Identification-Robust Nonparametric Inference in a Linear IV Model

Bertille Antoine* and Pascal Lavergne†

October 2, 2018

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 rely the conditionality principle used by Moreira (2003) to condition on a set of ICM statistics that inform on identification strength. Our two procedures 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.

Address correspondence: Pascal Lavergne, Toulouse School of Economics, 21 Allées de Brienne, 31000 Toulouse FRANCE, and Bertille Antoine, Department of Economics, 8888 University Drive, Burnaby, BC V5A1S6, Canada.

*Simon Fraser University. Email: bertille_antoine@sfu.ca.

†Toulouse School of Economics. Email: pascal.lavergne@univ-tlse1.fr

1 Introduction

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

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

where Y_2 are endogenous variables, X_1 are exogenous control variables, and X_2 are exogenous instrumental variables. We focus on inference on the parameter β of the endogenous variables. Over the last 30 years, it has become clear that standard asymptotic approximations may reflect poorly what is observed even for large samples when there is weak correlation between instrumental variables and endogenous explanatory variables. Alternative asymptotic frameworks have then been developed to account for potentially weak identification and tests have been proposed that deliver reliable inference about parameters of interest, see e.g. Staiger and Stock (1997), Stock and Wright (2000), Moreira (2003), Kleibergen (2002, 2005), Andrews and Cheng (2012), Andrews and Guggenberger (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). The existing inference procedures are robust to identification strength, but rely on a linear projection in the first-stage equation that links endogenous variables and instruments. In what follows, we argue that linear projection can either artificially create weak identification or exacerbate the issue. 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). 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 rather 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 rely on the

conditionality principle used by Moreira (2003) in the context of weak identification to condition upon another ICM statistic (when Y_2 is univariate, or a set of ICM statistics when Y_2 is multivariate) that is informative 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 and are valid for any identification stength. 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 nonparametric on the form of the first-stage equation, these properties are robust to its particular functional form. We also show how to obtain asymptotically valid critical values under heteroskedasticity of unknown form. Our tests retain the same power properties as under homoskedasticity, namely they are consistent under semi-strong identification and can have non trivial power under weak identification.

In a series of simulations, we find that the level of our tests is well controlled using simulated critical values. We also illustrate that our tests 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 (2017).Our tests have a significant power advantage compared to existing ones when the reduced form equation is non-linear.

Our paper is organized as follows. In Section 2, we introduce our framework, we recall the main existing procedures for inference under possibly weak identification, and we motivate our new tests from a power perspective. In Section 3, we recall the ICM principle and we describe our two procedures, namely the ICM test and the conditional ICM test. In Section 4, we state our main theoretical results on validity and power, allowing for heteroskedaticity of unknown form. In Section 5, we study the small sample performance of our tests through Monte-Carlo simulations and compare it to previous proposals. In Section 6, we explore two empirical 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 (2017). Second, we revisit the empirical study of Yogo (2004) on the elasticity of intertemporal substitution. Proofs are gathered in Section 7.

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, which include X_2 but can also include X_1 if one suspects some nonlinearities in X_1 in the function $\Pi(\cdot)$; see (1.1) for the definition of X_1 and X_2.

It is common in the literature to consider a linear projection of the form $Z'\Pi$, and to model weak identification as $\Pi = n^{-1/2}C$. In that case, the mean of the first-stage F statistic testing $\Pi = 0$ 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

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

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

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

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

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

the AR statistic is (up to a scale) the F statistic for the null hypothesis $\Delta = 0$. It tests at the same time H_0 and the correct specification of the model. The K test of Kleibergen

(2005) is derived as a score test of H_0 under the assumptions of joint normality of u and V_2. The Conditional Likelihood Ratio (CLR) test is based on

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

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. Dufour and Taamouti (2007) 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 if we see it as a linear approximation of the regression of Y_2 on the instruments. But, a linear approximation can yield little or no power for the tests. As an example, assume $Z \sim N(0,1)$ and

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

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

$$\min_{\pi_1,\pi_2} \mathbb{E}\ \left(\pi_1 Z + \pi_2(Z^2 - 1) - \Pi(Z)\right)^2$$

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

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 $r_n = o(\sqrt{n})$, 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 create or exacerbate weak identification.[2]

Typically little prior information is available for the link between the endogenous variable and the instruments. One may be tempted to estimate the reduced form nonparametrically, for instance to increase the number of approximating polynomials with the sample size. But as shown by Jun and Pinkse (2012), weak identification prevents inference on β using nonparametrically generated instruments. Basically the local nature of nonparametric estimation, which reduces information and yields a slower than $\sqrt{n}$ rate of convergence, interferes with identification weakness. Another solution would be to do a specification search for the best functional form of the reduced equation. But these tests may suffer from low power in case of weak identification. Therefore, finding a testing method that leaves the first stage equation unspecified while being robust to weak identification is extremely valuable from a practitioner's viewpoint.

3 ICM and Conditional ICM Tests Statistics

3.1 Test 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

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

[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 a $\frac{1}{\sqrt{n}}$ component.

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 an idea of Bierens (1982), who shows that $\tilde{H}_0$ holds if and only if

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

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

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

where μ is some probability measure with support $\mathbb{R}^q$ (except maybe a set of isolated points). This can be rewritten as $b_0' Y' W Y b_0$, where W is a matrix with generic elements $n^{-1} w\,(Z_i - Z_j)$, with

$$w(z) = \int_{\mathbb{R}^q} e^{is'z}\, d\mu(s)\,. \tag{3.6}$$

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

$$\mathbb{E}\left[(y - Y_2'\beta_0)\, e^{is'Z_i}\right] = 0$$

for s in a (arbitrary) neighborhood of 0 in $\mathbb{R}^q$. Hence μ in (3.5) can be taken as any 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 this equivalence result equally applies to a one-to-one transformation of Z, which can be chosen with bounded image.

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

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 focus here on a particular application of the ICM principle that is suitable for theoretical investigation and practical implementation, and we leave for future work the investigation of the relative merits of these ICM-type tests.

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

$$\mathrm{ICM}(\beta_0) = \frac{b_0' Y' W Y b_0}{b_0' \widehat{\Omega} b_0}\,. \tag{3.7}$$

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 these authors chose a different normalization. Our normalization does not affect the main properties of the ICM test, but is convenient when computing critical values. 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

$$\mathrm{CICM}(\beta_0) = \frac{b_0' Y' W Y b_0}{b_0' \widehat{\Omega} b_0} - \min_{b} \frac{b' Y' W Y b}{b' \widehat{\Omega} b}\,. \tag{3.8}$$

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. CICM 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

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

Then ICM $= \widehat{S}'W\widehat{S}$ and

$$\text{CICM}(\beta_0) = \widehat{S}'W\widehat{S} - \lambda_{\min}\left(\begin{bmatrix}\widehat{S}' \\ \widehat{T}'\end{bmatrix} W \left[\widehat{S}, \widehat{T}\right]\right), \tag{3.9}$$

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

$$\text{CICM}(\beta_0) = \frac{1}{2}\left[\widehat{S}'W\widehat{S} - \widehat{T}'W\widehat{T} + \sqrt{\left(\widehat{S}'W\widehat{S} - \widehat{T}'W\widehat{T}\right)^2 + 4\left(\widehat{S}'W\widehat{T}\right)^2}\right]. \tag{3.10}$$

To establish (3.9), it suffices to 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 then

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

The matrix J is orthogonal, i.e. $J'J = \mathbf{I}$, as $A_0'b_0 = \mathbf{0}$. The minimum eigenvalue of $\widehat{\Omega}^{-1/2}Y'WY\widehat{\Omega}^{-1/2}$ is thus also the one of $J'\widehat{\Omega}^{-1/2}Y'WY\widehat{\Omega}^{-1/2}J$, and $Y\widehat{\Omega}^{-1/2}J = [\widehat{S}, \widehat{T}]$.

