Identification-Robust Nonparametric Inference in a Linear IV Model*

Bertille Antoine[†] and Pascal Lavergne[‡]

April 4, 2019

Abstract

For a linear IV regression, we propose two new inference procedures on parameters of endogenous variables that are easy to implement and robust to any identification pattern. Our tests do not rely on a linear first-stage equation, they are powerful irrespective of the particular form of the link between instruments and endogenous variables, and they account for heteroskedasticity of unknown form. Building on Bierens (1982), we first propose an Integrated Conditional Moment (ICM) type statistic constructed by using the value of the coefficient 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 the conditionality principle used by Moreira (2003) to condition on a set of ICM statistics that inform on identification strength. Our two procedures uniformly control size irrespective of identification strength and have non-trivial power under weak identification. They are competitive with existing procedures in simulations and applications.

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

*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. Bertille Antoine acknowledges financial support from SSHRC.

[†] *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, 21 Allées de Brienne, 31000 Toulouse FRANCE.

1 Introduction

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

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

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 (2015), Andrews (2016), and Andrews and Mikusheva (2016a,b). Surveys on weak identification issues include Stock, Wright, and Yogo (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 linear projection of endogenous variables on instruments. We argue that this feature can artificially create a weak identification (or no identification) issue. If linear projection does not capture enough of the variation of the endogenous variable, tests have little power and sometimes no more than trivial one. As practitionners typically have little prior information on the form of the relation between endogenous variables and instruments, it is impossible to know ex-ante if instruments are sufficiently strong. It is unfortunately not possible to use nonparametric estimated optimal instruments under weak identification, see Jun and Pinkse (2012). Indeed, if identification is not strong enough, the statistical variability of a nonparametric estimator will dominate the signal we aim to estimate. Therefore, finding a testing method that leaves 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 methods are based 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 fixing the value of the coefficient under test. 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. For both the ICM test and the *Conditional* ICM test, 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 in the case of weak identification. Since we remain agnostic on the functional relation between endogenous and instrumental variables in the first-stage equation, these properties are independent of its particular form.

Our conditional ICM test is related to Andrews and Mikusheva (2016a), since it is conditional upon a functional nuisance parameter, and our method of proof is similar. A key difference is that we consider conditional moment restrictions while they focus on unconditional ones. Our orthogonalization procedure also differs. We cannot claim that our procedures will be optimal when identification is strong as theirs is. This is because, as detailed below, optimality under strong identification relies on nonparametric optimal instruments, while using nonparametric estimation under weak identification cannot result in a powerful test. We do not address admissibility of our procedure nor optimality in terms of weighted average power, see Chernozhukov, Hansen, and Jansson (2009) and Olea (2018), and we make no claim that our tests are optimal. The test statistics are chosen for practical convenience and their resemblance with standard statistics used in the presence of weak instruments. We do not either address subvector inference, though it is always possible to adopt a projection approach, see Dufour (1997) and Dufour and Taamouti (2005).

In a series of simulations, we found that the level of our tests is well controlled using simulated critical values. 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, Moreira, and Stock (2006) and Andrews, Marmer, and Yu (2019). In addition, we consider two applications. First, we investigate 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). Second, we revisit the empirical study of Yogo (2004) on the elasticity of intertemporal substitution. Overall, we found that our two inference procedures were easy to implement, well-behaved and competitive.

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 detail the motivation for 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. 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 extend to more general setups including heteroskedasticity of unknown form. We then prove uniform asymptotic validity and study uniform power under semi-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 two empirical applications. Proofs are gathered in Section 8.

2 Framework 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 can be projected out through orthogonal projection, which does not influence our reasoning, but simplifies exposition. Hence, in what follows, we consider a structural equation of the form

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

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, of dimension l, and include X_2 but can also include X_1 if one suspects some nonlinearities in X_1 in the function $\Pi(\cdot)$.

Most work considers a linear projection of the form $Z'\Pi$. The concentration parameter

$$\frac{\Pi' Z Z' \Pi}{\sigma^2_{V_2}}$$

is a unitless measure of the strength of the instruments. A useful interpretation is in terms of the first-stage statistic F, the Fisher statistic for testing the hypothesis $\Pi = \mathbf{0}$, as in large samples $\mathbb{E}\, F - 1$ is approximately proportional to the concentration parameter. If one models weak identification as $\Pi = n^{-1/2}C$, the mean of the first-stage F statistic testing $\Pi = \mathbf{0}$ stays small or moderate for n large. The test 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 $Y b_0$ 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 assumptions of homoskedasticity. Since under linearity one can rewrite the structural equation as

$$y_i - Y_{2i}'\beta_0 = X_{2i}'\Delta + \varepsilon_i, \qquad \text{where} \quad \Delta = \Pi\,(\beta - \beta_0) \quad \text{and} \quad \varepsilon_i = u_i + V_{2i}\,(\beta - \beta_0) \, ,$$

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

$$\text{CLR} = \frac{b_0' Y' P_Z Y b_0}{b_0' \widehat{\Omega} b_0} - \min_b \frac{b' Y' P_Z Y b}{b' \widehat{\Omega} b} \, ,$$

and is derived as an approximate likelihood ratio test statistic for H_0 in the normal case by Moreira (2003).

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 omitting 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. But, a linear approximation can yield little power for the tests. As an example, assume $Z \sim N(0,1)$ 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

$$\min_{\pi_1} \mathbb{E}\, (\pi_1 Z - \Pi(Z))^2$$

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.

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. If in our example 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 even strong, i.e. $r_n = 1$, the quadratic approximation will pick up only the weakest quadratic relation. Hence an inadequate functional form may artificially create a weak identification issue.[2]

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

[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 picks up only the $\frac{1}{\sqrt{n}}$ component.

One may be tempted to estimate the reduced form nonparametrically, for instance by increasing the number of approximating polynomials with the sample size. But the local nature of nonparametric estimation yields a slower than $\sqrt{n}$ rate of convergence, so that statistical variability of the nonparametric estimator can exceed the signal to estimate if identification is not strong enough. As shown by Jun and Pinkse (2012), this issue can appear even with semi-strong identification and yields inflated variance or inconsistency for estimators based on a first-step nonparametric estimation. Similarly, weak (or not strong enough) identification prevents inference on β using nonparametrically generated instruments. Consider for instance the AR statistic based on nonparametric instruments $\widehat{\Pi}$. If $\Pi = C/r_n$, then

$$\widehat{\Pi} = \frac{C}{r_n} + \frac{\nu}{a_n},$$

where a_n is the rate of convergence of the estimator and ν is estimation noise, which may also include a bias term as usual in nonparametric estimation. Whenever $a_n = o(r_n)$, the denominator of AR becomes random and unrelated to Π as

$$b_0' Y' \widehat{\Pi} \left(\widehat{\Pi}' \widehat{\Pi} \right)^{-1} \widehat{\Pi} Y b_0 = b_0' Y' \nu \left(\nu' \nu \right)^{-1} \nu Y b_0 (1 + o_p(1)),$$

and the AR test will have no more than trivial power. So, while nonparametric optimal instruments should be used for efficiency under strong identification, they cannot be relied upon under weak or even semi-strong identification. Thus it does not appear possible to build a procedure that would be nonparametric with respect to the reduced form and would be at the same time robust to weak identification and optimal under strong identification. A solution might be to do a specification search for the best functional form of the reduced 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. Since typically little prior information is available on the link between the endogenous variable and the instruments, finding a testing method that leaves the first stage equation unspecified while being robust to weak identification seems extremely valuable from a practitioner's viewpoint.

3 ICM and Conditional ICM Tests Statistics

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

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

We consider testing

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

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

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

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

$$\int_{\mathbb{R}^q} |n^{-1/2}\sum_{i=1}^{n}(y_i - Y_{2i}'\beta_0)\exp(is'Z_i)|^2 \, d\mu(s) \,, \tag{3.5}$$

where μ is some symmetric probability measure with support $\mathbb{R}^q$ (except maybe a set of isolated points). The statistic (3.5) can be rewritten in matrix form as

$$b_0'Y'WYb_0 \,,$$

where W is a matrix with generic element $n^{-1}w\left(Z_i - Z_j\right)$ and

$$w(z) = \int_{\mathbb{R}^q} \cos(s'z)\, d\mu(s) \,.$$

The condition for μ to have support $\mathbb{R}^q$ 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, Kotz, and Balakrishnan (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) and Antoine and Lavergne (2014), 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 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.[3]

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

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

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

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.

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. This would also yield a collection of test statistics that could be used under weak identification. 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.

