A Pivotal Nonparametric Test for Identification-Robust Nonparametric Inference in Linear IV Models

Hippolyte Boucher*

September 5, 2022

Abstract

In linear models with endogenous regressors it is well-known that weak instruments (IVs) bias the 2 Stage Least Squares (2SLS) and other k-class IV estimators and make standard Gaussian confidence intervals invalid. Inference can still be performed by inverting tests, however there are no known method to account for a non-linear first stage except Antoine and Lavergne (2019). Their method requires simulations of the distribution of the test statistic under the null which makes it difficult to apply when sample size is moderate to large. For the above reasons I build a pivotal test statistic based on a score of integrated conditional moments which allows to easily infer on the model's structural parameters regardless of instruments' strength and the shape of the first stage conditional mean. For heteroskedastic or independent and identically distribution data with normal or non-normal errors I prove that the test is valid regardless of the degree of identification of the structural parameter of interest, and also prove that the test is consistent as long if the parameter of interest is at least semi-strongly identified. I compare the performances of the test against competing ones and revisit the effect of education on wage using Angrist and Krueger (1991) data and prove that it is strictly positive.

Keywords: Weak Instruments, Hypothesis Testing, Semiparametric Model

JEL Codes: C12, C13, C14

* *Toulouse School of Economics. Email: hippolyte.boucher@ut-capitole.fr*

1 Introduction

Consider the linear model with endogenous variables popular in reduced form econometrics

$$y_i = x_i'\beta + z_{1i}'\gamma + u_i \qquad \mathrm{E}(u_i|z_{1i}, z_{2i}) = 0 \quad i = 1, \ldots, n \tag{1.1}$$

where x are endogenous variables, z_1 are exogenous control variables, and z_2 are exogenous instrumental variables. One would like to infer on the structural parameter β. Starting from a controversial application by Angrist and Krueger (1991) and a critique by Bound, Jaeger, and Baker (1995) it has been shown that if the correlation between instruments and endogenous regressors is small then standard asymptotic approximations of the distribution of IV estimators are unreliable both in small and large samples. From there, alternative asymptotic frameworks were developed to account for potentially weak identification or weak instruments, such as in the seminal paper by Staiger and Stock' (1997), so that robust tests and inference may still be performed, see e.g. , Anderson and Rubin (1949), Stock and Wright (2000), Moreira (2003), Kleibergen (2002, 2005), Andrews and Cheng (2012), Andrews (2016), and Andrews and Mikusheva (2016a,b). Important surveys on weak identification include Stock, Wright, and Yogo (2002), Dufour (2003), Hahn and Hausman (2003), and Andrews and Stock (2005). These procedures are robust to instrument strength and rely on a parametric and often linear approximation of the first-stage equation. But using a linear approximation of the first stage leads to a loss of information and thus lowers IV strength, in a recent paper Dieterle and Snell (2016) highlights how in a variety of applications adding polynomials and cross-products of the instruments change the 2SLS estimates significantly. Consequently, in the context of weak instruments, it would be preferable to consider a non-linear first stage but as Jun and Pinkse (2012) have shown, using a nonparametric estimate of the first-stage conditional mean does not allow to obtain a valid confidence interval for β when instruments are weak. This negative results extends to nonparametric IV estimation procedures such as Newey and Powell (2003) or Darolles, Florens, and Renault (2011). In fact estimating the first stage conditional mean nonparametrically typically results in a weak-identification robust confidence interval which is wider than if the first stage was considered linear because the conditional mean is too flat and the number of IVs too large, see Dieterle and Snell (2016). For the above reasons Antoine and Lavergne (2019) came up with an inference procedure based on a test which leaves the first stage equation unspecified

and unestimated while being robust to weak identification. Their test statistic uses the methodology of integrated conditional moments but is not pivotal hence its the distribution has to simulated. This makes their test hard to apply in practice when sample size becomes large ($n > 10,000$) as is most common in applied microeconomics papers and in the application of this paper.

Consequently I develop a pivotal test statistic for weak-identification robust inference in linear IV models with an unspecified first stage. This allows applied researchers to easily infer on the linear effect of endogenous variables on outcomes regardless of the nature of the relationship between the instruments and the endogenous variables. To create this test my approach resembles that of Antoine and Lavergne (2019): I combine the integrated conditional moment specification test of Bierens (1982) with the Lagrange multiplier test of Kleibergen (2005) (LM) to create KICM for Kleibergen integrated conditional moment, whereas they consider an integrated conditional moment version of Anderson and Rubin (1949) (AR) called ICM for integrated conditional moment and of the conditional likelihood ratio test of Moreira (2003) (CLR) called CICM for conditional integrated conditional moment. ICM and CICM are asymptotic tests, therefore they will perform well only in larger samples and yet they are also non-pivotal tests hence their critical values depend on the null hypothesis being tested and have to be simulated. This means that for large samples with possibly a few endogenous regressors it is very computationally costly to invert the ICM or CICM tests. On the contrary KICM is chi square with degrees of freedom equal to the number of endogenous regressors at the limit under the null with iid or heteroskedastic non-normal errors and a fixed number of instruments which makes inversion for inference relatively easy. In addition it is known that the LM test is more robust to many and many weak instruments, see Hansen, Hausman, and Newey (2008), compared the AR and the CLR, this result carries on for their integrated conditional moment versions. These advantages shine in the simulations and application of this paper: With 4 instruments KICM has no size distortion compared ICM and CICM. The ICM and CICM tests cannot be inverted to infer on the effect of schooling on wages because sample size is above $100,000$.

