen (2016): "Instrumental Variable Estimation with First Stage Heterogeneity." Working paper, MIT.

Anderson, T. W. and H. Rubin (1949): "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *Ann Math Stat*, 20, 46–63.

Andrews, D. W. K. (1994): "Empirical Process Methods in Econometrics," in *Handbook of Econometrics*, Elsevier, vol. 4, 2247 – 2294.

——— (1995): "Nonparametric Kernel Estimation for Semiparametric Models," *Econometric Theory*, 11, 560.

Andrews, D. W. K. and X. Cheng (2012): "Estimation and Inference With Weak, Semi-Strong, and Strong Identification," *Econometrica*, 80, 2153–2211.

Andrews, D. W. K. and P. Guggenberger (2019): "Identification and Singularity-Robust Inference for Moment Condition Models," *Quant. Econ.*, 10, 1703–1746.

Andrews, D. W. K., V. Marmer, and Z. Yu (2019): "On Optimal Inference in the Linear IV Model," *Quant. Econ.*, 10, 457–485.

Andrews, D. W. K., M. J. Moreira, and J. H. Stock (2006): "Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression," *Econometrica*, 74, 715–752.

Andrews, D. W. K. and J. H. Stock (2007): "Inference with Weak Instruments," in *Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society*, Cambridge University Press, vol. Volume 3 of *Econometric Society Monograph Series*, chap. 8.

Andrews, I. (2016): "Conditional Linear Combination Tests for Weakly Identified Models," *Econometrica*, 84, 2155–2182.

Andrews, I. and A. Mikusheva (2016a): "Conditional Inference With a Functional Nuisance Parameter," *Econometrica*, 84, 1571–1612.

——— (2016b): "A Geometric Approach to Nonlinear Econometric Models," *Econometrica*, 84, 1249–1264.

Antoine, B. and P. Lavergne (2014): "Conditional Moment Models under Semi-Strong Identification," *J. Econometrics*, 182, 59–69.

Bierens, H. (1982): "Consistent Model Specification Tests," *J. Econometrics*, 20, 105–134.

Bierens, H. J. (1990): "A Consistent Conditional Moment Test of Functional Form," *Econometrica*, 58, 1443–1458.

Bierens, H. J. and W. Ploberger (1997): "Asymptotic Theory of Integrated Conditional Moment Tests," *Econometrica*, 65, 1129–1151.

Burbage, J. B., L. Magee, and A. L. Robb (1988): "Alternative Transformations to Handle Extreme Values of the Dependent Variable," *J. Amer. Statist. Assoc.*, 83, 123–7.

Cattaneo, M. D. and M. H. Farrell (2013): "Optimal Convergence Rates, Bahadur Representation, and Asymptotic Normality of Partitioning Estimators," *J. Econometrics*, 174, 127–143.

Chen, X., S. Lee, and M. H. Seo (2021): "Powerful Inference," *arXiv:2008.11140 [econ, math, stat]*, arXiv: 2008.11140.

Chernozhukov, V. and C. Hansen (2008): "The Reduced Form: A Simple Approach to Inference with Weak Instruments," *Econom. Lett.*, 100, 68 – 71.

Chernozhukov, V., C. Hansen, and M. Jansson (2009): "Admissible Invariant Similar Tests for Instrumental Variables Regression," *Econometric Theory*, 25, 806.

de Wet, T. and J. H. Venter (1973): "Asymptotic Distributions for Quadratic Forms with Applications to Tests of Fit," *Ann. Statist.*, 1, 380–387.

Dieterle, S. G. and A. Snell (2016): "A Simple Diagnostic to Investigate Instrument Validity and Heterogeneous Effects When Using a Single Instrument," *Labour Econ*, 42, 76–86.

Dominguez, M. A. and I. N. Lobato (2004): "Consistent Estimation of Models Defined by Conditional Moment Restrictions," *Econometrica*, 72, 1601–1615.

Dreier, I. and S. Kotz (2002): "A Note on the Characteristic Function of the T-Distribution," *Statistics & Probability Letters*, 57, 221 – 224.