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

$$\text{ICM}(\beta_0) = \frac{b_0' Y' W Y b_0}{b_0' \widehat{\Omega} b_0} . \tag{3.6}$$

It is the ICM statistic that fixes the value of the parameter at β_0 and normalizes by an estimator of 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. 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. 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 second test is based on the statistic

$$\text{CICM}(\beta_0) = \frac{b_0' Y' W Y b_0}{b_0' \widehat{\Omega} b_0} - \min_b \frac{b' Y' W Y b}{b' \widehat{\Omega} b} . \tag{3.7}$$

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 CLR 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 $\text{ICM}(\beta_0) = \widehat{S}'W\widehat{S}$ and

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

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

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

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 will 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 $\widehat{T}$ with $S = Yb_0\left(b_0'\Omega b_0\right)^{-1/2}$ 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, Chernozhukov and Hansen (2008) adapt the Anderson-Rubin statistic using an heteroskedasticity-robust estimator of the covariance matrix.

We instead consider the same statistic ICM, but we allow for unknown heteroskedasticity when simulating critical values. We assume that we know the conditional variance function

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

so that we can compute $\Sigma = \mathrm{Var}(Yb_0|Z) = \mathrm{diag}\left(b_0'\Omega_1 b_0, \ldots, b_0'\Omega_n b_0\right)$. Then

$$\mathrm{ICM} = \frac{b_0'Y'\Sigma^{-1/2}\Sigma^{1/2}W\Sigma^{1/2}\Sigma^{-1/2}Yb_0}{b_0'\Omega b_0},$$

and ICM follows under H_0 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.

Heteroskedasticity-robust versions of the CLR have been proposed by Andrews et al. (2006) (in the working paper version of their article), Moreira and Moreira (2015), Moreira and Ridder (2017), Kleibergen (2007), and Andrews (2016). Andrews and Mikusheva (2016a) note that CLR could be used in heteroskedastic contexts by conditionning on the statistic of Kleibergen (2005), and more generally that a wide class of QLR tests are valid when conditionning on a nuisance process. Hence, we chose to work with the QLR-type statistic CICM, and to adapt critical values to heteroskedasticity. There may well be modified versions of the statistic that could account for heteroskedasticity, but they would not be of the form (3.7), and thus would not have the same intuitive interpretation.

The null distribution of CICM depends only of the asymptotic covariance structure of S and T conditional on Z under Lindeberg-type conditions, see Rotar' (1979). Under homoskedasticity, we have used the uncorrelation of S and T to simulate critical values. Under heteroskedasticity, S and T are not conditionally independent anymore. We can however condition on the part of T that is uncorrelated with S. Specifically, let

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

Then with normal errors S_i and R_i are conditionally jointly Gaussian and independent under H_0. Moreover R contains only information about $\Pi(\cdot)$, and none about β. We can simulate the distribution of CICM keeping R and Z fixed. We generate G_i, $i =, \ldots n$, as independent normal with mean 0 and variance $\mathrm{Var}(S_i|Z_i)$ for each i, and we compute CICM with drawings of G_i in place of S_i and

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

in place of T_i.

The above orthogonalisation 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. They thus 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 $n^{-1}\sum_{i=1}^{n} Z_i\left(y_i - Y_{2i}'\beta_0\right)$. This can be done by using their covariance function. The issue with CICM is similar but more intricate, as we are interested in the mean function $\mathbb{E}\left[(y - Y_2'\beta_0)\exp(is'Z)\right]$ for all s, and we consider as a nuisance parameter $\mathbb{E}\left[(y - Y_2'\beta)\exp(it'Z)\right]$ for all other values of β and all t. To orthogonalize the process $n^{-1}\sum_{i=1}^{n}\left(y_i - Y_{2i}'\beta\right)\exp(it'Z_i)$ with respect to the process $n^{-1}\sum_{i=1}^{n}\left(y_i - Y_{2i}'\beta_0\right)\exp(is'Z_i)$, 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

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

Lemma 4.1 *In the normal case with known $\Omega(\cdot)$,*

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

The ICM test is similar because $\Sigma^{-1/2}S \sim N(\mathbf{0}, \mathbf{I})$ conditionally on Z. The result for CICM 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, thus conditionally independent of, the components 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 remain asymptotically valid when errors are not Gaussian, and conditional variances are estimated instead of known.

5.1 Homoscedastic 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 quadractic form in S, and its asymptotic distribution depends only on the two first (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 confidence set obtained by inverting the ICM test is

$$\left\{\beta_0 : ICM(\beta_0) < c_{1-\alpha}\left(\beta_0, Z, \widehat{\Omega}(\cdot)\right)\right\} ,$$

where $c_{1-\alpha}(\beta_0, Z, \widehat{\Omega})$ is the $1-\alpha$ quantile of the statistic obtained by simulations. Under homoskedasticty, this critical value is independent of the particular value of β_0. When β_0 is scalar, ICM$\,(\beta_0)$ is a ratio of two quadratic forms in β_0, and the confidence set is obtained by solving a quadratic inequality, as is the AR confidence interval. We thus obtain as in Dufour and Taamouti (2005) and Mikusheva (2010) that it can be of four possible forms.

Lemma 5.1 *For homoskedastic errors, and when β is scalar, the asymptotic ICM confidence interval can have one of four possible forms:*

1. *a finite interval (β_1, β_2);*

2. *a union of two infinite intervals $(-\infty, \beta_2) \cup (\beta_1, +\infty)$;*

3. the whole real line $(-\infty, +\infty)$*;*

4. an empty set $\emptyset$*.*

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

$$\left\{\beta_0 : ICM(\beta_0) < c_{1-\alpha}(\beta_0, Z, \widehat{\Omega}(\cdot), \widehat{R}(\beta_0))\right\} ,$$

where $c_{1-\alpha}(\beta_0, Z, \widehat{\Omega}, \widehat{R}(\beta_0))$ 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) with the CLR.

5.2 Heteroskedastic Case

Accounting for unknown heteroskedasticity requires to estimate conditional variances of Y. One of our main tasks in the next section will be to establish asymptotic results accounting for estimation of $\Omega = \mathbb{E}\,\text{Var}(Y|Z)$ and $\Omega(\cdot) = \text{Var}(Y|Z = \cdot)$. 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, Gasser, and Wolf (1993) among others. The conditional variance 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, let us focus on kernel smoothing, which is used in our simulations and applications. 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 $e = (1, \dots 1)'$, 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, Geng, Li, and Wang (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 $\mathbb{E}\left(Y'Y\right)$ and then avoid preliminary estimation of $\mathbb{E}\left(Y|Z\right)$. Indeed $\mathbb{E}\left(S|Z\right) = 0$ a.s. under H_0 so that $\operatorname{Var}\left(S|Z\right) = \mathbb{E}\left(S^2|Z\right)$ and $\operatorname{Cov}\left(T, S|Z\right) = \mathbb{E}\left(T'S|Z\right)$.

With at hand a parametric or nonparametric estimator of $\Omega(\cdot)$, one can estimate the conditional variance of S_i by $\widehat{\operatorname{Var}}(S_i|Z_i) = b_0'\widehat{\Omega}_i b_0 \left(b_0\widehat{\Omega}b_0\right)^{-1}$, where $\widehat{\Omega}_i \equiv \widehat{\Omega}(Z_i)$. To approximate the asymptotic distribution of ICM $= S'WS$, we generate independent Gaussian $\widehat{G}_i$, $i =, \dots n$, with mean 0 and variance $\widehat{\operatorname{Var}}(S_i|Z_i)$ 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 $\widehat{R} = \left[\widehat{R}_1 \dots \widehat{R}_n\right]$, with

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

then recomputing CICM with drawings of G_i in place of $\widehat{S}_i$ and

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

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 semi-strong identification comes from 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 $\{Y_1, \dots Y_n\}$ the independent copies from Y for a sample size n.

Assumption B $\Pi(Z) = D_n^{-1} C(Z)$*, where D_n is a $l \times l$ matrix*

$$D_n = \begin{bmatrix} r_{1,n}\mathbf{I}_{p_1} & \mathbf{0} & \dots & \dots & \mathbf{0} \\ \mathbf{0} & r_{2,n}\mathbf{I}_{p_2} & \mathbf{0} & \dots & \mathbf{0} \\ \dots & & & & \dots \\ \dots & & & \mathbf{0} & r_{s,n}\mathbf{I}_{p_s} \end{bmatrix}, \qquad \sum_{j=1}^{s} p_s = l\,,$$

where $1 \leq r_{j,n}$ for all j and $C(\cdot)$ is a fixed matrix such that $\mathbb{E}\, C(Z)C'(Z)$ is bounded and positive definite.