In the second section of this paper I formally introduce the model, the existing tests, and motivate KICM. The third section is devoted to the derivation of the KICM test statistic from the null hypothesis and its implementation. The fourth establishes

the validity and consistency of KICM for iid and heteroskedastic data. In the fifth section I perform an simulation exercise to assess KICM performances. In the sixth section I perform inference on the return to schooling on salary using the data from Angrist and Krueger (1991). I conclude in a seventh and final section. Proofs are in section A, B, C and D of the appendix, tables and plots from the simulations and the application are in section E of the appendix.

2 Framework

The objective is to infer on the effect β of l endogenous variables x_i on an outcome y_i by testing null hypotheses of the form $H_0 : \beta = \beta_0$ for some $\beta_0 \in \mathbb{R}^l$. Without loss of generality exogenous control variables are projected out'a la Frisch-Waugh consequently in the rest of the paper I consider the following structural equation

$$y_i = x_i'\beta + u_i \qquad \mathrm{E}(u_i|z_i) = 0 \qquad i = 1, \ldots, n \tag{2.2}$$

which is augmented by a first-stage reduced form equation for x_i with $k > l$ exogenous instruments z_i

$$x_i = \Pi(z_i) + v_i \qquad \mathrm{E}(v_i|z_i) = 0 \qquad i = 1, \ldots, n \tag{2.3}$$

z_i may also include some of the exogenous controls if one also suspects $\Pi(\cdot)$ to be non-linear in those. I denote by y, x, z and $\Pi(z)$ the stacked versions the versions of $(y_i, x_i', z_i', \Pi(z_i)')$ over the observations $i = 1, \ldots, n$ so that y is of dimension $n \times 1$, x of dimension $n \times l$, z of dimension $n \times k$, and $\Pi(z)$ of dimension $n \times l$.

$\Pi(\cdot)$ may be "close to zero" so that z is weakly related to x. This weak instruments problem prevents consistent estimation of β and renders inference using standard Gaussian confidence interval invalid. Valid inference is still possible by inverting weak-identification robust tests but those may yield conservative confidence interval as they don't account for non-linearities in the first stage. In the next subsections I first briefly present the weak instruments problem, second I review the most popular methods for weak-identification robust inference, and third I motivate the use of KICM.

2.1 The weak instruments problem

Consider the setting described by (2.2) and (2.3) and assume for exposition that $(y_i, x_i, z_i)_{i=1}^n$ is iid and that $\Pi(\cdot)$ is linear and injective, ie $\Pi(z_i) = \Pi' z_i$ with Π a

full rank $k \times l$ matrix. To estimate β one will use a k-class estimator such as 2SLS but when the instruments are weak as in Π is close to being singular the estimators mentioned above are biased and the traditional inference procedures become unreliable even in large samples. The literature has largely expanded upon these types of problems, see the surveys by Stock et al. (2002), Dufour (2003), Hahn and Hausman (2003), and Andrews and Stock (2005), and has coined different types of weak instruments asymptotics in order to model these problems. Because I consider a finite number of instruments k, I follow the terminology of Andrews and Cheng (2012) and without loss of generality allow instruments to be very weak, weak, semi-strong and strong

$$\Pi \equiv \frac{1}{n^a} C \tag{2.4}$$

where C is a $k \times l$ full rank matrix and a is positive or infinite. a represents instruments strength so that when $a = 0$ the instruments are deemed strong and β is strongly identified, when $a \in (0; 1/2)$ the instruments are deemed semi-strong and consequently β is semi-strongly identified, when $a = 1/2$ the instruments are deemed weak and thus β is weakly identified, and when $a > 1/2$ the instruments are deemed very weak and therefore β is very weakly identifiedthis is an abuse of words, as long as $a < +\infty$ the structural parameter β is point-identified however depending of the strength of the instruments β may not be consistently estimated which is why this terminology is used. When the instruments are weak or very weak $a \geqslant 1/2$ then k-class estimators such as 2SLS lose their consistency and their asymptotic normality.

Because consistent estimation is too difficult in case of weak instruments even with regularization, the literature has focused on providing inference robust to weak instruments by inverting tests. I introduce the most famous tests in the literature then show with a simple example that, because they do not take into account many non-linearties in the first stage, they decrease identification strength of β which is why KICM is needed.

2.2 Existing tests

Define

$$Y = \begin{pmatrix} y_1 & x_1' \\ y_1 & x_2' \\ \vdots & \vdots \\ y_n & x_n' \end{pmatrix}, \ \Omega_i \equiv \Omega(z_i) = \mathrm{Var}(Y_i|z_i) = \begin{pmatrix} \mathrm{Var}(y_i|z_i) & \mathrm{Cov}(y_i, x_i|z_i) \\ \mathrm{Cov}(x_i, y_i|z_i) & \mathrm{Var}(x_i|z_i) \end{pmatrix} = \mathrm{Var}(v_i'\beta + u_i \quad v_i'|z_i)$$

$$b_0 = [1 \ -\beta_0']', \qquad A_0 = [\beta_0 \ I_l]'$$

Assuming the data is homoskedastic, ie $\Omega(z_i) = \Omega$, then one can also define

$$S \equiv S(\beta_0) = Yb_0 \left(b_0'\Omega b_0\right)^{-1/2}, \qquad T \equiv T(\beta_0) = Y\Omega^{-1}A_0 \left(A_0'\Omega^{-1}A_0\right)^{-1/2}$$

Notice that b_0 has dimension $(l+1) \times 1$ and A_0 has dimension $(l+1) \times l$ so T has dimension $n \times l$ whereas S has dimension $n \times 1$. Most notably, these notations imply that $Yb_0 = y - x\beta_0$ and that if the data is homoskedastic $\mathrm{Var}(y_i - x_i'\beta_0|z_i) = b_0'\Omega b_0$.