Dufour, J.-M. (1997): "Some Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models." *Econometrica*, 65, 1365–1388.

——— (2003): "Identification, Weak Instruments, and Statistical Inference in Econometrics," *Can J Economics*, 36, 767–808.

Dufour, J.-M. and M. Taamouti (2005): "Projection-Based Statistical Inference in Linear Structural Models with Possibly Weak Instruments," *Econometrica*, 73, 1351–1365.

——— (2007): "Further Results on Projection-Based Inference in IV Regressions with Weak, Collinear or Missing Instruments," *J. Econometrics*, 139, 133–153.

Escanciano, J. C. (2018): "A simple and robust estimator for linear regression models with strictly exogenous instruments," *Econom. J.*, 21, 36–54.

Hahn, J. and J. Hausman (2003): "Weak Instruments: Diagnosis and Cures in Empirical Economics," *Am. Econ. Rev.*, 93, 118–125.

Han, C. and P. Phillips (2006): "GMM with Many Moment Conditions," *Econometrica*, 74, 147–192.

Hansen, C., J. Hausman, and W. Newey (2008): "Estimation With Many Instrumental Variables," *J. Bus. Econom. Statist.*, 26, 398–422.

Hausman, J. A., W. K. Newey, T. Woutersen, J. C. Chao, and N. R. Swanson (2012): "Instrumental Variable Estimation with Heteroskedasticity and Many Instruments," *Quant. Econ.*, 3, 211–255.

Johnson, N., S. Kotz, and N. Balakrishnan (1995): *Continuous Univariate Distributions*, vol. 2 of *Wiley series in probability and mathematical statistics: Applied probability and statistics*, Wiley & Sons.

Jun, S. J. and J. Pinkse (2012): "Testing Under Weak Identification with Conditional Moment Restrictions," *Econometric Theory*, 28, 1229–1282.

Kasy, M. (2018): "Uniformity and the Delta Method," *J. Econom. Methods*, 8.

Kleibergen, F. (2002): "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression," *Econometrica*, 70, 1781–1803.

——— (2005): "Testing Parameters in GMM Without Assuming that They Are Identified," *Econometrica*, 73, 1103–1123.

——— (2007): "Generalizing Weak Instrument Robust IV Statistics Towards Multiple Parameters, Unrestricted Covariance Matrices and Identification Statistics," *J. Econometrics*, 139, 181–216.

Kosorok, M. R. (2008): *Introduction to Empirical Processes and Semiparametric Inference*, Springer Series in Statistics, Springer-Verlag New York.

Lavergne, P. and V. Patilea (2013): "Smooth Minimum Distance Estimation and Testing with Conditional Estimating Equations: Uniform in Bandwidth Theory," *J. Econometrics*, 177, 47–59.

Mikusheva, A. (2010): "Robust Confidence Sets in the Presence of Weak Instruments," *J. Econometrics*, 157, 236 – 247.

Mikusheva, A. and L. Sun (2020): "Inference with Many Weak Instruments," Working paper, MIT.

Montiel Olea, J. L. (2020): "Admissible, Similar Tests: A Characterization," *Econometric Theory*, 36, 347–366.

Moreira, H. and M. J. Moreira (2019): "Optimal Two-Sided Tests for Instrumental Variables Regression with Heteroskedastic and Autocorrelated Errors," *J. Econometrics*, 213, 398 – 433.

Moreira, M. J. (2003): "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica*, 71, 1027–1048.

Moreira, M. J. and G. Ridder (2017): "Optimal Invariant Tests in an Instrumental Variables Regression With Heteroskedastic and Autocorrelated Errors," *Working paper, FPG.*

Newey, W. and F. Windmeijer (2009): "Generalized Method of Moments With Many Weak Moment Conditions," *Econometrica*, 77, 687–719.

Rice, J. (1984): "Bandwidth Choice for Nonparametric Regression," *Ann. Statist.*, 12, 1215–1230.