Our condition on $C(\cdot)$ is an identifiability assumption. When it fails, the model only provides set identification, and we may possibly identify only some linear combinations of the coefficients, even under strong identification. We allow for different identification strengths across the different components of β ranging from strong, i.e. $r_n = 1$ to weak, i.e. $r_n = n^{1/2}$, and beyond. In practice, we do not need to know or estimate the matrix D_n.

Let $\mathcal{O}$ be a class of matrix-valued functions and let $N(\varepsilon, \mathcal{O}, L_2(Q))$ be the covering number of $\mathcal{O}$, that is the minimum number of $L_2(Q)$ ε-balls needed to cover $\mathcal{O}$, where an $L_2(Q)$ ε-ball around Ω is the set of matrix functions $\left\{h \in L_2(Q) \ : \ \int \|h - \Omega\|^2 \, dQ < \varepsilon\right\}$. We denote by $\mathcal{P}$ the class of distributions that fulfills our Assumption A as well as the following.

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*

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

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 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).

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 (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 accurate).

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 under the null hypothesis.

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

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

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

More general setup where $\Pi(\cdot)$ belongs to a set of smooth functions for the continuous components of Z would allow in particular for "localized" functions, which are identically zero but in the neighborhood of some separated points of the support of the continuous Z. In such a case, identification would come only from the behavior of the $\Pi(\cdot)$ around these points. While it is debatable whether this is relevant from an empirical viewpoint, such a setup would also raise technical issues, such as how to measure identification strength, and whether the set of ICM statistics $\widehat{T}'W\widehat{T}$ is suitable to evaluate identification strength. These issues would be very similar to the ones encountered when the marginal distribution of Z becomes concentrated on a few points. In most of the literature, with the exception of some examples discussed in Han and Phillips (2006), this possibility is implicitly ruled out by regularity assumptions.

5.4 Asymptotic Power

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

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

With β_0 the true value of β, we consider a test of $H_0 : \; \beta = \beta_1$ versus $H_1 : \; \beta \neq \beta_1$, where $\beta_1 \neq \beta_0$ is fixed. The object of interest is the asymptotic power of our two tests when $\tilde{c} \to \infty$.

Theorem 5.3 *Under Assumptions A, E, C and D, for any β_0 and any $\beta_1 \neq \beta_0$,*

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

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

The above result 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 ICM(β_1) is given by $\sum_{i=1}^{n} \lambda_i (G_i + c_i)^2$, 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. Our proof's strategy is different so as to encompass the study of our two tests. The behavior of CICM is indeed more involved because it depends on the behavior of the whole process ICM(β) for any β.

6 Small Sample Behavior

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

$$\begin{aligned} y_i &= \alpha_0 + Y_{2i}\beta_0 + \sigma(Z_i)u_i \,, \\ Y_{2i} &= \gamma_0 + \frac{c}{\sqrt{n}} f(Z_i) + \sigma(Z_i)v_{2i} \,. \end{aligned} \tag{6.11}$$

where c is a constant that controls the strength of the identification and Y_{2i} is univariate. The joint distribution of (u_{1i}, v_{2i}) is a bivariate normal with mean $\mathbf{0}$, unit unconditional variances, and unconditional correlation ρ. In all our simulations, $\alpha_0 = \beta_0 = \gamma_0 = 0$ and $\rho = 0.8$. We consider three different specifications for the function $f(\cdot)$: (i) a polynomial function of degree 3; (ii) a function compatible with first-stage group heterogeneity, see Abadie, Gu, and Shen (2016); (iii) a linear function. More specifically, we consider the following three cases, where each function is centered and standardized:

(i) $f(z) \propto z - 2z^3/5$

(ii) $f(z) \propto z$

(iii) $f(z_1, z_2) \propto (2z_2 - 1)(z_1 - 2z_1^3/5)$.

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. We consider heteroskedasticity depending on the first component of Z of the form

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

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 five inference procedures: the similar tests based on AR, K, and CLR; the conditional LR robust to heteroskedasticity (RCLR) proposed by Andrews et al. (2006); the robust version of AR (CH) proposed by Chernozhukov and Hansen (2008). Only CH and RCLR are robust to heteroskedasticity. We consider 5000 replications for each value under test, and 299 simulations to compute our tests' p-values.

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 form of the reduced form. We consider in turn the following variations of our benchmark model: an 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 7 inference procedures for these 6 versions of the model. In Figure 1, we display the power curves for different values in the null hypothesis for the parameter β.

Starting with the benchmark model, AR, K, and CLR are oversized without much surprise, as these tests are not robust to heteroskedasticity. On the other hand, CH and RCLR are oversized, while ICM is undersized. In terms of power, only ICM and CICM have excellent power properties; all the other methods have trivial power. For the homoskedastic case, AR, K, and CLR exhibit better size control as expected, they are oversized as CH ad RCLR are, while ICM is still undersized. The power curves are very similar to the benchmark case.

When increasing the sample size, the over-rejection of CH and RCLR disappear, but ICM and CICM are undersized. There is little improvement for AR, K, and CLR. Doubling the sample size does not improve the power properties of our competitors.

When increasing the number of instruments to 3 and 7, by fitting piecewise linear functions, size control deteriorates for RCLR and CH. All methods now have good power. The most powerful ones are CICM and RCLR, but RCLR does not control the size well: its size is 0.144 and 0.266 with 3 and 7 IV, respectively, instead of 0.107 for CICM. Increasing the sample size with 3 IV, we observe that CH and RCLR do control size well, and that the best power is obtained with RCLR and CICM.

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 (iii). Empirical sizes are reported in Table 1, and power curves are gathered in Figure 2.
Starting with the benchmark model, AR, K, and CLR are severely oversized, CH, RCLR, and CICM are somewhat oversized, while ICM is undersized. In terms of power, all methods have good power properties: the most powerful ones are AR and CLR, while CICM, RCLR, and CH are not far behind. In the homoskedastic model, the standard procedures have the highest power, but CICM is close by. When increasing the number of instruments to 3 and 7, fitting piecewise linear functions, size control deteriorates for RCLR and CH. When increasing identification, all the methods display similar power curves, while noticeable differences only relate to size control. 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 CICM maintains 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 show 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. 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 others tests are oversized. When we increase the number of instruments to 7 and to 15, the size distortions mentioned for the competitors worsen.

Our results show that our tests are more powerful than competitors when the functional form of the link between instrumental variables and endogneous regressors is nonlinear. When trying to account for nonlinearities, the standard procedures do not control size for small sample sizes. Our tests also perform well with heteroskedasticity of unknown form. Overall, our inference procedures have high power overall together with good size control.

7 Empirical illustrations

7.1 Short-term effects of Mexico's 16th-century demographic collapse

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 (namely measures of drought, rainfall abundance, and the difference between the two) 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) the sum of the 2 lowest consecutive PDSI values between 1570 and 1575 (more negative numbers indicate severe and prolonged drought), (ii) the maximum PDSI between 1576 and 1580 (as a measure of excess rainfall), and (iii) the difference between the former and the latter.

We focus here on the short-term effects of the above 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 use the data constructed in Sellars and Alix-Garcia[4] (2018) to estimate the following 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[5], Y_{2i} is the population decline in municipality i measured as the 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[6]. This specification corresponds to Column 6 in Table 2 in Sellars and Alix-Garcia.

Our results are presented in Panel A.1 of Table 2, where we report the 95% confidence intervals for the population decline constructed from the 2 tests proposed in this paper, ICM and the conditional ICM (CICM) estimated over the whole population (1030 observations). We also present corresponding confidence regions computed with two-stage-least squares (TSLS), as well as the 4 weak-identification robust inference procedures considered in our simulation study, AR, CLR, RCLR, and CH: all these inference procedures rely on the following (linear) first-stage equation,

$$Y_{2i} = \Pi Z_i + \delta' X_{1i} + v_i\,, \qquad \mathbb{E}\left(v_i | X_{1i}, Z_i\right) = 0 \tag{7.12}$$

where the vector of instruments Z_i corresponds either to the three above-mentioned climate instruments, or - to account for nonlinearities - to the first two powers of these three instruments with cross-products of order 2 (a total of 9 instruments), or the first three powers with cross-products of order 2 and 3 (a total of 18 instruments), or the first five powers of these three instruments with cross-products of order 2 and 3 (a total of 24 instruments).[7] We also report associated F-test statistics and adjusted R^2: while the F-test statistic drops significantly as the number of instrument increases, it remains

[4]See also their sections 3 and 4 for a detailed description of the data and their identification strategy.

[5]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, Magee, and Robb (1988).