Anderson and Rubin (1949) were the first to address the issue of inference with weak instruments under the assumption of homoskedacity and linearity in the first stage, without resorting to estimating the first stage correlation coefficient Π. The principle is the following: for different β_0, $y - x\beta_0$ is regressed on z and a test of joint significance of z is performed, then all the values of β_0 for which the test is not rejected form the confidence interval of β. To test $H_0 : \beta = \beta_0$ rewrite model (2.2)

$$Yb_0 = y - x\beta_0 = x(\beta - \beta_0) + u \equiv z\delta_0 + u_0$$

then the AR test statistic is the Wald statistic which tests $H_0 : \delta_0 = 0$

$$\mathrm{AR} \equiv \mathrm{AR}(\beta_0) = \frac{b_0'Y'P_zYb_0}{b_0'\Omega b_0} = S'P_zS$$

As z has rank k so does P_z which implies that under the null $H_0 : \beta = \beta_0$ and assuming linearity in the first stage and homoskedasticity $\mathrm{AR} \xrightarrow{d} \chi^2_k$. This holds whether or not the errors are normals and if Ω is replaced by a consistent estimator such as $\frac{1}{n-k}Y'M_zY$.

Later came Kleibergen (2002) Lagrange Multiplier statistic

$$\text{LM} = S'P_{P_zT}S = S'P_zT(T'P_zT)^{-1}T'P_zS$$

derived from limited information maximum likelihood criterion. P_{P_zT} has rank l thus $\text{LM} \xrightarrow{d} \chi^2_l$ under the null, linearity of $\Pi(\cdot)$ and homoskedastic errors.

Moreira (2003) coined a conditional likelihood-ratio statistic

$$\text{CLR} = S'P_zS - \lambda_{min}(\begin{pmatrix} S'P_zS & S'P_zT \\ T'P_zS & T'P_zT \end{pmatrix})$$

where $\lambda_{min}(\cdot)$ is the minimum eigenvalue. In general the asymptotic distribution of CLR has to be simulated and depends on β_0.

Note that $H_0 : \beta = \beta_0$ is equivalent to $\delta_0 = 0$ if and only if Π is non-singular hence if Π is singular or if the instruments are very weak $a > 1/2$, see (2.4), then the CI derived from any of these procedures will be of infinite length. Because the limiting distribution of LM does not depend on the number of instruments k but on the number of endogenous regressors l inference using LM yields confidence intervals with better coverage than with the AR and CLR if k is moderate or large. One downside of LM is that the confidence intervals for β may be the union of 2 or 3 intervals. When k is small it can be shown that CLR has better power than the other 2 and thus the confidence interval built from it has smaller length.

There exists heteroskedasticity robust and some auto correlation robust versions of the AR, LM and CLR tests, see e.g. Andrews, Moreira, and Stock (2004), Kleibergen (2007), Chernozhukov and Hansen (2008), Moreira and Moreira (2015) and Andrews and Mikusheva (2016a), tests which are similar in purpose but specific to other types of models such as the generalized empirical likelihood test of Guggenberger and Smith (2005) also exists. As noted by Dufour and Taamouti (2007) the AR test is "robust to misspecification" in the first stage unlike the LM and CLR tests: As long as there is one instrument left, if an instrument is not included in the first stage then the AR test will still be χ^2 distributed under the null thus it will have correct size. Tests which allow for non conservative inference on subvectors of β also exist, see Guggenberger, Kleibergen, Mavroeidis, and Chen (2012), Guggenberger, Kleibergen, and Mavroeidis (2019).

2.3 Motivation

An important issue with these tests is that they only consider the linear relationship between the endogenous variables and the instruments hence non linearities remain in part undetected leading to a loss of power thus larger confidence intervals.

As an example consider the following scalar-IV model $l = k = 1$ with iid data homoskedastic data $(y_i, x_i, z_i)_{i=1}^n$ with Ω known and $z_i \sim \mathcal{N}(0, 1)$

$$y_i = x_i\beta + u_i, \qquad x_i = z_i^2 + v_i, \qquad \mathrm{E}(v_i|z_i) = \mathrm{E}(u_i|z_i) = 0$$

Then notice that the best linear projection of x_i on z_i denoted as $BLP(x_i|z_i)$ equals 0

$$BLP(x_i|z_i) = z_i' \,\mathrm{E}(z_i z_i')^{-1} \,\mathrm{E}(z_i x_i) = z_i' \,\mathrm{E}(z_i z_i')^{-1} \,\mathrm{E}(z_i^3) = 0$$

because $\mathrm{E}(z_i^3) = 0$. Similarly the projection y_i on z_i equals 0

$$BLP(y_i|z_i) = z_i' \,\mathrm{E}(z_i z_i')^{-1} \,\mathrm{E}(z_i y_i) = z_i' \,\mathrm{E}(z_i z_i')^{-1} \,\mathrm{E}(z_i^3\beta + z_i v_i\beta + z_i u_i) = 0$$

As a consequence, instruments are considered irrelevant by the AR, LM and CLR tests because all of them are quadratic functions of $Y'P_zY$, thus confidence intervals built from them are the whole real line. More precisely using the law of large numbers (LLN) and the central limit theorem (CLT) it can be shown that

$$Y'P_zY = \frac{1}{\sqrt{n}}Y'z\left(\frac{1}{n}z'z\right)^{-1}\frac{1}{\sqrt{n}}z'Y = O_{\mathrm{P}}(1)$$

where $O_{\mathrm{P}(1)}$ is the big O in probability notation for bounded in probability[1]. Thus AR cannot explode under the alternative $H_1 : \beta \neq \beta_0$ because it is bounded in probability, it has no power hence a confidence interval built from it will be very large.