3.2 Critical Values

We now discuss how to obtain critical and P-values. We first assume normal errors with known covariance structure. We will then relax both assumptions.

3.2.1 Normal Errors

We first consider that Ω is known, and we replace $\widehat{S}$ and $\widehat{T}$ with $S = Yb_0\left(b_0'\Omega b_0\right)^{-1/2}$ and $T = Y\Omega^{-1}A_0\left(A_0'\Omega^{-1}A_0\right)^{-1/2}$.

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 distrbuted. 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.

Heteroskedastic Case Heteroskedasticity is often encountered in microeconometric models. The usual way to account for potential unknown heteroskedasticity is to modify the test statistic at the outset. For instance, Chernozhukov and Hansen (2008) use an Anderson-Rubin-Wald type statistic that is made robust to heteroskedasticity by using a heteroskedasticity-robust estimator of the covariance matrix. We instead consider the same statistic ICM, but we allow for unknown heteroskedasticity when simulating critical values. If we know the conditional variance function

$$\Omega_i \equiv \Omega(Z_i) = \mathrm{Var}\left(Y_i | Z_i\right) = \begin{pmatrix} \mathrm{Var}(y_i|Z_i) & \mathrm{Cov}(y_i, Y_{2i}|Z_i) \\ \mathrm{Cov}'(Y_{2i}, y_i|Z_i) & \mathrm{Var}(Y_{2i}|Z_i) \end{pmatrix}, \tag{3.11}$$

we can compute $\Sigma = \mathrm{Var}(Yb_0) = \mathrm{diag}\left(b_0'\Omega_1 b_0, \ldots, b_0'\Omega_n b_0\right)$. 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 by Kleibergen (2005), and more generally that a wide class of QLR tests are valid when conditionning on a nuisance process. Instead of modifying our statistic,

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 (3.8), and thus would not have the same intuitive interpretation. However, we make no claim about the optimality of our procedure.

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

$$E = [E_1 \dots E_n] \qquad E_i = T_i - \frac{\text{Cov}(T_i, S_i|Z_i)}{\text{Var}(S_i|Z_i)} S_i \,.$$

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

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

in place of T_i. As will be shown in the next section, when the covariance structure is unknown, estimation of conditional variances has no first-order asymptotic effect on the validity of critical values of our tests.

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 propose to condition a test of the null hypothesis on the process of sample moments evaluated at values of β, that informs on $\mathbb{E}\left(Z(y - Y_2'\beta)\right)$. In the heteroskedastic case, the sample process $n^{-1}\sum_{i=1}^n Z_i\,(y_i - Y_{2i}'\beta)$ needs to be orthogonalized with respect to the sample mean under the null hypothesis $n^{-1}\sum_{i=1}^n Z_i\,(y_i - Y_{2i}'\beta_0)$. This can be done by using the covariance of these processes for different values of the parameter. The issue with CICM is similar but more intricate, as we are interested

in the mean function $\mathbb{E}\left[(y - Y_2'\beta_0)e^{-it'Z}\right]$, and we consider as a nuisance parameter $\mathbb{E}\left[(y - Y_2'\beta)e^{-it'Z}\right]$ for all other values of β and all t. To orthogonalize the corresponding sample mean process, we then need a transformation that removes correlation at the level of individual observations.

3.2.2 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.

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 conditional ICM statistic depends on $S'WS$, $S'WT$, and $T'WT$ as seen from (3.9), which are linear and quadratic forms in S. Under homoskedasticity, S is uncorrelated with the columns of T (conditional on Z), and the same method provides asymptotically correct critical values.

Accounting now 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 with estimation of $\Omega = \mathbb{E}\,\mathrm{Var}(Y|Z)$ and $\Omega(\cdot) = \mathrm{Var}(Y|Z = \cdot)$. We should first note that weak identification does not preclude consistent estimation of the latter. 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. In heteroskedastic cases, we need nonparametric conditional variance estimators. 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[4] $K(\cdot)$, and a bandwidth b_n. With $e =$

[4] In our simulations and applications, we used the triangular kernel; we also used the Gaussian kernel and the results were very similar.

$(1, \dots 1)'$, 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 $\Omega(\cdot)$,a and a consistent estimator of Ω using $\widehat{\Omega} = n^{-1}\sum_{i=1}^n \widehat{\Omega}(Z_i)$.

With at hand a nonparametric estimator of $\Omega(\cdot)$, one can estimate the conditional variance of S_i by $\widehat{\text{Var}}(S_i|Z_i) = b_0'\widehat{\Omega}_i b_0 \left(b_0\widehat{\Omega}b_0\right)^{-1}$. To approximate the asymptotic distribution of ICM $= S'WS$, we generate independent $\widehat{G}_i$, $i =, \dots n$, as normal with mean 0 and variance $\widehat{\text{Var}}(S_i|Z_i)$ for each i, and proceeds similarly as above. The intuition carries over for CICM, provided we condition on the part of $\widehat{T}$ which is asymptotically uncorrelated with $\widehat{S}$ conditional on Z. The conditional covariance of $\widehat{T}_i$ and $\widehat{S}_i$ can be estimated as

$$\left(A_0'\widehat{\Omega}^{-1}A_0\right)^{-1/2} A_0'\widehat{\Omega}_i b_0 \left(b_0'\widehat{\Omega}b_0\right)^{-1/2}.$$

Then the asymptotic distribution of CICM will be approximated by first computing $\widehat{E} = \left[\widehat{E}_1 \dots \widehat{E}_n\right]$, with

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

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

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

in place of $\widehat{T}_i$.

3.3 Confidence Regions

To obtain a confidence set for β, one needs to invert the ICM test. The confidence set for β at significance level α is

$$\{\beta_0 : ICM(\beta_0) < c_{1-\alpha}(\beta_0)\},$$

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

Lemma 3.1 *For homoskedastic errors, and when β is scalar, the 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.

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 at level α is defined as

$$\{\beta_0 : \text{CICM}(\beta_0) < c_{1-\alpha}(\beta_0)\} ,$$

where $c_{1-\alpha}(\beta_0) \equiv c_{1-\alpha}(Z, \widehat{E}(\beta_0), \beta_0)$ is the $1-\alpha$ quantile of the statistic obtained by simulations. However, it does not seem possible to obtain a simple characterization of confidence intervals, as done by Mikusheva (2010) for the CLR test.