[6]We follow Stellars and Alix-Garcia (2018) and include 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-level fixed effects.

[7]For each instruments $Z_{k,i}$, we consider orthogonalized polynomials which is key in practice to avoid multicollinearity.

moderate around 18 with 24 instruments; at the same time, the adjusted R^2 increases moderately from 0.22 to 0.29.

Our results indicate a significant and negative impact of the ratio of 1650 to 1570 density on the dependent variable: in other words, a decrease in the ratio of 1650 to 1570 density increases the likelihood of having more large estates per area in 1900. This is in line with the results of Sellars and Alix-Garcia (2018). It is interesting to mention that the confidence regions of CLR and RCLR vary substantially with the instrument set that is used: overall, as nonlinearities are accounted for through the use of powers of the original climate variables, the confidence regions become narrower as expected; however, a richer set of instruments does not always yield a confidence region that is a subset of the one obtained with a poorer set of instruments - in fact, these regions do not always overlap. Moreover, when comparing CLR and RCLR using 3 instruments, the two confidence regions are almost identical, suggesting that heteroskedasticity is not a concern in this application; however, when larger sets of instruments are considered to account for potential nonlinearities (see e.g. the case with 18 instruments), the two associated confidence regions are quite different, suggesting the presence of heteroskedasticity.

When comparing confidence regions obtained by ICM, AR, and CH, it is important to recall that these tests can be interpreted as specification tests of the model. In particular, an empty confidence interval can be interpreted as a rejection of the model: in other words, there does not exist a parameter value of the model that cannot be rejected. All models are rejected by ICM, AR, and CH which may suggest that the simplicity of the linear structural model may not be appropriate, that the instruments may not all be valid, or that the parameter values may be heterogenous over the population.
First, to mitigate concerns about the heterogeneity of the population, we re-estimate the model over the sub-population corresponding to the largest region (NE); our results are presented in Panel A.2 of Table 2. The model is still rejected by ICM, AR, and CH. In addition, the remaining inference procedures now report confidence regions that either contain zero, or that are infinitely large.
Second, we re-estimate the above model using only the most reliable of the three climate instruments - drought-rainfall gap - as done in Table A11 of Sellars and Alix-Garcia. The first-stage equation (7.12) is updated accordingly,

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

where the vector of instruments Z_i corresponds to the first k powers of the drought-rainfall gap instrument (with k taking values from 1 to 5). Our results are reported in

Panel B of Table 2, both on the whole population and the restricted population in region NE. Overall, the model is not rejected anymore by ICM, AR, or CH. However, most inferences procedures display some important differences when the model is estimated over the whole population, or over the restricted subpopulation. For instance, AR, CLR, CH, and RCLR indicate a significant and negative impact of the ratio of 1650 to 1570 density on the dependent variable when estimated over the whole population, while such effect cannot always be distinguished from zero, or is even sometimes positive when estimation is done over the restricted population. In addition, the results are quite sensitive to the exact form of the first-stage (e.g. the order of the polynomial approximation) when estimation is done over the restricted population, while it is not the case over the whole population.

Overall, our empirical study reveal some important practical messages that emphasize the advantage of using an inference procedure - such as CICM - that is robust to the presence of heteroskedasticity of unknown form and relies on the exogeneity of the instruments, without having to specify or pin down the (potentially nonlinear) relationship between endogenous variable and instruments.

7.2 Elasticity of Intertemporal Substitution (EIS)

We reproduce and extend some of the results presented by Yogo (2004), who studied instrumental variables estimation of the Elasticity of Intertemporal Substitution (EIS), considering the linearized Euler equation,

$$\Delta c_{t+1} = \nu + \psi r_{t+1} + u_{t+1}\,,$$

where ψ is the EIS, Δc_{t+1} the consumption growth at time $(t+1)$, r_{t+1} a real asset return at time $(t+1)$, ν a constant. We used the quarterly data for 11 countries used in Yogo (2004). The set of instrumental variables is composed of the nominal interest rate, inflation, consumption growth, and log dividend price-ratio that are lagged twice.

Results are gathered in Table 3, where we report the 95% confidence intervals for the EIS constructed from the following 7 inference procedures: the 2 tests proposed in this paper, ICM and CICM, TSLS as well as the 4 weak-identification robust inference procedures considered in our simulation study, AR, CLR, RCLR, and CH.[8] The weak-

[8]The confidence intervals based on the TSLS are not robust to weak identification and are presented for comparison purposes only.

identification robust confidence intervals indicate that the EIS is below 1, but small and not significantly different from 0 for most countries.

When comparing CICM confidence interval to the CLR confidence interval - which is known to be tighter due to the good power properties of CLR - CICM always delivers tighter bounds, but for Sweden (SWD). Focusing on CICM confidence intervals, the EIS is found to be less than 0.33 across all 11 countries. In addition, it is positive and significantly different from zero for the USA using both the long and short samples. Perhaps, more surprisingly, it is negative and significantly different from zero for Italy (ITA).

When comparing confidence intervals obtained through ICM, AR, and CH, it is important to recall that these tests can be interpreted as specification tests of the model. In particular, an empty confidence interval can be interpreted as a rejection of the model: in other words, there does not exist a parameter value of the model that cannot be rejected. Most models are rejected by ICM: noticeable exceptions include Switzerland (SWT) and France (FR) which cannot be rejected either by AR or CH. The fact that ICM rejects many more models than AR and CH can easily be understood since ICM has power against many more alternatives (e.g. against many nonlinear specifications of the model).

Focusing now on Switzerland (SWT), the only model that cannot be rejected by ICM and reveals an EIS that is significantly different from zero, we re-estimate the model using an extended set of instruments that contains the first two powers of the 4 instruments previously considered as well as their cross-products (for a total of 14 instruments). The results of the estimation of the model for SWT are presented in Table 4. When the first-stage accounts for quadratic nonlinearities, a significant and negative EIS is obtained by RCLR which is similar to ICM. Other inference procedures obtain a tighter confidence interval, except AR.

8 Proofs

8.1 Proof of Lemma 5.1

Let $\Gamma = W - c_{1-\alpha}\widehat{\Omega}$ with elements $\gamma_{i,j}$, $i, j = 1, 2$. The value of β_0 belongs to the confidence set if and only if $b_0'\Gamma b_0 = \gamma_{1,1} + 2\gamma_{1,2}\beta_0 + \gamma_{2,2}\beta_0^2 < 0$. Let $\Delta = \gamma_{1,2}^2 - \gamma_{1,1}\gamma_{2,2} = -\det\Gamma$. There are 4 cases:

1. If $\Delta > 0$ and $\gamma_{2,2} > 0$, the confidence set is (β_1, β_2), where
$$\beta_1 = \frac{-\gamma_{1,2} - \sqrt{\Delta}}{\gamma_{2,2}} \qquad \beta_2 = \frac{-\gamma_{1,2} + \sqrt{\Delta}}{\gamma_{2,2}} .$$
2. If $\Delta > 0$ and $\gamma_{2,2} < 0$, the confidence set is $(-\infty, \beta_2) \cup (\beta_1, +\infty)$.
3. If $\Delta < 0$ and $\gamma_{2,2} < 0$, the confidence set is the whole real line.
4. If $\Delta < 0$ and $\gamma_{2,2} > 0$, the confidence set is empty.

8.2 Proof of Theorem 5.2

To simplify exposition, we consider the case where Ω is known and the statistic is based on $S = Yb_0\,(b_0'\Omega b_0)^{-1/2}$. 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.

8.2.1 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). Hence
$$\sup_P \sup_s \|n^{-1} \sum_{i=1}^n C(Z_i) \exp(is'Z_i) - \mathbb{E}\, C(Z) \exp(is'Z)\| \xrightarrow{p} 0\,, \tag{8.14}$$
because $\|\mathbb{E}\, C(Z)C'(Z)\|^2 < \infty$. Since $\mathbb{E}\,\|Y\|^{2+\delta} < \infty$, we have by van der Vaart and Wellner (2000, Lemma 2.8.3) that
$$\begin{pmatrix} n^{-1/2} \sum_{i=1}^n (Y_i - \mathbb{E}\,(Y_i|Z_i)) \cos(s'Z_i) \\ n^{-1/2} \sum_{i=1}^n (Y_i - \mathbb{E}\,(Y_i|Z_i)) \sin(s'Z_i) \end{pmatrix} \rightsquigarrow \begin{pmatrix} \mathbb{G}_1(s) \\ \mathbb{G}_2(s) \end{pmatrix},$$
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
$$\sup_{P \in \mathcal{P}} d_{BL}(\mathbb{G}_n, \mathbb{G}) \to 0 \quad \text{where} \quad d_{BL}(\mathbb{G}_n, \mathbb{G}) = \sup_{f \in BL_1} |\mathbb{E}\, f(\mathbb{G}_n) - \mathbb{E}\, f(\mathbb{G})|$$
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
$$n^{-1/2} \sum_{i=1}^n (Y_i - \mathbb{E}\,(Y_i|Z_i)) \exp(is'Z_i) \rightsquigarrow \mathbb{G}(s) = \mathbb{G}_1(s) + \mathbb{G}_2(s) \tag{8.15}$$

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

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

From Assumption C and Kosorok (2008, Lemma 9.13), these functions forms a BUEI class. Consider now the class of functions $\mathcal{B} = \{a'\Omega(\cdot)b/b'\Omega(\cdot)b, \|a\|, \|b\| \leq M, \Omega \in \mathcal{O}\}$. 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 (8.15) and (8.16) follows.