In a more general case with the possibility of semi-strong and weak instruments, only considering linearities in the first stage as in the AR, LM and CLR can only exacerbate the issue of instrument weakness even in the best of cases while in the worst as in the example above it can make the instruments completely irrelevant. For this reason a new test which takes into account non-linearities in the first stage such as KICM is needed.

[1]Formally if $X = O_{\mathrm{P}(1)}$ then $\forall\varepsilon > 0 \exists M : \mathrm{P}(|X| > M) \leqslant \varepsilon$. If $X = o_{\mathrm{P}(1)}$ then X is degenerate in probability and $\forall\varepsilon > 0, \mathrm{P}(|X| > \varepsilon) \to 0$.

3 Building KICM

I derive the KICM statistic in two steps: First, I consider a conditional moment null hypothesis and prove that it is equivalent to an integrated conditional moment hypothesis as in Bierens (1982). Second, using the ICM statistic of Antoine and Lavergne (2019) as a criterion I build KICM which is a transformation of ICM's score. Then I present the feasible versions of KICM for both homoskedastic and heteroskedastic data. From now on and in the rest of the paper I consider the model characterized by (2.2) and (2.3) and unless specifically mentioned I do not assume that $\Pi(\cdot)$ is linear and that the data is iid.

3.1 From a conditional moment to an integrated conditional moment

Recall the model characterized by (2.2) and (2.3)

$$y_i = x_i'\beta + u_i \qquad \mathrm{E}(u_i|z_i) = 0 \tag{2.2}$$

$$x_i = \Pi(z_i) + v_i \qquad \mathrm{E}(v_i|z_i) = 0 \tag{2.3}$$

Then to test $H_0 : \beta = \beta_0$ the structural equation should be rewritten

$$y - x\beta_0 = \Pi(z)(\beta - \beta_0) + u_0, \qquad u_0 = v(\beta - \beta_0) + u$$

As a consequence $H_0 : \beta = \beta_0$ implies that $H_0^1 : \mathrm{E}(y_i - x_i'\beta_0|z_i) = 0$ a.s which turns into an equivalence under specific conditions.

Using H_0^1 directly is not possible so instead I use the "Fourier" transformation from Bierens (1982) to obtain an equivalent many moments condition $H_0^2 : \mathrm{E}((y_i - x_i'\beta_0)e^{it'z_i}) = 0 \ \forall t \in \mathbb{R}^k$. One may interpret H_0^2 as the true error being 0 on average for any possible direction of the instruments or equivalently for any possible "additive combination" of the moments of the instruments. Indeed

$$\forall t \in \mathbb{R}^k \quad exp(it'z_i) = cos(t'z_i) + i\ sin(t'z_i) = exp(i\sum_{j=1}^{k} z_i^{t_k})$$

Alternatives to H_0^2 could be used such as a many moments condition with check functions, however using the complex exponential will allow to formulate the test in a

simple matrix form and makes it pivotal.

The condition H_0^2 is equivalent to $H_0^3 : |\,\mathrm{E}((y_i - x_i'\beta_0)e^{it'z_i})|^2 = 0\ \forall t \in \mathbb{R}^k$ where $|\ \cdot\ |$ denotes the modulus. Finally H_0^3 is equivalent to H_0^4, an integrated version of the many moments conditions over the t to only have 1 final moment so that the null $H_0 : \beta = \beta_0$ is equivalent to

$$H_0^4 : \int_{\mathbb{R}^k} |\,\mathrm{E}((y_i - x_i'\beta_0)e^{is'z_i})|^2 d\mu(s) = 0$$

where μ is a (finite) measure with support $\mathbb{R}^k$ which is positive almost everywhere to account for all the moments. These equivalences are summarized in the following proposition.

Proposition 3.1

Assuming that (2.2) and (2.3) hold and that μ is a positive measure almost everywhere on $\mathbb{R}^k$ then

$$\begin{aligned} H_0 : \beta = \beta_0 \Rightarrow H_0^1 : \mathrm{E}(y_i - x_i'\beta_0|z_i) = 0\ a.s\ &\Leftrightarrow H_0^2 : \mathrm{E}((y_i - x_i'\beta_0)e^{it'z_i}) = 0\ \forall t \in \mathbb{R}^k \\ &\Leftrightarrow H_0^3 : |\,\mathrm{E}((y_i - x_i'\beta_0)e^{it'z_i})|^2 = 0\ \forall t \in \mathbb{R}^k \\ &\Leftrightarrow H_0^4 : \int_{\mathbb{R}^k} |\,\mathrm{E}((y_i - x_i'\beta_0)e^{is'z_i})|^2 d\mu(s) = 0 \end{aligned}$$

Moreover if $l = 1$ and $\mathrm{P}(\Pi(z_i) = 0) = 0$ then

$$H_0 \Leftrightarrow H_0^1 \Leftrightarrow H_0^2 \Leftrightarrow H_0^3 \Leftrightarrow H_0^4$$

The proof is in A.1 of the appendix. It is then straightforward to build a test statistic for H_0 from H_0^4. Note that in general H_0^4 tests an implication of H_0 like the previously mentioned tests, when $\Pi(\cdot)$ is small the test has low power and the confidence interval built from it is large.