Robins, J. and A. van der Vaart (2006): "Adaptive Nonparametric Confidence Sets," *Ann. Statist.*, 34, 229–253.

Rotar', V. (1979): "Limit Theorems for Polylinear Forms," *J. Multivariate Anal.*, 9, 511 – 530.

Seifert, B., T. Gasser, and A. Wolf (1993): "Nonparametric Estimation of Residual Variance Revisited," *Biometrika*, 80, 373–383.

Sellars, E. A. and J. Alix-Garcia (2018): "Labor scarcity, land tenure, and historical legacy: Evidence from Mexico," *J Dev Econ*, 135, 504–516.

Staiger, D. and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

Stinchcombe, M. B. and H. White (1998): "Consistent Specification Testing With Nuisance Parameters Present Only Under the Alternative," *Econometric Theory*, 14, 295–325.

Stock, J. H. and J. H. Wright (2000): "GMM with Weak Identification," *Econometrica*, 68, 1055–1096.

Stock, J. H., J. H. Wright, and M. Yogo (2002): "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business and Economic Statistics*, 20, 518–529.

van der Vaart, A. (1994): "Bracketing Smooth Functions," *Stoch Proc Appl*, 52, 93–105.

van der Vaart, A. W. and J. A. Wellner (2000): *Weak Convergence and Empirical Processes: with Applications to Statistics*, New York: Springer.

WOOD, S. (2017): *Generalized Additive Models: An Introduction with R*, Boca Raton, FL: Chapman and Hall, 2nd ed.

YIN, J., Z. GENG, R. LI, AND H. WANG (2010): "Nonparametric Covariance Model," *Statist. Sinica*, 20, 469–479.

Supplementary Appendix

We provide here additional simulation results. Specifically, we consider the following variations of our benchmark cases for polynomial and linear models presented in Section 6: (a) a moderate level of endogeneity with $\rho = 0.3$ (in our original design $\rho = 0.8$); (b) a larger sample size $n = 1,001$ (in our original design $n = 101$); (c) Z normally distributed (in our original design Z is deterministic between -2 and 2). In Table 4, we present the empirical sizes associated with ICM, CICM, S, CH, RCLR and JP; in Figure 5, we present the corresponding power curves.

The main qualitative findings reported in Section 6 are not affected by the above-mentioned changes in our DGP. Increasing the sample size to $n = 1,001$ does not affect the power curves of any of the test procedures as they remain practically the same as those obtained when $n = 101$. This could be expected with weak identification. We notice, however, that ICM is slightly more undersized. When considering normally distributed instruments, power curves are virtually the same as with a fixed design; in addition, all tests but S are slightly oversized under the null. When we increase the sample size to $n = 1,001$, all procedures control size appropriately.

	ICM	CICM	S	CH	RCLR
Polynomial Model (i)					
$\rho = 0.3$	0.0844	0.1036	0.1068	0.1168	0.1148
$n = 1,001$	0.0606	0.0868	0.0944	0.0952	0.0962
$Z \sim \mathcal{N}(.)$	0.1286	0.1252	0.1004	0.1250	0.1264
$Z \sim \mathcal{N}(.)$ and $n = 1,001$	0.0838	0.1066	0.0964	0.0998	0.1004
Linear Model (ii)					
$\rho = 0.3$	0.0844	0.1182	0.1068	0.1168	0.1148
$n = 1,001$	0.0606	0.0926	0.0944	0.0952	0.0962
$Z \sim \mathcal{N}(.)$	0.1286	0.1310	0.1004	0.1250	0.1264
$Z \sim \mathcal{N}(.)$ and $n = 1,001$	0.0838	0.1060	0.0964	0.0998	0.1004

Table 4: Empirical sizes associated with 5 inference procedures for simulations designs (i) and (ii) and three variations for a theoretical 10% level.