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

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

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).

8.2.2 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 $\text{ICM}(\beta_0) = \text{ICM}(h_{\beta_0,S}) = \|h_{\beta_0,S}\|^2$. Let

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

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

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

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

Lemma 8.1 *Over the set* $\{h : \|h\| \leq C\}$*, (a)* ICM(h) *is bounded and Lipschitz continuous in* h*. (b)* 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 $a^* = (a_S^*, a_T^{*\prime})'$ be the value of a that optimizes (8.17). Let $a_i^*, i = 1, 2$ be the value that optimizes CICM(h, g_i). Then

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

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

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

A similar inequality holds true for $a = a_2^*$. Hence

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

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

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

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

$$
\begin{aligned}
&|\text{CICM}(h_1,g) - \text{CICM}(h_2,g)| \\
&\qquad = \left| \|h_1\|^2 - \min_{\|a\|=1} \|a_S h + a_T' g_1\|^2 - \|h_2\|^2 + \min_{\|a\|=1} \|a_S h + a_T' g_2\|^2 \right| \\
&\qquad \le \left| \|h_1\|^2 - \|h_2\|^2 \right| + \max_{a \in \left\{a_1^*, a_2^*\right\}} \left| \langle a_S (h_1 - h_2), a_S (h_1 + h_2) + 2g' a_T \rangle \right| \\
&\qquad \le \langle h_1 - h_2, h_1 + h_2 \rangle + 2 \max_{a \in \left\{a_1^*, a_2^*\right\}} \|a_S (h_1 - h_2)\| \|a_S (h_1 + h_2) + 2g' a_T\| \\
&\qquad \le 2\|h_1 - h_2\| \left(C + \max_{a \in \left\{a_1^*, a_2^*\right\}} \|a_S (h_1 + h_2) + 2g' a_T\| \right) .
\end{aligned}
$$

Since

$$
\begin{aligned}
\|a_{1,S}^* (h_1 + h_2) + 2g' a_{1,T}^*\| &\le 2\|a_{1,S}^* h_1 + g' a_{1,T}^*\| + \|a_{1,S}^* (h_1 - h_2)\| \\
&\le 2C + \|h_1 - h_2\| ,
\end{aligned}
$$

and a similar inequality obtains for $a = a_2^*$,

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

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

Lemma 8.2 *Under Assumption A and D,* $\lim_{M\to\infty} \sup_{\beta_0} \sup_{P\in\mathcal{P}_{\beta_0}} \Pr[\text{ICM}(\beta_0) > M] \to 0$.

Proof. By definition

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

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

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

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

8.2.3 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 Gaussian process. Let $\widehat{\Omega}_i = \widehat{\Omega}(Z_i)$ and $\widehat{G}_i = (b_0'\Omega b_0)^{-1/2}\left(b_0'\widehat{\Omega}_i b_0\right)^{1/2}\varepsilon_i$, where the ε_i are independent standard Gaussian. From (8.16),

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

uniformly in $P \in \mathcal{P}$. We now follow the terminology of Kasy (2018) and say that $h_{\beta_0,S}$ converges in distribution to $h_{\widehat{G}}$ as

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

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 $\text{ICM}(h_S)$ defined by $\text{ICM}_F(h_S) = \text{ICM}(h_S)F(\|h_S\|)$. Consider the conditional quantile of ICM_F

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

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

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

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 5.2 does not hold. Then there exists some $\delta > 0$, an infinitely increasing subsequence of sample sizes n_j, a sequence of probability measures $P_{n_j} \in \mathcal{P}_{\beta_{0,n_j}}$ with corresponding sequence of β_{0,n_j} 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. Since for any β_0

$$\Pr\left[\mathrm{ICM}(h_{\beta_0,S}) > x\right] \le \Pr\left[\mathrm{ICM}_F(h_{\beta_0,S}) > x\right] + \Pr\left[\mathrm{ICM}(h_{\beta_0,S}) \ge C_1\right]$$

and $c_{F,1-\alpha}(h) \le c_{1-\alpha}(h)$,

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

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

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

Therefore for n_j large enough

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

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

8.2.4 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

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

From our previous results, we have joint uniform weak convergence of $\left(h_{\beta_0,S}, h_{\beta_0,\tilde{S}}, h_{\beta_0,U}\right)$ to a Gaussian complex process, with zero asymptotic covariance between $\left(h_{\beta_0,S}, h_{\beta_0,\tilde{S}}\right)$ and $h_{\beta_0,U}$. Moreover

$$n^{-1/2}D_n h_{\beta_0,E}(s) = \left(A_0'\Omega^{-1}A_0\right)^{-1/2}\left[n^{-1}\sum_{i=1}^{n}A_0'\Omega^{-1}C(Z_i)\exp(is'Z_i)\right]$$

$$\sup_{P\in\mathcal{P}}\|n^{-1/2}D_n h_{\beta_0,E}(s) - L(\beta_0,s)\|_\infty \xrightarrow{as} 0$$

$$\text{with} \quad L(\beta_0,s) = \left(A_0'\Omega^{-1}A_0\right)^{-1/2}\mathbb{E}\left(A_0\Omega^{-1}C(Z)\exp(is'Z)\right)\,,$$

by (8.15). 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 standard normal. Then $\left(h_{\widehat{G}}, h_{\tilde{G}}, h_{\beta_0,U}, n^{-1/2}D_n h_{\beta_0,E}\right)$ has the same joint uniform weak limit as $\left(h_{\beta_0,S}, h_{\beta_0,\tilde{S}}, h_{\beta_0,U}, n^{-1/2}D_n h_{\beta_0,E}\right)$. Moreover the components of $n^{-1/2}D_n$ have their limits in $\mathbb{R}_+^l \cup +\infty$.

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

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

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 5.2 does not hold. Then there exists some $\delta > 0$, an infinitely increasing subsequence of sample sizes n_j, and a sequence of probability measures $P_{n_j} \in \mathcal{P}_{\beta_{0,n_j}}$ 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 $\Pr_{n_j} \left[\mathrm{ICM}(h_{\beta_{0,n_j},S}) \ge C_1 \right] < \delta$. Since for any β_0

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

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

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

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

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

Therefore for n_j large enough

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

which contradicts the definition of the quantile.

Proof of Theorem 5.3

Write

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

with $a = (a_1 a_2')' = Q b_1 \left(b_1' \Omega b_1\right)^{-1/2}$ and

$$Q = \left[\left(b_0' \Omega b_0\right)^{-1/2} b_0' \Omega \quad \left(A_0' \Omega^{-1} A_0\right)^{-1/2} A_0'\right] .$$

Since $\beta_1 \neq \beta_0$, $a_2 \neq 0$ and

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

From our previous results, $\|h_{\beta_0,S}\|$ is uniformly bounded, $\|\tilde{c}_n^{-1} h_{\beta_0,T}(s) - \tilde{c}_n^{-1} h_{\beta_0,E}(s)\|_\infty \xrightarrow{as} 0$ as $\tilde{c}_n \to \infty$, and

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

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

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

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 conclusion would contradicts Assumption E. Then

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

Assume now that the conclusion of Theorem 5.3 does not hold. Then there exists some $\delta > 0$, an infinitely increasing subsequence of sample sizes n_j, a sequence of probability measures $P_{n_j} \in \mathcal{P}_{\beta_0}$ and a corresponding sequence $\tilde{c}_{n_j}$ 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 $\text{ICM}(h_{\beta_0,S})$ is uniformly bounded in probability by Lemma 8.2 and so is the critical value $c_{1-\alpha}(h_{\widehat{G}})$, and this contradicts (8.19).