3.2 From ICM to KICM

To test $H_0 : \beta = \beta_0$ an empirical counterpart of $H_0^4 : \int_{\mathbb{R}^k} |\,\mathrm{E}((y_i - x_i'\beta_0)e^{is'z_i})|^2 d\mu(s) = 0$ is taken, then multiplying by n and standardizing allows the CLT to apply to the integrand, this is the ICM statistic of Antoine and Lavergne (2019) which writes

$$\text{ICM} \equiv \text{ICM}(\beta_0) = \int_{\mathbb{R}^k} |n^{-1/2} \sum_{i=1}^{n} \frac{y_i - x_i'\beta_0}{\text{Var}(y_i - x_i'\beta_0|z_i)^{1/2}} e^{is'z_i}|^2 \mu(s)$$

ICM can actually be written as a function of S, let W be a $n \times n$ matrix with elements $W_{ij} = n^{-1} w(z_i - z_j)$ such that

$$w(z) = \int_{\mathbb{R}^k} e^{is'z} d\mu(s) \tag{3.5}$$

The condition for μ to have support $\mathbb{R}^k$ translates into the restriction that $w(\cdot)$ should have a Fourier transform which is strictly positive almost everywhere, or if the support of the instruments z is bounded, that its Fourier transform is well-defined in a neighborhood of 0, see theorem 1 in Bierens (1982). The choice of $w(\cdot)$ thus includes products of densities such as triangular, normal, or logistic, see Johnson, Kotz, and Balakrishnan (1995), Student, including Cauchy, see Dreier and Kotz (2002), or Laplace. Using properties of the modulus it is simple to show that $\text{ICM} = S'WS$.

Thus ICM ressembles $\text{AR} = S'P_zS$ but W is not a projection matrix hence ICM is not pivotal asymptotically and its distribution depends on β_0. Similarly from Antoine and Lavergne (2019) CICM is the integrated conditional moment equivalent of the CLR of Moreira (2003) and is not pivotal asymptotically. To derive the KICM test statistic which is pivotal asymptotically I use ICM as a criterion function and derive its score which I then standardize: Taking (y, x, z) as deterministic notice that the convex function

$$\beta \mapsto \frac{b'Y'WYb}{b'\Omega b}, \quad b = (1 \ -\beta')'$$

is minimized uniquely at $\beta = \beta_0$ under the null hypothesis. Thus taking the first order condition at β_0 yields

$$\begin{aligned}
\frac{\partial}{\partial\beta} b_0 \times \left(\frac{Y'WYb_0}{b_0'\Omega b_0} - \frac{\Omega b_0 b_0' Y'WYb_0}{(b_0'\Omega b_0)^2}\right) = 0 \ &\Leftrightarrow \frac{Y'WYb_0}{b_0'\Omega b_0} - \frac{\Omega b_0 b_0' Y'WYb}{(b_0'\Omega b_0)^2} = 0 \\
&\Leftrightarrow \frac{(A_0'\Omega^{-1}A_0)^{-1/2} A_0'\Omega^{-1} Y'WYb_0}{(b_0'\Omega b_0)^{1/2}} = 0 \\
&\Leftrightarrow T'WS = 0
\end{aligned}$$

where the second line is obtained by multiplying by $(A_0'\Omega^{-1}A_0)^{-1/2}A_0'\Omega^{-1}$ and using the fact that $A_0'b_0 = 0_l$. I prove later that $\text{E}(S'WT|z) = 0$ when (y, x, z) is random. Finally KICM is a quadratic version of the score $S'WT$ standardized with respect to WT

$$\text{KICM} = S'WT(T'W^2T)^{-1}T'WS = S'P_{WT}S$$

where P_{WT} is the orthogonal projection matrix on WT which gives the statistic its chi square l degrees of freedom asymptotic distribution.

3.3 KICM in practice

In practice in order to use KICM properly several elements are still needed:

First I greatly simplify the choice of $w(\cdot)$ by imposing that its Fourier transform is a real symmetric density which is strictly positive almost everywhere, or around 0 if z has bounded support, and that the L_2 norm of $w(\cdot)$ equals 1. As a consequence $w(\cdot)$ is also a symmetric real bounded density, thus possible choices for $w(\cdot)$ are Triangular, Logistic, Cauchy or Laplace distribution densities (see Johnson et al. (1995)). Imposing that the Fourier transform is a centered density cleverly prevents having too many instruments, puts more weights on lower moments of z and make sure no cardinal direction of moments of z is favored. On the other hand making sure that the squared norm of $w(\cdot)$ equals 1 ensures that the elements of W do not scale with sample size.

Second $\Omega(z_i) = \text{Var}(Y_i|z_i)$ must be estimated consistently. Assuming it is linear in z then simply using the parametric estimator $\hat{\Omega} = \frac{1}{n}Y'M_zY$ is a good idea. If this assumption is too strong one may use a semi-parametric or non-parametric estimator, e.g. from Seifert, Gasser, and Wolf (1993) or from Yin, Geng, Li, and Wang (2010). I use the later in the simulations and application. It writes

$$\hat{\Omega}(z) = \frac{\frac{1}{nh}\sum_{i=1}^{n}(Y_i - \bar{Y}(z))(Y_i - \bar{Y}(z))'K((z_i - z)/h)}{\frac{1}{nh}\sum_{i=1}^{n}K((z_i - z)/h)},\ \bar{Y}(z) = \frac{\frac{1}{nh}\sum_{i=1}^{n}Y_iK((z_i - z)/h)}{\frac{1}{nh}\sum_{i=1}^{n}K((z_i - z)/h)}$$

with the bandwidth h chosen properly to allow convergence. If data is homoskedastic the estimator is the following average $\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n}\hat{\Omega}(z_i)$.

Third the instruments should be standardized. A desirable property of weak-identification robust tests is that of invariance to orthogonal transformations of the instruments which allows the tests to be invariant to instruments' scale (see Andrews and Stock (2007)). KICM cannot satisfy this property however by standardizing the instruments a priori the same effect can be obtained. Additionally in the literature on nonparametric estimation via Kernels, regressors are standardized through the bandwidth (see Li and Racine (2006)) and here $w(\cdot)$ has the role of a Kernel function.