4 Theoretical Results

4.1 Similarity of CICM Test

Our first result shows that the CICM test is similar conditionally on $E(\beta_0)$ and Z in the simple case of normal errors with known covariance structure. 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 statistic ultimately depends on the empirical processes that are jointly asymptotically Gaussian whatever the particular error distribution; see our proofs in Section 7. Hence the ideal setup allows to study the properties of our test abstracting from finite-sample considerations.

Define the conditional critical value

$$c_{1-\alpha}(Z, E(\beta_0), \beta_0) = \min \{c : \Pr [\text{CICM} (\beta_0) > c | Z, E(\beta_0)] \leq \alpha\} .$$

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

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

4.2 Asymptotic Validity

We consider the following assumptions.

Assumption A *(i) The observations* (y_i, Y_{2i}, Z_i) *are iid and follow (2.2) and (2.3). (ii)* $\mathbb{E}(|y|^2|Z)$ *and* $\mathbb{E}(\|Y_2\|^2|Z)$ *are uniformly bounded away from infinity.*

Assumption B $w(\cdot)$ *is symmetric, bounded, has a non-negative Fourier transform (almost everywhere), and is such that* $\int w^2(x)\, dx = 1$*.*

We consider the set of functions on $\mathbb{R}^k$ that possess uniformly bounded derivatives up to order $\lfloor \alpha \rfloor$ and whose highest partial derivatives are Lipschitz of order $\alpha - \lfloor \alpha \rfloor$. More formally, for any vector $q = (q_1, \dots q_k)'$ of integers, define the differential operator

$$D_q = \frac{\partial^{\sum_{j=1}^k q_j}}{\partial z_1^{q_1} \dots \partial z_k^{q_k}} ,$$

and

$$\|f\|_\alpha = \max_{\sum_{j=1}^k q_j \leq \lfloor \alpha \rfloor} \sup_z |D_q f(z)| + \max_{\sum_{j=1}^k q_j = \lfloor \alpha \rfloor} \sup_{z \neq z'} \frac{|D_q f(z) - D_q f(z')|}{\|z - z'\|^{\alpha - \lfloor \alpha \rfloor}} .$$

Then $\mathcal{C}_M^\alpha$ is the set of continuous functions with $\|f\|_\alpha \leq M$ for some finite M. The class $\mathcal{F}$ is defined as the set of matrix functions that have eigenvalues uniformly bounded away from zero and infinity and whose elements belong to $\mathcal{C}_M^\alpha$ for some $\alpha > k/2$ and some finite M.

Assumption C *(i)* $\Omega(\cdot) \equiv \mathrm{Var}(Y|Z=.)$ *belongs to the class* $\mathcal{F}$. *(ii) For a partition of* $\mathbb{R}^k = \cup_{j=1}^{\infty} I_j$ *into bounded convex sets with nonempty interior and some* $s < 1/2$,

$$\sum_{j=1}^{\infty} \left(\Pr\left(Z \in I_j\right)\right)^s < \infty\,.$$

Assumption D *(i)* $\int \|\widehat{\Omega}(Z) - \Omega(Z)\|^2\, dP(Z) \xrightarrow{p} 0$ *(ii)* $\Pr\left(\widehat{\Omega}(\cdot) \in \mathcal{F}\right) \to 1$ *as* $n \to \infty$ *(iii)* $\widehat{\Omega} - \Omega \xrightarrow{p} 0$.

Under these conditions the estimation of $\Omega(.)$ does not affect the asymptotic behavior of our test statistics. Conditions imposed by Assumption D-(i) and (ii) are standard in the literature, and primitive conditions on estimators have been derived (see e.g. Andrews (1994)). Assumption C-(ii) is a condition on the tails of the distribution of Z, and is trivially satisfied if Z is bounded. When $k = 1$, it is fulfilled as soon as $\mathbb{E}\,|Z|^{2+\delta} < \infty$ for some $\delta > 0$. Under Assumption C the results from van der Vaart (1994) on the entropy of the class $\mathcal{F}$ apply.

Assumption E *(i)* $\Pi(Z) = C(Z)/r_n$, *where* $1 \leq r_n \leq \sqrt{n}$ *(ii)* $\mathbb{E}\,\|C(Z)\|^2 < \infty$.

We allow for strong identification, i.e. $r_n = 1$, weak identification, i.e. $r_n = n^{1/2}$, and no identification when $C(.) = 0$. One could also allow for different identification strengths across the different equations of (2.3), but this would lead to very cumbersome calculations.

We respectively denote by $c(Z, \alpha)$ and $c(Z, \widehat{E}(\beta_0), \alpha)$ 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).

Theorem 4.2 *Let Assumptions A to E hold. Under* H_0, $\Pr\left[\mathrm{ICM}(\beta_0) > c\left(Z, \alpha\right)\right] \to \alpha$ *and* $\Pr\left[\mathrm{CICM}(\beta_0) > c\left(Z, \widehat{E}\left(\beta_0\right), \alpha\right)\right] \to \alpha$ *as* $n \to \infty$.

4.3 Asymptotic Power

We adopt here the large local alternatives setup considered by Bierens and Ploberger (1997). Namely, we assume

$$\Pi(Z) = \tilde{c}\, r_n^{-1} C(Z_i) \qquad \mathbb{E}\,\|C(Z)\|^2 = 1\,, \tag{4.12}$$

where $\tilde{c}$ is a positive constant. We consider that β_0 is the true value of β and that we entertain a test of

$$H_0 : \ \beta = \beta_1 \qquad H_1 : \ \beta \neq \beta_1 \,,$$

where $\beta_1 \neq \beta_0$ is fixed. The object of interest is the asymptotic power of our two tests when $r_n << \sqrt{n}$ (semi-strong instruments) or $r_n = \sqrt{n}$ but $\tilde{c}$ becomes large.

Theorem 4.3 *Under Assumptions A to D and (4.12), when* $\beta_1 \neq \beta_0$*,*

- *If* $r_n/\sqrt{n} \to 0$*, both* $\Pr\left[\mathrm{ICM}(\beta_1) > c\left(Z, \alpha\right)\right]$ *and* $\Pr\left[\mathrm{CICM}(\beta_1) > c\left(Z, \widehat{E}(\beta_1), \alpha\right)\right]$ *tend to 1.*

- *If* $r_n = \sqrt{n}$*, then*

$$\lim_{\tilde{c}\to+\infty} \Pr\left[\mathrm{ICM}(\beta_1) > c\left(Z, \alpha\right)\right] = \lim_{\tilde{c}\to+\infty} \Pr\left[\mathrm{CICM}(\beta_1) > c\left(Z, \widehat{E}(\beta_1), \alpha\right)\right] = 1 \,.$$

The above result shows that under weak identification power is non trivial for a large enough $\tilde{c}$. As already mentionned, the power properties of ICM directly follows from the ones established by Bierens and Ploberger (1997). Namely, the asymptotic distribution of $\mathrm{ICM}(\beta_1)$ is the same as the one of $G'WG$, where G is a vector of independent normal with the same individual means and variances as $Y_i'\Omega_i b_0 \left(b_0'\Omega_i b_0\right)^{-1/2}$, where $b_1 = (1, -\beta_1)'$, that is the one of $\sum_{i=1}^n \lambda_i \left(G_i + c_i\right)^2$, where $\lambda_i, i = 1, \ldots n$, are strictly positive real numbers, $G_i, i = 1, \ldots n$, are independent standard normals, and $c_i, i = 1, \ldots n$, are non-zero real numbers. It easily follows that this distribution stochastically dominates at first order the asymptotic distribution of $\mathrm{ICM}(\beta_0)$, which is similar but with $c_i = 0$ for all i. The behavior of CICM is related, but more involved because it depends on the behavior of the whole process $\mathrm{ICM}(\beta)$ where β varies over the parameter space.