For CICM, we can apply a similar reasonning because $\text{ICM}(\beta_1) - \text{ICM}(\beta_0) = \text{CICM}(\beta_1) - \text{CICM}(\beta_0)$, $0 \leq \text{CICM}(\beta_0) \leq \text{ICM}(\beta_0)$ is uniformly bounded, and thus its critical value is uniformly bounded as well.

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," *Annals of Mathematical Statistics*, 20, 46–63.

Andrews, D., V. Marmer, and Z. Yu (2019): "A Note on Optimal Inference in the Linear IV Regression Model." *Quantitative Economics, forthcoming*, cowles Foundation Discussion Papers 2073, Cowles Foundation for Research in Economics, Yale University.

Andrews, D. W. (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 (2015): "Identification- and Singularity-Robust Inference for Moment Condition," Cowles Foundation Discussion Papers 1978, Cowles Foundation for Research in Economics, Yale University, Working Paper, Yale University.

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," *Journal of 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," *Journal of the American Statistical Association*, 83, 123–7.

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

Chernozhukov, V. and C. Hansen (2008): "The Reduced Form: A Simple Approach to Inference with Weak Instruments," *Economics Letters*, 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.

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," *Canadian Journal of 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," *Journal of Econometrics*, 139, 133–153.

Hahn, J. and J. Hausman (2003): "Weak Instruments: Diagnosis and Cures in Empirical Economics," *American Economic Review*, 93, 118–125.

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

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," *Journal of Econometric 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," *Journal of 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," *Journal of Econometrics*, 177, 47–59.

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

Moreira, H. and M. J. Moreira (2015): "Optimal Two-Sided Tests for Instrumental Variables Regression with Heteroskedastic and Autocorrelated Errors," Working paper, FPG, arXiv: 1505.06644.

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.

Olea, J. L. M. (2018): "Admissible, Similar Tests: A Characteization," *Working Paper*.

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

Rotar', V. (1979): "Limit Theorems for Polylinear Forms," *Journal of Multivariate Analysis*, 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," *Journal of Development Economics*, 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. W. and J. A. Wellner (2000): *Weak convergence and empirical processes: with applications to statistics*, New York: Springer, oCLC: 45749647.

Yin, J., Z. Geng, R. Li, and H. Wang (2010): "Nonparametric Covariance Model," *Statistica Sinica*, 20, 469–479.

Yogo, M. (2004): "Estimating the Elasticity of Intertemporal Substitution when instruments are weak." *Review of Economics and Statistics*, 86, 797–810.

	AR	K	CLR	CH	RCLR	ICM	CICM
Polynomial Model (i)							
Benchmark	0.1874	0.1874	0.1850	0.1168	0.1148	0.0844	0.1068
Homoscedastic	0.1104	0.1104	0.1112	0.1180	0.1152	0.0644	0.1024
Sample size 401	0.1672	0.1672	0.1678	0.0998	0.0986	0.0624	0.0888
3 IV	0.1426	0.0646	0.0854	0.1484	0.1442	0.0844	0.1068
7 IV	0.1030	0.1116	0.1130	0.2966	0.2658	0.0844	0.1068
3 IV and sample size 401	0.1216	0.0550	0.0662	0.0982	0.1078	0.0624	0.0888
Linear Model (ii)							
Benchmark	0.1874	0.1874	0.1850	0.1168	0.1148	0.0844	0.1302
Homoskedastic	0.1104	0.1104	0.1112	0.1180	0.1152	0.0644	0.1120
3 IV	0.1426	0.1784	0.1766	0.1484	0.1522	0.0844	0.1302
7 IV	0.1030	0.1744	0.1668	0.2966	0.2370	0.0844	0.1302
Stronger identif.	0.1874	0.1874	0.1850	0.1168	0.1148	0.0844	0.1334
No identif.	0.1874	0.1874	0.1850	0.1168	0.1148	0.0844	0.1002
Group Heterogeneity Model (iii)							
Benchmark	0.1854	0.1504	0.1758	0.1188	0.2806	0.1004	0.1050
7 IV	0.1354	0.0728	0.0978	0.1606	0.1866	0.1004	0.1050
15 IV	0.1110	0.1208	0.1200	0.3684	0.3260	0.1004	0.1050

Table 1: Empirical sizes associated with the 7 inference procedures for the three models and their different variations considered in Section 6 for a theoretical 10% level.

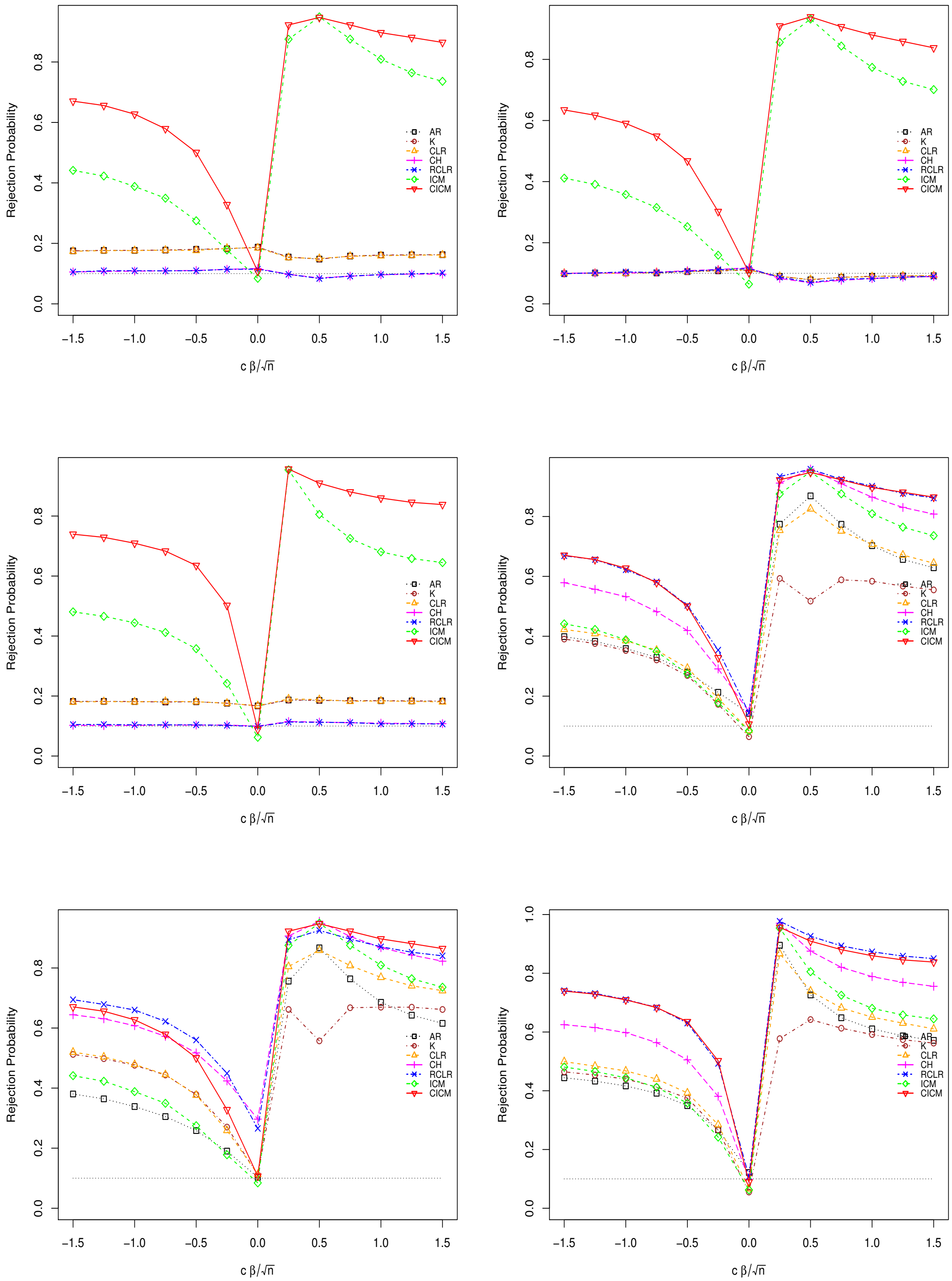

Figure 1: Power curves for Polynomial Model (i): benchmark (top left), homoskedastic case (top right), sample size 401 (middle left), 3 IV (middle right), 7 IV (bottom left) and 3 IV with sample size 401 (bottom right).