Feasible tests Based on the above the feasible KICM statistic for homoskedastic data $\text{Var}(Y_i|z_i) = \Omega$ writes

$$KICM_f = S_f' P_{WT_f} S_f, \quad S_f = Y b_0 (b_0' \hat{\Omega} b_0)^{-1/2}, \quad T_f = Y \hat{\Omega}^{-1} A_0 (A_0' \hat{\Omega}^{-1} A_0)^{-1/2}$$

where W has elements $W_{ij} = \frac{1}{n} w(z_i - z_j)$ with $w(\cdot)$ a density satisfying the aforementioned conditions and where $\hat{\Omega}$ is a consistent estimator of Ω.

In case of heteroskedastic data $\text{Var}(Y_i|z_i) = \Omega(z_i) = \Omega_i$ I first define the heteroskedasticity robust version of KICM

$$\text{KICM}_h = S_h' P_{WT_h} S_h, \quad \forall i \;\; S_{ih} = Y_i' b_0 (b_0' \Omega(z_i) b_0)^{-1/2}, \quad T_{ih}' = Y_i' \Omega(z_i)^{-1} A_0 (A_0' \Omega(z_i)^{-1} A_0)^{-1/2}$$

with S_h and T_h the stacked versions of S_{ih} and T_{ih} respectively. This change allows S_{ih} and T_{ih} to be properly standardized in the heteroskedastic case. Then the feasible version of the heteroskedasticity robust KICM statistic writes

$$\text{KICM}_{hf} = S_{hf}' P_{WT_{hf}} S_{hf}, \quad \forall i \;\; S_{ihf} = Y_i' b_0 (b_0' \hat{\Omega}(z_i) b_0)^{-1/2}, \quad T_{ihf}' = Y_i' \hat{\Omega}(z_i)^{-1} A_0 (A_0' \hat{\Omega}(z_i)^{-1} A_0)^{-1/2}$$

where $\hat{\Omega}(\cdot)$ is a consistent estimator of $\Omega(\cdot)$.

Weak-identification robust inference With a feasible KICM test statistic in hand it is now possible to infer on β regardless of its degree of identification. To do so the econometrician has to invert the KICM test. First they need to select a nominal coverage $1-\alpha$ for the confidence interval and a grid over $\mathbb{R}^l$ from which they will test different values of β_0. Then the econometrician must compute the KICM feasible statistic for all values of β_0 over the grid. Finally all values of β_0 for which the statistic is above the $1-\alpha$ quantile of a chi square with l degrees of freedom will constitute the $1-\alpha$ confidence interval of β.

The next section is devoted to formal results on the distibution and the asymptotic behavior of KICM.

4 KICM validity and consistency

In this section I first introduce assumptions necessary for the asymptotic theory then I prove that KICM is chi square distributed under normality of the errors under the null,

and lastly I prove validity and consistency of KICM without normality of the errors. The theoretical coverage probabilities of the confidence interval built from KICM are direct implications of the propositions and theorem introduced in this section which is why they are omitted.

4.1 Assumptions

I first assume that the data is either iid or heteroskedastic in the sense that errors' variance are functions of the instruments. Then to obtain asymptotic results and a consistent estimator of the conditional variance $\Omega(\cdot)$ I require Y_i to have strictly more than a second conditional moment which is bounded in order to use Berry-Esseen inequalities to prove convergence.

Assumption A
(i) Observations $(y_i, x_i', z_i')_{i=1}^n$ *are independent and identically distributed*
(ii) Observations $(y_i, x_i', z_i')_{i=1}^n$ *are independent with* $(z_i')_{i=1}^n$ *also identically distributed*
(iii) $\exists \delta > 0, M > 0 : \mathrm{E}(||Y_i||^{2+\delta}|z_i) < M$

Second I make assumptions on the parameters. I assume the unique existence of a structural parameter of interest β and of some reduced form parameter $\Pi(\cdot)$. Then in order to model strong, weak and very weak identification I allow $\Pi(\cdot)$ to depend on n in two ways: Either $\Pi(\cdot) = n^{-a}C(\cdot)$ where $C(\cdot)$ is a function which does not depend on n and a represents the degree of identification of β or equivalently the degree of weakness of the instruments. The coefficient a is just a theoretical tool to study size and power when parameters have different identification strength, it is unknown in practice. Then β is strongly identified or equivalently instruments are strong when $a = 0$, β is semi-strongly identified and the instruments are semi-strong when $0 < a < 1/2$, β is weakly identified and the instruments are said to be weak in the sense of Staiger and Stock' (1997) when $a = 1/2$, β is very weakly identified and instruments are very weak when $a > 1/2$, and when $a = \infty$ instruments are irrelevant and β is not identified at all. Either $\Pi(\cdot) = N_C^{-1}C(\cdot)$ where N_C is a $l \times l$ diagonal matrix with entries which correspond to the degree of identification of each element in the vector β or equivalently the degree of weakness of the instruments with regards to each element of β

$$N_C = \begin{pmatrix} n^{a_1} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & n^{a_l} \end{pmatrix}, \quad (a_1, \dots, a_l) \in \bar{\mathbb{R}}_+^l$$

where a_j represents the degree of identification of β_j.Thus if $a_j = 0$ instruments are strong for j and β_j is strongly identified, if $0 < a_j < 1/2$ then β_j is semi-strongly identified, etc... With either of those assumptions β and its elements can be strongly identified, weakly identified or not identified at all if $a = \infty$. Lastly I assume $C(z_i)$ to have a strictly positive and finite second moment.