5 Small Sample Behavior

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

$$\begin{aligned} y_i &= \alpha_0 + Y_{2i}\beta_0 + \sigma(Z_i)u_i \,, \\ Y_{2i} &= \gamma_0 + \frac{c}{\sqrt{n}} f(Z_i) + \sigma(Z_i)v_{2i} \,. \end{aligned} \tag{5.13}$$

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.

6 Empiricals illustrations

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

We extend some of the results presented in Sellars and Alix-Garcia (2017) 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 dependence between the two) are used as instrumental variables. Sellars and Alix-Garcia (2017) 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 ratio of the former to 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 by Sellars and Alix-Garcia (2017)[5] (2017) 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 an indicator for whether the municipality i had an above-median number of estates per square kilometer in 1900, 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].

Our results are presented in 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. We also present confidence regions computed with TSLS, as well as 4 weak-identification robust inference procedures, the similar tests AR and conditional LR (CLR); the conditional LR robust to heteroskedasticity (RCLR); the

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

[6] We follow Sellars and Alix-Garcia (2017) and control for the standard deviation of PDSI, for a measure of maize productivity, for various measures of elevation and slope, and include governorship-level fixed effects.

method proposed by Chernozukhov and Hansen (CH): all these inference procedures rely on the following (linear) first-stage equation,

$$Y_{2i} = \Pi Z_i + \delta' X_{1i} + v_i \,, \qquad \mathbb{E}\left(v_i | X_{1i}, Z_i\right) = 0 \tag{6.14}$$

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[7] with cross-products of order 2 (a total of 9 instruments), or the first three powers of these three instruments and cross-products of order 3 (a total of 18 instruments). Table 2 considers the full sample of municipalities across central Mexico (a total of 1080 observations). 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. It is interesting to mention that the confidence regions of CLR and RCLR vary substantially with the instrument set that is used: moreover, as nonlinearities are accounted for through the use of powers of the original climate variables, these confidence regions become closer to the one of CICM. This is an important practical message that emphasizes the advantage of using an inference procedure - such as CICM - that relies on the exogeneity of the instrument, without having to specify or pin down the (potentially nonlinear) relationship between endogenous variable and instruments.

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 is not too surprising given that the dependent variable is binary and that the structural model is linear in a continuous variable.

6.2 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,

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

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

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 similar tests AR and conditional LR (CLR); the conditional LR robust to heteroskedasticity (RCLR); the method proposed by Chernozukhov and Hansen (CH); the 2 tests proposed in this paper, ICM and the conditional ICM (CICM); as well as the TSLS[8]. The weak-identification robust confidence intervals indicate that the EIS is below 1, but small and not significantly different from 0 for most countries.

When comparing the conditional ICM confidence interval to the conditional LR confidence interval - which is known to be tighter due to the good power properties of CLR - the conditional ICM always delivers tighter bounds, but for Sweden (SWD). Focusing on the conditional ICM confidence interval, the EIS is 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). Unreported estimations of a first stage equation of r_{t+1} on various sets

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

of instruments revealed the presence of non-linear quadratic effects, but not cubic. 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 and CICM. Other inference procedures obtain a tighter confidence interval, except AR.

7 Proofs

Proof of Lemma 3.1

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.

Proof of Lemma 4.1

The result follows from the following facts (i) the components of $[S, E]$ are jointly normal (ii) $S \sim N(\mathbf{0}, \mathbf{I})$ is independent of Z and $\Pi(\cdot)$ (ii) S is also uncorrelated, thus independent, with all components of E.

Proof of Theorem 4.2

Lemma 7.1 *Under Assumptions A and C-(ii)*

$$n^{-1/2}\sum_{i=1}^{n}\left(Y_i - \mathbb{E}\left(Y_i|Z_i\right)\right)e^{is'Z_i} \rightsquigarrow \mathbb{G}_Y(s), \tag{7.15}$$

where $\mathbb{G}_Y$ is a gaussian process with mean 0 and covariance

$$\mathbb{E}\,\mathbb{G}_Y(s)\mathbb{G}_Y(t) = \mathbb{E}\left(\mathrm{Var}(Y|Z)e^{i(s+t)'Z}\right).$$

Let $B(\cdot)$ and $D(\cdot)$ be (possibly vector) functions of Z with finite second moments, such as $B(\cdot)$ belongs to $\mathcal{C}_M^\alpha$ for some $\alpha > k/2$ and finite M, and let $\widehat{B}(\cdot)$ an estimator of $B(\cdot)$ such that Assumption D-(i) and (ii) hold. Then

$$\sup_s \left| n^{-1/2} \sum_{i=1}^n \left(\widehat{B}_i - B_i \right) \left(Y_i - \mathbb{E}\left(Y_i | Z_i \right) \right) e^{is'Z_i} \right| \xrightarrow{p} 0 \ . \tag{7.16}$$

Proof. The class of functions $\mathcal{E} = \left\{ e^{is'Z}, s \in \mathbb{R}^k \right\}$ is uniformly bounded by 1 in modulus and infinitely differentiable, and then belongs to $\mathcal{C}_M^\alpha$, for some finite M and $\alpha > k/2$. From Assumption C-(ii) and van der Vaart (1994), it then has bracketing entropy bounded by a polynomial with exponent less than 2, that is

$$\log N_{[]}(\varepsilon, \mathcal{E}, L_2(P)) \leq K \left(\frac{1}{\varepsilon} \right)^V \quad V < 2 \quad \text{for some } K \quad \text{and for any } \epsilon > 0 \, .$$

The convergence in 7.15 thus follows from Assumption A.

The class of functions $\{ B(\cdot) e^{is'\cdot}, B(\cdot) \in \mathcal{C}_M^{k/2}, s \in \mathbb{R}^k \}$ has, up to some constant, polynomial bracketing entropy with exponent less than 2, see e.g. Kosorok (2008, Lemma 9.25), and the second convergence follows. ■

Let us first deal with

$$\text{ICM}(\beta_0) = \widehat{S}' W \widehat{S} = \int_{\mathbb{R}^q} \left| n^{-1/2} \sum_{i=1}^n \widehat{S}_i e^{is'Z_i} \right|^2 d\mu(s) \, .$$