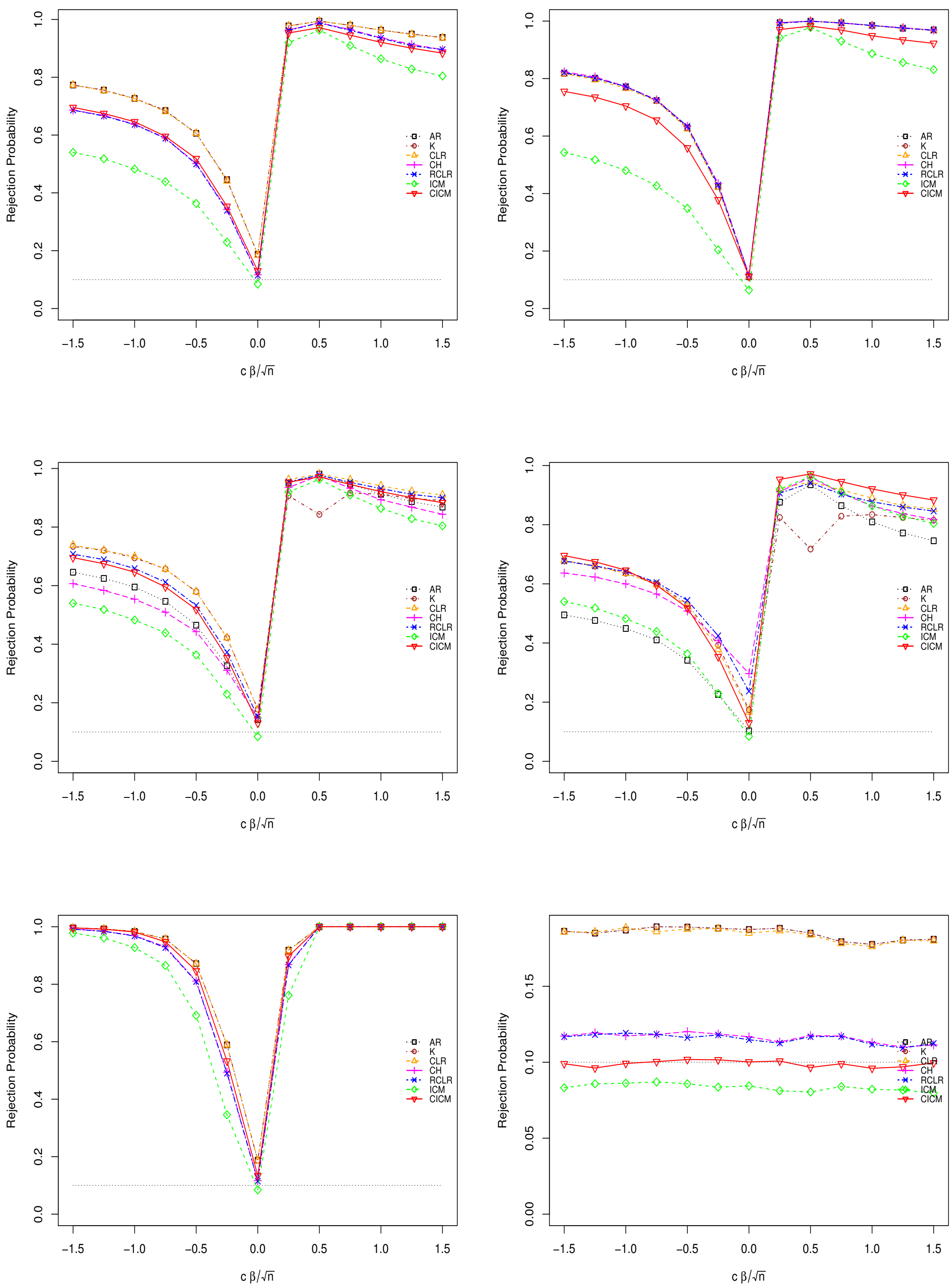

Figure 2: Power curves for Linear Model (i): benchmark (top left), homoskedastic case (top right), 3 IV (middle left), 7 IV (middle right), stronger identification (bottom left) and no identification (bottom right).

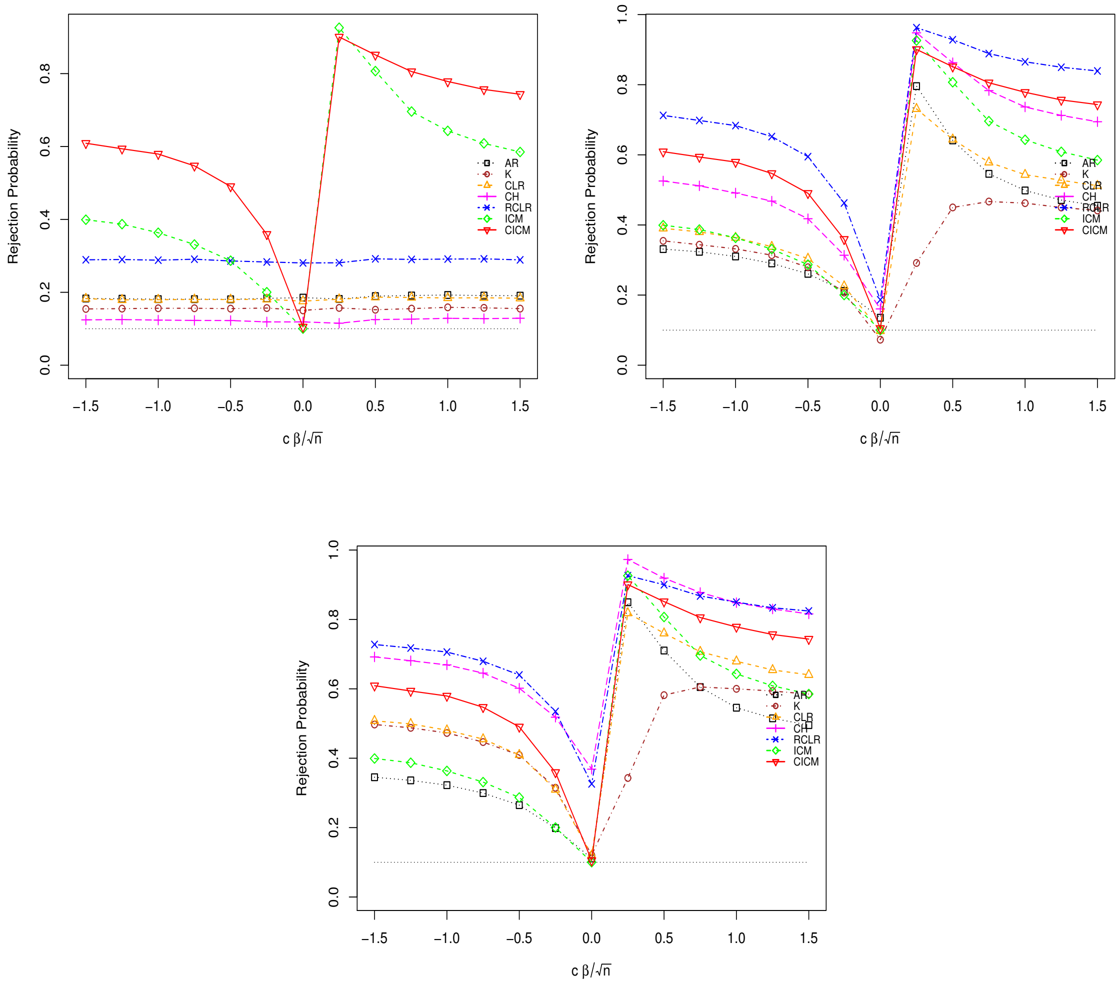

Figure 3: Power curves for Group Heterogeneity Model (ii): benchmark (top left), 7 IV (top right), and 15 IV (bottom).