Assumption B
(i) There exists a unique β *and some* $\Pi(\cdot)$ *such that (2.2) and (2.3) hold*
(ii) $\Pi(\cdot) = n^{-a}C(\cdot)$ *where* $a \in \bar{\mathbb{R}}_+$
(iii) $\Pi(\cdot) = N_C^{-1}C(\cdot)$ *where* N_C *is a diagonal matrix with entries* $(n^{a_j})_{j=1}^{l}$ *with* $a_j \in \bar{\mathbb{R}}_+$
(iv) $C(\cdot)$ *does not depend on* n *and* $0 < \mathrm{E}(C(z_i)C(z_i)') < +\infty$

Next I impose conditions on $w(\cdot)$ so that by Bochner's theorem $\mu(\cdot)$ is finite and strictly positive almost everywhere. Furthermore, I assume that W is positive definite which holds in practice but cannot be formally proven without imposing more conditions on $w(\cdot)$ and z.

Assumption C
(i) $w(\cdot)$ *has a Fourier transform which is strictly positive almost everywhere, or which is strictly positive in a neighborhood of* 0 *if the support of* z_i *is bounded*
(ii) W *is positive definite almost surely for any* n

Lastly I impose conditions on the conditional covariance estimator $\hat{\Omega}$. In the homoskedastic case a simple consistent estimator for any possible DGP is needed.

In the heteroskedastic case proving the validity and consistency of KICM involves looking at random processes of Ω and t such as $(t, \Omega) \mapsto Y_i'\Omega^{-1}cos(t'z_i)$. These processes must be sufficiently smooth in Ω which is why I restrict the covariances to the class $\mathcal{O}$. Each Ω in the class $\mathcal{O}$ should be "uniformly bounded" using their minimal and maximal eigenvalues. Functions in that class also have to be smooth enough so that the class is not "too large", for any DGP $\mathcal{O}$ has a finite covering number[2] denoted $N(\varepsilon, \mathcal{O}, L_2(P))$ which should not to explode when ε gets bigger for any DGP. This assumption is necessary in order to obtain asymptotic equicontinuity uniform of the random processes involving Ω so that the difference between the feasible version KICM

[2]A ball $L_2(P)$ of size $\varepsilon > 0$ centered in $g \in L_2(P)$ writes $\{f \in L_2(P) : \int ||f - g||_2^2 dP < \varepsilon\}$, then the covering number of $\mathcal{O}$ for DGP P with interval of size ε is the minimal number of ε-balls necessary to cover all of $\mathcal{O}$

and KICM vanishes in the heteroskedastic case, see Vaart and Wellner (2000), Kosorok (2008) and the proof of theorem 4.3 in appendix B for more details. In addition estimator $\hat{\Omega}(\cdot)$ also needs to converge uniformly towards $\Omega(\cdot)$ in the L_2 sense and to belong uniformly to $\mathcal{O}$ almost surely at the limit. In the literature on convergence of nonparametric statistics with nuisance parameters this type of condition is common, see Andrews (1995), these are also necessary conditions for the difference between feasible KICM and KICM to vanish uniformly at the limit.

Assumption D
Let $\mathcal{P}$ denote the set of all distributions which satisfy assumptions A(i), A(ii) or A(iii), B(i), B(ii) or B(iii), B(iv), C(i) and C(ii)
(i) $\forall P \in \mathcal{P}\ \hat{\Omega} \xrightarrow{P} \Omega$
(ii) Define $\mathcal{O}$ such that $\forall \Omega(\cdot) \in \mathcal{O}$

$$0 < \underline{\lambda} = \inf_{s \in \mathbb{R}^l, \Omega \in \mathcal{O}} \{\lambda(\Omega(s))\} < \sup_{s \in \mathbb{R}^l, \Omega \in \mathcal{O}} \{\lambda(\Omega(s))\} = \bar{\lambda} < +\infty$$

$$\forall P \in \mathcal{P},\ \forall \varepsilon > 0,\ N(\varepsilon, \mathcal{O}, L_2(P)) < K' e^{-K\varepsilon}$$

for some $K' > 0$, $K < 2$
(iii) $\sup_{P \in \mathcal{P}} ||\hat{\Omega}(\cdot) - \Omega(\cdot)||_{L_2(P)} \to 0, \quad \sup_{P \in \mathcal{P}} \mathrm{P}(\hat{\Omega}(\cdot) \in \mathcal{O}) \to 1$

4.2 Normal Errors

For exposition I first consider conditionally normal errors or equivalently that $Y_i|z_i \sim \mathcal{N}(\mathrm{E}(Y_i|z_i), \Omega)$. Under this assumption it is relatively simple to show that under the null KICM $\sim \chi^2_l$ if the data is homoskedastic, and $\mathrm{KICM}_h \sim \chi^2_l$ if the data is heteroskedastic.

In the homoskedastic case $\mathrm{Var}(Y_i|z_i) = \Omega$ it can be shown that WT has rank l and therefore that the orthogonal projection matrix P_{WT} also has rank l. In turn using the fact that S and T are orthogonal this implies that KICM is χ^2_l distributed conditionally on (z, T) which implies that KICM is χ^2_l distributed. This result is summarized in the following proposition.

Proposition 4.1
For any $P \in \mathcal{P}$ such that $H_0 : \beta = \beta_0$, assumptions A(i), B(i), B(iii), B(iv), C(i) and C(ii) hold and assuming that $Y_i|z_i \sim \mathcal{N}(\mathrm{E}(Y_i|z_i), \Omega)$ holds then KICM $\sim \chi^2_l$

The proof is in A.2 of the appendix. As a direct corollary the feasible statistic under normal errors KICM_f is also asymptotically χ^2_l when assumptions A(iii) and D(i) are added. The reason is that, under the null, KICM_f is bounded in probability, and that S_{if} and T_{if} have covariance identity at the limit and therefore S_f and T_f are orthogonal at the limit. This result is an implication of the continuous mapping theorem (CMT) so its formal proof is omitted.