Now

$$n^{-1/2} \sum_{i=1}^n \widehat{S}_i e^{is'Z_i} = (1 + o_p(1)) \left(b_0' \widehat{\Omega} b_0 \right)^{-1/2} n^{-1/2} \sum_{i=1}^n Y_i' b_0 e^{is'Z_i} \rightsquigarrow \mathbb{G}_S(s) \, ,$$

using Assumption D-(iii) and (7.15), where $\mathbb{G}_S(s)$ is a Gaussian process. But for $G_i = \text{Var}(S_i|Z_i)\varepsilon_i$ with independent $\varepsilon_i \sim N(0,1), i = 1, \ldots n$, $n^{-1/2} \sum_{i=1}^n G_i e^{is'Z_i} \rightsquigarrow \mathbb{G}_S(s)$. Let now $\widehat{G}_i = \left(b_0' \widehat{\Omega} b_0 \right)^{-1/2} \left(b_0' \widehat{\Omega}_i b_0 \right)^{1/2} \varepsilon_i$. As $\Omega(\cdot)$ is uniformly bounded from above and below, the class of functions $\left\{ \left(b_0' \widehat{\Omega} b_0 \right)^{-1/2} \left(b_0' \widehat{\Omega}_i b_0 \right)^{1/2}, \Omega \in \mathcal{F} \right\}$ satisfy Assumption C for some finite M', see Kosorok (2008, Lemma 9.25). From Lemma 7.1, Equation (7.16), $n^{-1/2} \sum_{i=1}^n \widehat{G}_i e^{is'Z_i} \rightsquigarrow \mathbb{G}_S(s)$. By the continous mapping theorem, $\widehat{G}' W \widehat{G}$ has thus the same asymptotic distribution than $S'WS$.

Let's now deal with CICM, and first with $\widehat{T}' W \widehat{S}$. We have

$$\widehat{T}' W \widehat{S} = n^{-1} \sum_{i=1}^n \left(A_0' \widehat{\Omega}^{-1} A_0 \right)^{-1/2} \frac{A_0' \widehat{\Omega}^{-1} \widehat{\Omega}_i b_0}{b_0' \widehat{\Omega}_i b_0} Y_i' b_0 \widehat{S}_j w_{ij} + \left(\widehat{E} - \widehat{F} \right)' W \widehat{S} + \widehat{F}' W \widehat{S} \, ,$$

where $\widehat{F}_i = A_0'\widehat{\Omega}^{-1}\mathbb{E}\left(Y_i|Z_i\right)$. The first term can be written as

$$\left(A_0'\widehat{\Omega}^{-1}A_0\right)^{-1/2}\int_{\mathbb{R}^q}\left[n^{-1/2}\sum_{i=1}^n\frac{A_0'\widehat{\Omega}^{-1}\widehat{\Omega}_i b_0}{b_0'\widehat{\Omega}_i b_0}Y_i'b_0e^{is'Z_i}\right]\left[n^{-1/2}\sum_{i=1}^n\widehat{S}e^{-is'Z_i}\right]\,d\mu(s)\,.$$

The second process has been dealt with previously. For the first one, $\left(A_0'\widehat{\Omega}^{-1}\widehat{\Omega}(\cdot)b_0\right)\,\left(b_0'\widehat{\Omega}(\cdot)b_0\right)$ belongs to $\mathcal{C}_M^\alpha$ for some $\alpha > k/2$ and some finite M under Assumptions C and D, and from Lemma 7.1, we can replace estimated variances by their true values without affecting its asymptotic behavior. The process thus weakly converges to a Gaussian process. The second term is

$$\left(\widehat{E}-\widehat{F}\right)'W\widehat{S} = \int_{\mathbb{R}^q}\left[n^{-1/2}\sum_{i=1}^n\left(\widehat{E}_i-\widehat{F}(Z_i)\right)'e^{is'Z_i}\right]\left[n^{-1/2}\sum_{i=1}^n\widehat{S}_ie^{-is'Z_i}\right]\,d\mu(s)$$

By similar arguments, the first process is

$$\left(1+o_p(1)\right)\left(A_0'\Omega^{-1}A_0\right)^{-1/2}n^{-1/2}\sum_{i=1}^n\left(A_0'\Omega^{-1}\left(Y_i-\mathbb{E}\left(Y_i|Z_i\right)\right)-\frac{A_0'\Omega^{-1}\Omega_i b_0}{b_0'\Omega_i b_0}Y_i'b_0\right)e^{is'Z_i}\,,$$

and converges to a Gaussian process. Rescaling the third term, we obtain

$$\frac{r_n}{\sqrt{n}}\widehat{F}'W\widehat{S} = \left(A_0'\Omega^{-1}A_0\right)^{-1/2}\int_{\mathbb{R}^q}\left[n^{-1}\sum_{i=1}^n A_0'\widehat{\Omega}^{-1}C(Z_i)e^{is'Z_i}\right]\left[n^{-1/2}\sum_{i=1}^n\widehat{S}_ie^{-is'Z_i}\right]\,d\mu(s)\,.$$

Under Assumptions D, the first process above the integral is asymptotically equivalent to

$$\left[n^{-1}\sum_{i=1}^n A_0'\Omega^{-1}C(Z_i)e^{is'Z_i}\right]$$

and converges to $\mathbb{E}\left(A_0\Omega^{-1}C(Z)e^{is'Z}\right)$

The term $T'WT$ can be also decomposed and dealt with similarly, see the next proof. Let's gather our results. For simplicity of exposition, let's assume $l = 1$, and denote $\widehat{B} = \text{diag}\left(\widehat{\text{Cov}}(T_i,S_i|Z_i)/\widehat{\text{Var}}(S_i|Z_i), i=1,\dots n\right)$, then we have

$$\widehat{T}'W\widehat{S} = \widehat{S}\widehat{B}'W\widehat{S} + \left(\widehat{E}-\widehat{F}\right)'W\widehat{S} + \widehat{F}'W\widehat{S} \tag{7.17}$$

$$\widehat{T}'W\widehat{T} = \widehat{S}\widehat{B}'W\widehat{B}\widehat{S} + \left(\widehat{E}-\widehat{F}\right)'W\left(\widehat{E}-\widehat{F}\right) + 2\left(\widehat{E}-\widehat{F}\right)'W\widehat{B}\widehat{S} \tag{7.18}$$

$$+ 2\widehat{F}'W\widehat{B}\widehat{S} + 2\left(\widehat{E}-\widehat{F}\right)'W\widehat{F} + \widehat{F}'W\widehat{F}\,. \tag{7.19}$$

The probability orders of the different terms are given in the following table.

Probability orders	1	$\sqrt{n}/r_n$	n/r_n^2
$\widehat{S}'W\widehat{S}$	x		
$\widehat{S}'\widehat{B}W\widehat{S}$	x		
$\left(\widehat{E}-\widehat{F}\right)'W\widehat{S}$	x		
$\widehat{F}'W\widehat{S}$		x	
$\widehat{S}'\widehat{B}W\widehat{B}\widehat{S}$	x		
$\left(\widehat{E}-\widehat{F}\right)'W\left(\widehat{E}-\widehat{F}\right)$	x		
$\left(\widehat{E}-\widehat{F}\right)'W\widehat{B}\widehat{S}$	x		
$\widehat{F}'W\widehat{B}\widehat{S}$		x	
$\left(\widehat{E}-\widehat{F}\right)'W\widehat{F}$		x	
$\widehat{F}'W\widehat{F}$			x

Each of these terms in turn depends on the four processes

$$n^{-1/2}\sum_{i=1}^{n} Y_i'b_0e^{is'Z_i} \qquad n^{-1/2}\sum_{i=1}^{n}\frac{A_0'\widehat{\Omega}^{-1}\widehat{\Omega}_ib_0}{b_0'\widehat{\Omega}_ib_0}Y_i'b_0e^{is'Z_i}$$
$$n^{-1/2}\sum_{i=1}^{n}\left(\widehat{E}_i-\widehat{F}(Z_i)\right)e^{is'Z_i} \qquad n^{-1}\sum_{i=1}^{n}A_0'\widehat{\Omega}^{-1}C(Z_i)e^{is'Z_i}\,.$$

From Lemma 7.1, the three former processes jointly weakly converge to a Gaussian process, with zero asymptotic covariance between the components depending on $Y_i'b_0$ and those depending on $\widehat{E}_i-\widehat{F}$.