Panel A: estimation using the 3 climate instruments							
A.1: estimation on the whole population				**A.2: estimation on the restricted population (region NE)**			
ICM	$\emptyset$			ICM	$\emptyset$		
CICM	[-2.26, -1.00]			CICM	[-1.775, 0.000]		
Inference procedures based on (7.12) with 3 instruments ($k = 1$)							
TSLS	[-1.21, -0.59]			TSLS	[-1.83, 0.04]		
AR	$\emptyset$	F-stat	97.25	AR	$\emptyset$	F-stat	13.7
CLR	[-1.51, -0.79]	Adj. R^2	0.22	CLR	$(-\infty, -5.51] \cup [48.48, \infty)$	Adj. R^2	0.05
CH	$\emptyset$			CH	$\emptyset$		
RCLR	[-1.54, -0.75]			RCLR	$(-\infty, -7.74] \cup [25.52, \infty)$		
Inference procedures based on (7.12) with 9 instruments ($k = 2$)							
TSLS	[-1.16, -0.55]			TSLS	[-0.58, 0.55]		
AR	$\emptyset$	F-stat	34.7	AR	$\emptyset$	F-stat	12.6
CLR	[-1.48, -0.77]	Adj. R^2	0.23	CLR	[-0.13, 7.21]	Adj. R^2	0.12
CH	$\emptyset$			CH	$\emptyset$		
RCLR	[-1.18, -0.58]			RCLR	[-0.4, 3.4]		
Inference procedures based on (7.12) with 18 instruments ($k = 3$)							
TSLS	[-1.04, -0.48]			TSLS	[-0.89, 0.04]		
AR	$\emptyset$	F-stat	20.6	AR	$\emptyset$	F-stat	10.1
CLR	[-1.37, -0.70]	Adj. R^2	0.25	CLR	[-2.08, -0.32]	Adj. R^2	0.17
CH	$\emptyset$			CH	$\emptyset$		
RCLR	[-0.92, -0.47]			RCLR	[-2.6, 0.3]		
Inference procedures based on (7.12) with 24 instruments ($k = 5$)							
TSLS	[-0.91, -0.39]			TSLS	[-0.79, 0.07]		
AR	$\emptyset$	F-stat	18.6	AR	$\emptyset$	F-stat	9.2
CLR	[-1.16, -0.55]	Adj. R^2	0.29	CLR	[-1.50, -0.10]	Adj. R^2	0.20
CH	$\emptyset$			CH	$\emptyset$		
RCLR	[-1.06, -0.40]			RCLR	$(-\infty, 1.48] \cup [4.68, \infty)$		
Panel B: estimation using only 1 climate instrument, the drought-rainfall gap							
B.1: estimation on the whole population				**B.2: estimation on the restricted population (region NE)**			
ICM	[-4.40, 0.65]			ICM	$(-\infty, +\infty)$		
CICM	[-2.62, -0.08]			CICM	$(-\infty, +\infty)$		
Inference procedures based on (7.13) with 1 instrument ($k = 1$)							
TSLS	[-2.01, -0.56]			TSLS	[-5.96, 1.24]		
AR	[-2.13, -0.61]	F-stat	55.9	AR	$(-\infty, 0.53] \cup [642.35, \infty)$	F-stat	3.8
CLR	[-2.13, -0.61]	Adj. R^2	0.05	CLR	$(-\infty, 0.53] \cup [642.35, \infty)$	Adj. R^2	0.004
CH	[-2.22, -0.66]			CH	$(-\infty, 0.17]$		
RCLR	[-2.20, -0.67]			RCLR	$(-\infty, 0.17]$		
Inference procedures based on (7.13) with 2 instrument ($k = 2$)							
TSLS	[-1.76, -0.39]			TSLS	[-0.15, 1.80]		
AR	[-1.40, -0.96]	F-stat	29.8	AR	[0.42, 1.61]	F-stat	18.6
CLR	[-2.02, -0.49]	Adj. R^2	0.05	CLR	[-0.03, 2.29]	Adj. R^2	0.04
CH	[-1.38, -0.75]			CH	$\emptyset$		
RCLR	[-1.62, -0.52]			RCLR	[1.05, 2.80]		
Inference procedures based on (7.13) with 3 instruments ($k = 3$)							
TSLS	[-1.65, -0.31]			TSLS	[0.83, 1.75]		
AR	$\emptyset$	F-stat	20.2	AR	[0.12, 1.96]	F-stat	12.6
CLR	[-2.03, -0.44]	Adj. R^2	0.06	CLR	[-0.06, 2.27]	Adj. R^2	0.04
CH	$(-\infty, -0.45]$			CH	$\emptyset$		
RCLR	[-1.76, -0.57]			RCLR	[1.12, 2.59]		
Inference procedures based on (7.13) with 5 instruments ($k = 5$)							
TSLS	[-1.47, -0.26]			TSLS	[-0.11, 1.81]		
AR	$\emptyset$	F-stat	14.6	AR	[-0.13, 2.64]	F-stat	7.7
CLR	[-1.79, -0.36]	Adj. R^2	0.06	CLR	[0.00, 2.41]	Adj. R^2	0.04
CH	$\emptyset$			CH	$\emptyset$		
RCLR	$[-1.67, -0.62] \cup [0.10, 0.47]$			RCLR	[0.84, 1.84]		

Table 2: 95% Confidence Intervals for the population collapse, using either the 3 climate instruments (Panel A), or only 1 climate instrument (Panel B), over the full sample of size equal to 1030 (whole population, left-hand side panel) and the restricted sample of size 780 (restricted population in region NE, right-hand side panel).

Country	AR	CLR	CH	RCLR	ICM	CICM	TSLS
USA (long)	$\emptyset$	[-0.20, 0.21]	[-0.25, -0.01]	[-0.77, 0.16]	$\emptyset$	[0.10, 0.31]	[-0.12, 0.23]
AUL	[-0.16, 0.22]	[-0.21, 0.27]	[-0.11, 0.22]	[-0.17, 0.28]	$\emptyset$	[-0.21, 0.11]	[-0.18, 0.27]
CAN	[-0.57, -0.12]	[-0.71, -0.00]	[-0.56, -0.16]	[-0.83, 0.09]	$\emptyset$	[-0.50, 0.01]	[-0.62, 0.01]
FR	[-0.70, 0.53]	[-0.48, 0.30]	[-0.57, 0.31]	[-0.40, 0.16]	[-0.73, 0.60]	[-0.32, 0.26]	[-0.47, 0.28]
GER	[-1.80, 0.26]	[-1.49, 0.04]	[-1.73, 0.66]	[-1.40, 0.33]	$\emptyset$	[-1.01, 0.23]	[-1.34, 0.07]
ITA	[-0.30, 0.19]	[-0.24, 0.11]	[-0.30, 0.19]	[-0.24, 0.11]	$\emptyset$	[-0.25, -0.00]	[-0.23, 0.09]
JAP	[-0.64, 0.43]	[-0.60, 0.40]	[-0.88, 0.25]	[-0.77, 0.20]	$\emptyset$	[-0.42, 0.36]	[-0.48, 0.34]
NTH	[-0.96, 0.69]	[-0.78, 0.50]	$\emptyset$	[-0.55, 0.22]	$\emptyset$	[-0.57, 0.19]	[-0.71, 0.41]
SWD	[-0.30, 0.25]	[-0.22, 0.17]	[-0.27, 0.26]	[-0.21, 0.20]	$\emptyset$	[-0.35, 0.12]	[-0.20, 0.16]
SWT	[-1.77, 0.35]	[-1.26, 0.06]	[-1.34, 0.26]	[-1.04, 0.05]	[-0.84, -0.15]	[-1.21, 0.05]	[-1.03, 0.05]
UK	[0.02, 0.30]	[-0.11, 0.43]	[0.20, 0.27]	[-0.69, 0.45]	$\emptyset$	[-0.12, 0.23]	[-0.08, 0.41]
USA (short)	$\emptyset$	[-0.22, 0.23]	$\emptyset$	[-0.24, 0.12]	$\emptyset$	[0.02, 0.27]	[-0.12, 0.24]

Table 3: 95%- confidence interval for the EIS using the Interest Rate. ICM and CICM regions are obtained with a grid of size 401 and 4,999 replications. The regions for TSLS, AR, CLR, CH and RCLR are computed using the following instruments that are lagged twice: the nominal interest rate, inflation, consumption growth, and log dividend price-ratio. Note that the regions for TSLS, AR, and CLR are obtained using `ivmodel` from R.

ICM	[-0.838, -0.148]
CICM	[-1.205, 0.048]
Inference procedures using the original set of 4 instruments, $Z1$	
TSLS	[-1.030, 0.050]
AR	[-1.767, 0.348]
CLR	[-1.256, 0.057]
CH	[-1.333, 0.258]
RCLR	[-1.055, 0.048]
Inference procedures using the extended set of 14 instruments, $Z2$	
TSLS	[-0.81, 0.09]
AR	[-1.734, 0.596]
CLR	[-1.09, 0.17]
CH	[-0.942, 0.085]
RCLR	[-0.680, -0.148]

Table 4: 95%-confidence interval for the EIS using the Interest Rate for Switzerland (SWT) with 91 quarterly observations from 1976Q2 to 1998Q4. The regions for ICM, CICM, CH and RCLR are obtained with a grid of size 401 evenly spread over [-2,1] and 999 replications, while the regions for TSLS, AR, and CLR are obtained using `ivmodel` from R. In addition, the regions for TSLS, AR, CLR, CH and RCLR are computed using the following sets of instruments: $Z1$ includes the nominal interest rate, inflation, consumption growth, and log dividend price-ratio, and all are lagged twice; $Z2$ includes the first two powers of the previously listed instruments as well as cross-products.