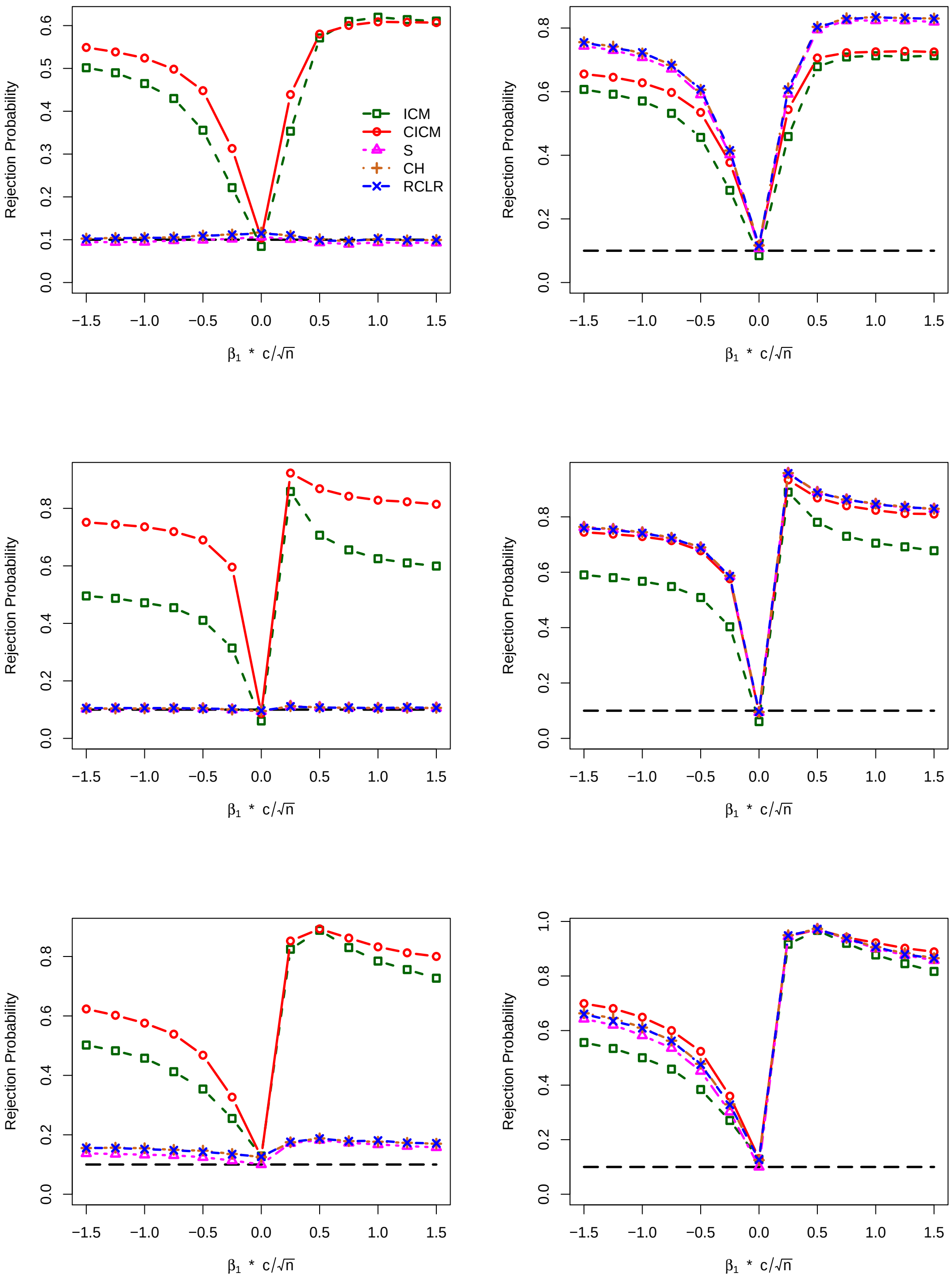

Figure 5: Power curves associated with 6 inference procedures for simulations designs (i) and (ii) and the three above-mentioned variations. Left: Polynomial Model (i); Right: Linear Model (ii). First row: Lower endogeneity $\rho = .3$. Second row: $n = 1,001$. Third row: Normal instruments.