Let's now consider separately two cases. Under weak identification, $r_n=\sqrt{n}$ and each term has the same probability order. Then we can conclude, as we did above, that replacing $\widehat{S}$ by $\widehat{G}$ results in the same joint weak limit for these processes, and thus for CICM. Under semi-strong identification, we will use the result proven below that under semi-strong identification, the asymptotic behavior of CICM is driven by $\widehat{S}'W\widehat{T}\left(\widehat{T}'W\widehat{T}\right)^{-1}\widehat{T}'W\widehat{S}$. But from our previous results

$$\begin{aligned}\widehat{S}'W\widehat{T}\left(\widehat{T}'W\widehat{T}\right)^{-1}\widehat{T}'W\widehat{S} &= \left(\frac{r_n}{n^{1/2}}\widehat{S}'W\widehat{T}\right)\left(\frac{r_n^2}{n^1}\widehat{T}'W\widehat{T}\right)^{-1}\left(\frac{r_n}{n^{1/2}}\widehat{T}'W\widehat{S}\right)\\ &= \left(\frac{r_n}{n^{1/2}}\widehat{S}'W\widehat{F}\right)\left(\frac{r_n^2}{n}\widehat{F}'W\widehat{F}\right)^{-1}\left(\frac{r_n}{n^{1/2}}\widehat{F}'W\widehat{S}\right)(1+o_p(1))\,.\end{aligned}$$

We can again conclude that replacing $\widehat{S}$ by $\widehat{G}$ results in the same asymptotic distribution for CICM.

Lemma 7.2 *Under the assumptions of Theorem 4.2, if* $r_n/\sqrt{n} \to 0$,

$$\text{CICM} - \widehat{S}'W\widehat{T}\left(\widehat{T}'W\widehat{T}\right)^{-1}\widehat{T}'W\widehat{S} \xrightarrow{p} 0\,.$$

Proof. By definition,

$$\text{CICM} = \widehat{S}'W\widehat{S} - \lambda_{\min}(A_n) \quad A_n = \begin{pmatrix} \widehat{S}'W\widehat{S} & \widehat{S}'W\widehat{T} \\ \widehat{T}'W\widehat{S} & \widehat{T}'W\widehat{T} \end{pmatrix}.$$

We have to determine the behavior of $\lambda_{\min}(A_n)$. Note that A_n is positive definite and thus invertible. Now

$$A_n^{-1} = \begin{pmatrix} A_n^{11} & A_n^{21'} \\ A_n^{21} & A_n^{22} \end{pmatrix} \qquad A_n^{11} = \left(\widehat{S}'W\widehat{S} - \widehat{S}'W\widehat{T}\left(\widehat{T}'W\widehat{T}\right)^{-1}\widehat{T}'W\widehat{S}\right)$$

$$A_n^{21} = -\left(\widehat{T}'W\widehat{T}\right)^{-1}\widehat{T}'W\widehat{S}A_n^{11} \qquad A_n^{22} = \left(\widehat{T}'W\widehat{T} - \widehat{T}'W\widehat{T}\left(\widehat{S}'W\widehat{S}\right)^{-1}\widehat{S}'W\widehat{T}\right)^{-1}.$$

Under semi-strong identification, A_n^{11}, $\widehat{S}'W\widehat{T}$, and $\widehat{T}'W\widehat{T}$ are of respective probability order 1, $\sqrt{n}/r_n$ and n/r_n^2. Hence

$$A_n^{-1} - \begin{pmatrix} A_n^{11} & \mathbf{0}' \\ \mathbf{0} & \mathbf{0}_{l\times l} \end{pmatrix} \xrightarrow{p} 0.$$

The maximum eigenvalue of the latter matrix is A_n^{11}, which is strictly positive and bounded in probability. By the continuity of the maximum eigenvalue, it must be that $\lambda_{\max}A_n^{-1} - A_n^{11} \xrightarrow{p} 0$ and thus $\lambda_{\min}A_n - \left(A_n^{11}\right)^{-1} \xrightarrow{p} 0$. ■

Proof of Theorem 4.3

We have $\text{ICM}(\beta_1) - \text{ICM}(\beta_0) = \text{CICM}(\beta_1) - \text{CICM}(\beta_0)$. Write

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

with $a' = (a_1, a_2') = \widehat{Q}b_1$ and

$$\widehat{Q} = \left[b_0'\widehat{\Omega}\ A_0'\right]/(b_1'\widehat{\Omega}b_1)^{-1/2} \xrightarrow{p} Q = \left[b_0'\Omega\ A_0'\right]/(b_1'\Omega b_1)^{-1/2}.$$

Since $\beta_1 \neq \beta_0$, $a_2 \neq 0$. We thus can write

$$\text{ICM}(\beta_1) - \text{ICM}(\beta_0) = (a_1^2 - 1)\widehat{S}'W\widehat{S} + a_2'\widehat{T}'W\widehat{T}a_2 + 2a_1a_2'\widehat{T}'W\widehat{S}\,.$$

From the previous proof, $\widehat{S}'W\widehat{S} = O_p(1)$, $\widehat{T}'W\widehat{S} = O_p(r_n/\sqrt{n})$, and the dominant term is $\widehat{T}'W\widehat{T}$. Now assume $l = 1$ for exposition sake, consider the decomposition (7.19), and use Lemma 7.1 to obtain

$$\begin{aligned}
\widehat{S}'\widehat{B}W\widehat{B}\widehat{S} &= O_p(1) \\
\left(\widehat{E} - \widehat{F}\right)' W \left(\widehat{E} - \widehat{F}\right) &= O_p(1) \\
\left(\widehat{E} - \widehat{F}\right)' W\widehat{B}\widehat{S} &= O_p(1) \\
\frac{r_n}{\sqrt{n}}\widehat{F}'W\widehat{B}\widehat{S} &= O_p(1) \\
\frac{r_n}{\sqrt{n}}\widehat{F}'W \left(\widehat{E} - \widehat{F}\right)' &= O_p(1)\,.
\end{aligned}$$

Moreover, $r_n^2 n^{-1}\widehat{F}'W\widehat{F} \xrightarrow{p} \tilde{C}$ where

$$\tilde{C} = \tilde{c}^2(A_0'\Omega^{-1}(Z)A_0)^{-1/2}\mathbb{E}\left[A_0'\Omega^{-1}C(Z_1)C'(Z_2)\Omega^{-1}A_0 w(Z_1 - Z_2)\right](A_0'\Omega^{-1}(Z)A_0)^{-1/2}\,.$$

Gathering results, for $r_n^2/n = O_p(1)$,

$$\mathrm{ICM}(\beta_1) - \mathrm{ICM}(\beta_0) = o_p\left(\frac{n}{r_n^2}\tilde{c}^2\right) + \frac{n}{r_n^2}a_2'\tilde{C}a_2\,.$$

For strong or semi-strong instruments, this difference tends in probability to $+\infty$ because $\frac{n}{r_n^2} \to \infty$. For weak instruments, it tends in probability to $+\infty$ when $\tilde{c} \to \infty$. Since $\mathrm{CICM}(\beta_0) \leq \mathrm{ICM}(\beta_0) = O_p(1)$, and both $c(Z,\alpha)$ and $c(Z,\widehat{E}(\beta_1),\alpha)$ are stochastically bounded, we obtain the desired result.

References

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

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 (2017): "A Note on Optimal Inference in the Linear IV Regression Model." *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.

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.

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. Chapter 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.

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

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. (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.

Johnson, N., S. Kotz, and N. Balakrishnan (1995): *Continuous Univariate Distributions*, vol. 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.

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," *arXiv:1505.06644 [math, stat]*, 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," *ArXiv e-prints*.

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. and J. Alix-Garcia (2017): "Labor Scarcity, Land Tenure, and Historical Legacy: Evidence from Mexico," *Texas A&M University*.

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, 295325.

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

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

van der Vaart, A. (1994): "Bracketing Smooth Functions," *Stochastic Processes and Their Applications*, 52, 93–105.

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
Homoskedastic	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 5 for a theoretical 10% level.

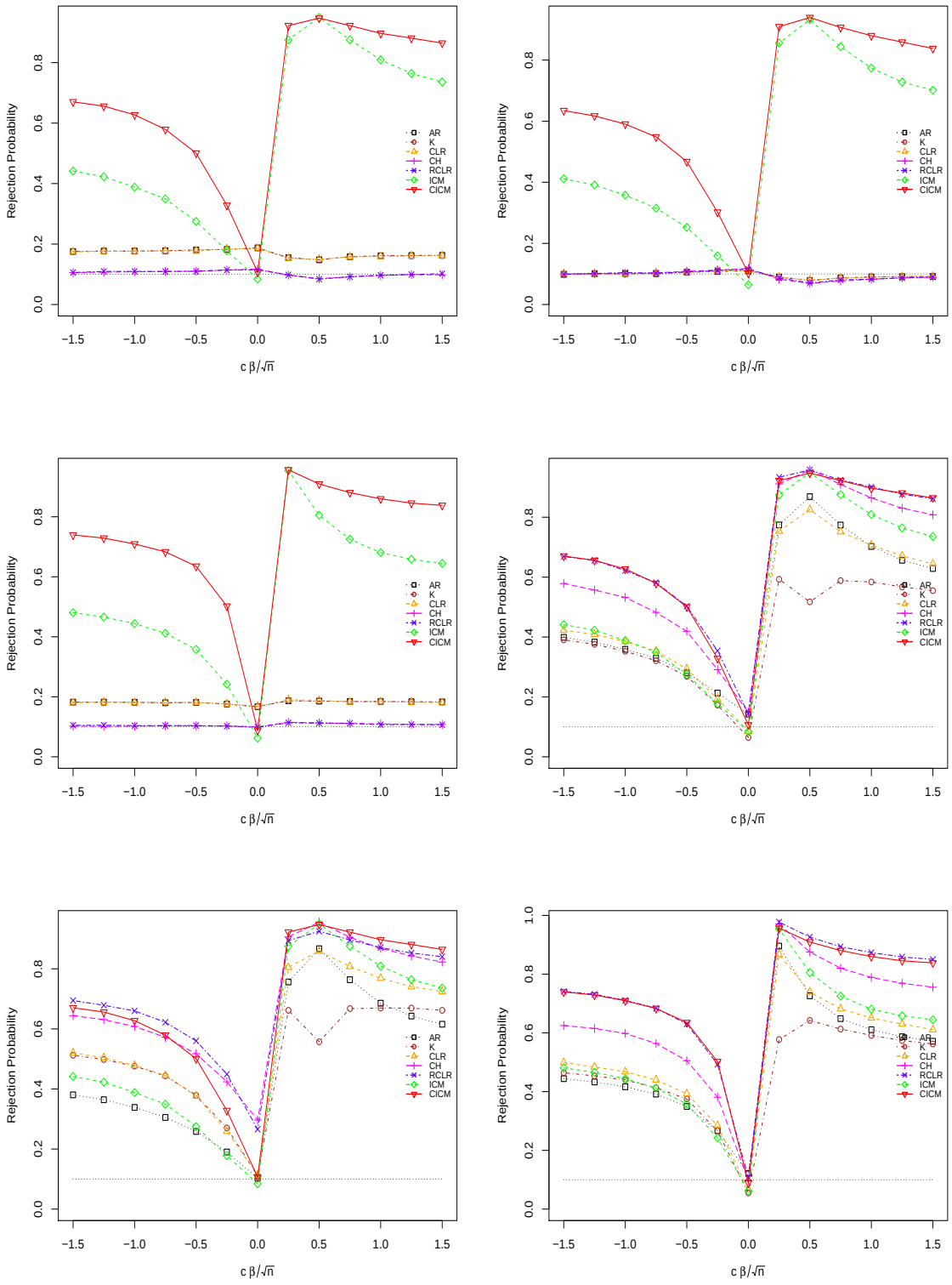

Figure 1: Power curves for Polynomial Model (i): benchmark (top left), homoskedastic case (top right), sample size 401 (middle left), 3 IV (middle right), 7 IV (bottom left) and 3 IV with sample size 401 (bottom right).

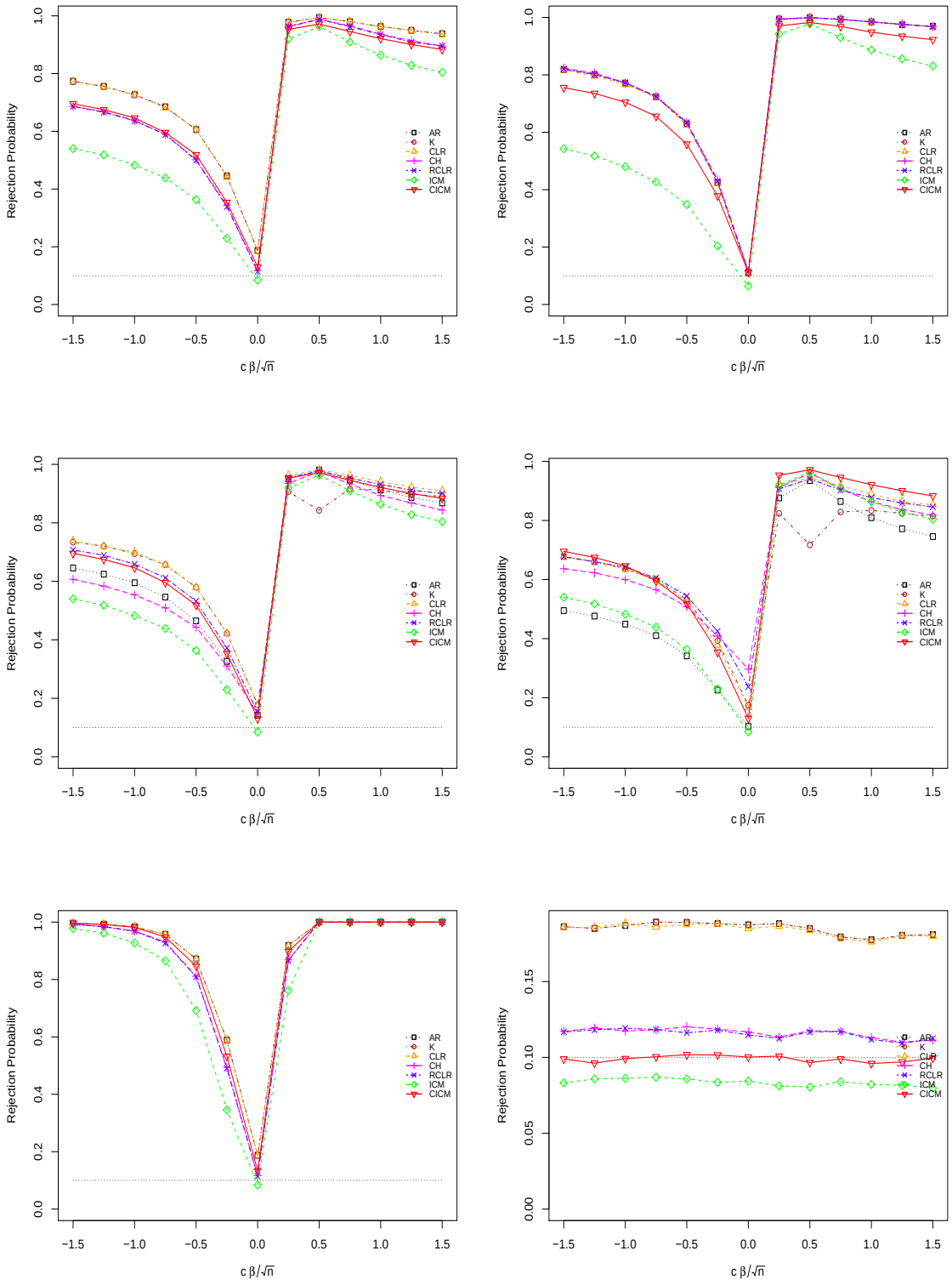

Figure 2: Power curves for Linear Model (i): benchmark (top left), homoskedastic case (top right), 3 IV (middle left), 7 IV (middle right), stronger identification (bottom left) and no identification (bottom right).

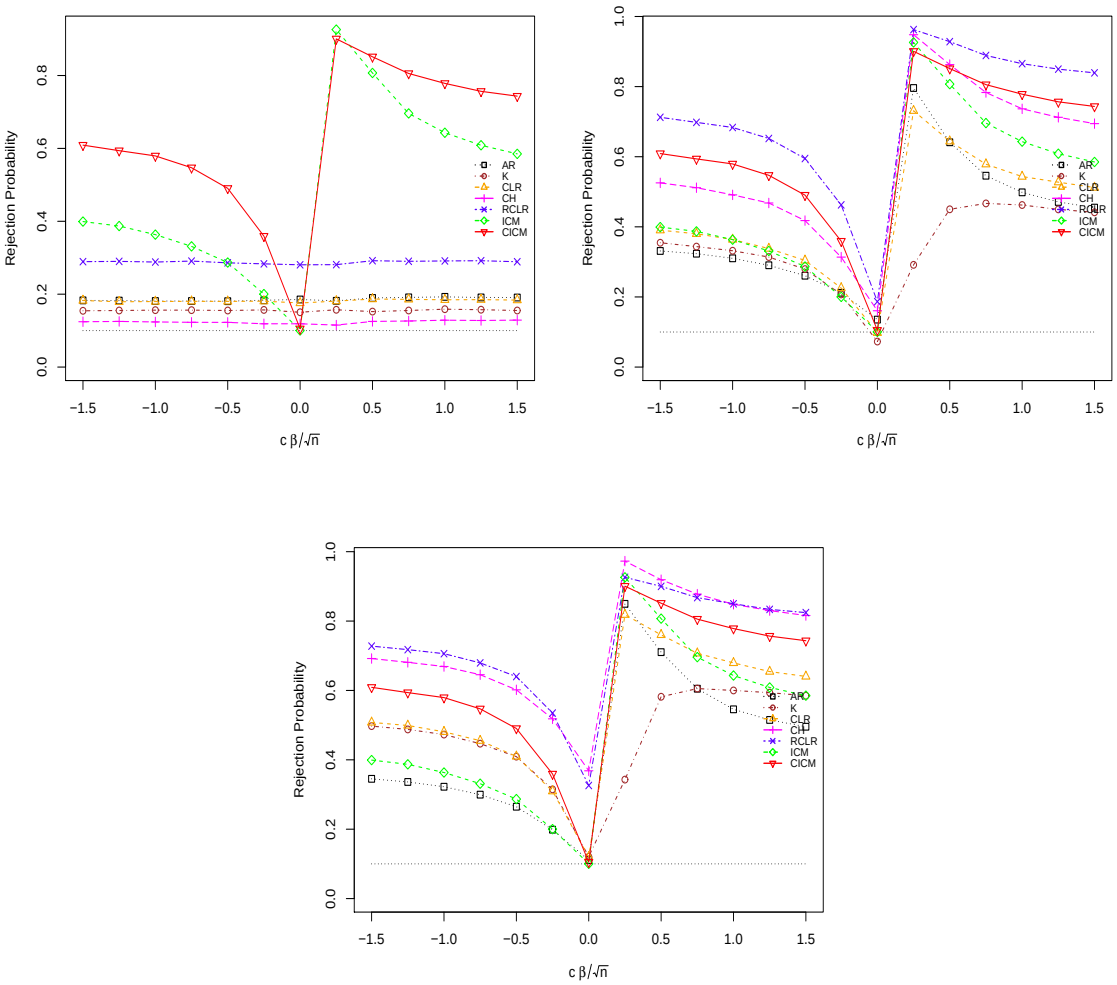

Figure 3: Power curves for Group Heterogeneity Model (ii): benchmark (top left), 7 IV (top right), and 15 IV (bottom).

ICM	$\emptyset$
CICM	[-0.615, -0.220]
Inference procedures based on (6.14) with 3 instruments	
TSLS	[-0.306, -0.075]
AR	$\emptyset$
CLR	[-0.899, -0.366]
CH	$\emptyset$
RCLR	[-1.980, -0.945]
Inference procedures based on (6.14) with 9 instruments	
TSLS	[-0.301, -0.100]
AR	$\emptyset$
CLR	[-0.610, -0.280]
CH	$\emptyset$
RCLR	[-0.773, 0.153]
Inference procedures based on (6.14) with 18 instruments	
TSLS	[-0.216, -0.034]
AR	$\emptyset$
CLR	[-0.419, -0.132]
CH	$\emptyset$
RCLR	[-0.110, 0.063]

Table 2: 95% Confidence Intervals for the population collapse with a sample size equal to 1030. The regions for ICM, CICM, CH and RCLR are obtained with 599 replications, while the regions for TSLS, AR, and CLR are obtained using `ivmodel` from R.

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.