In the heteroskedastic case $\text{Var}(Y_i|z_i) = \Omega(z_i)$ the heteroskedasticity robust KICM statistic KICM_h also follows a chi square distribution with l degrees of freedom

Proposition 4.2
Given Ω, for any $P \in \mathcal{P}$ such that $H_0 : \beta = \beta_0$, assumptions A(i), B(ii), B(iii), B(iv), C(i) and C(ii) hold and assuming that $Y_i|z_i \sim \mathcal{N}(\text{E}(Y_i|z_i), \Omega(z_i))$ holds then $\text{KICM}_h \sim \chi^2_l$

The proof is very similar to that of proposition 4.1 and is omitted. Once again the key argument is that (S_{ih}, T_{ih}) has variance identity conditionally on z_i. The feasible statistic KICM_{hf} can also be proven to be χ^2_l asymptotically taking $\Omega(\cdot)$ as given. The formal proof of this result is also omitted.

4.3 Non-normal Errors

With non-normal errors I use empirical process theory to prove that the heteroskedasticity robust feasible KICM test is uniformly valid and uniformly consistent. In the heteroskedastic case $\Omega(z_i)$ is random and could be unbounded thus it is very desirable that KICM is valid and consistent for all possible data generating processes and not specific ones. Note that the following results also hold under homoskedastic data and / or normal data.

Theorem 4.3 shows that the feasible heteroskedasticity robust KICM test is uniformly valid regardless of instruments strength. Indeed no statement is made about $(a_j)_{j=1}^l$ thus by inverting this test one automatically obtains a confidence interval with at least nominal coverage. Of course in practice this interval may be large, especially if instruments are weak.

Theorem 4.3 (Uniform Validity of KICM)
Denote by $q_{1-\alpha}$ the $1-\alpha$ quantile of the chi-square distribution with l degrees of freedom.

Then, under the null $H_0 : \beta = \beta_0$ and assumptions A(ii), A(iii), B(i), B(iii), B(iv), C(i), C(ii), D(ii) and D(iii):

$$\lim_{n\to\infty} \sup_{\beta_0} \sup_{P\in\mathcal{P}:\beta=\beta_0} \mathrm{P}(\mathrm{KICM}_{hf} > q_{1-\alpha}) \leqslant \alpha$$

The size of KICM *is lesser or equal than its nominal size asymptotically.*

The proof of theorem 4.3 is in B of the appendix.

Regarding the power of the KICM test it is determined by the identification strength of the elements of β which differs from β_0, in addition it is the element which is best identified which will drive power. Intuitively, to reject H_0 the part of β which is different from β_0 must be at least semi-strongly identified. Mmore explicitly if for element j and j' $\beta_j = \beta_{j0}$ and $\beta_{j'} \neq \beta_{j'0}$, then element j can never contribute to rejecting the null whereas j' will contribute if $a_{j'} \leqslant 1/2$. Thus if B(ii) is assumed the test is consistent if $a < 1/2$ but if B(iii) is assumed then the test is consistent if $min\{a_j : \beta_{j0} \neq \beta_j\} < 1/2$. For simplicity corollary 4.4 presents the asymptotic uniform power properties of the feasible heteroskedasticity robust KICM test under assumption B(ii), ie $\Pi = n^{-a}C(\cdot)$. The power properties of the KICM test under assumption B(iii), ie $\Pi(\cdot) = N_C^{-1}C(\cdot)$, are presented and discussed in appendix D.

Corollary 4.4 (Uniform consistency of KICM)

Denote by $q_{1-\alpha}$ the $1-\alpha$ quantile of the chi-square distribution with l degrees of freedom. Then under assumptions A(ii), A(iii), B(i), B(ii), C(i), C(ii), D(ii), and D(iii),

- $\lim_{n\to\infty} \inf_{\beta_0} \inf_{P\in P:\beta\neq\beta_0, a<1/2} \mathrm{P}(\mathrm{KICM}_{hf} > q_{1-\alpha}) = 1$;
  The test is consistent when the instruments are at least semi-strong.

- $\lim_{n\to\infty} \inf_{\beta_0} \inf_{P\in\mathcal{P}:\beta\neq\beta_0, a=1/2} \mathrm{P}(\mathrm{KICM}_{hf} > q_{1-\alpha}) \in [\alpha; 1)$;
  The test has more than trivial power when the instruments are weak.

- $\lim_{n\to\infty} \sup_{\beta_0} \sup_{P\in\mathcal{P}:\beta\neq\beta_0, a>1/2} \mathrm{P}(\mathrm{KICM}_{hf} > q_{1-\alpha}) \leqslant \alpha$;
  The test has trivial power when the instruments are very weak.

The proof of corollary 4.4 is in C of the appendix. Next, I study the empirical performances of KICM.

5 Simulations

Setting I perform simulations in order to evaluate the empirical performances (size, power, confidence interval length) of the KICM test in small samples in case of strong, semi-strong or weak instruments, for 4 different first stages, and in case of homoskedastic or heteroskedastic data. The specification in the simulations is the following: I assume that there is one regressor $l = 1$ and either one or two instruments $k \in \{1, 2\}$ in (2.2) and (2.3), the true β is 0, the instrument z_i are standard normal thus centered, uncorrelated, and with a symmetric distribution, sample size is either 100 or 400. Thus there are 24 possible setups:

- 4 possible first stages: linear; non-linear; polar polynomial; semi-polar polynomial

$$\Pi_1(z) = \frac{z_1}{n^a}, \qquad \Pi_2(z) = \frac{1}{n^a}\frac{z_1 + z_2 + z_1 z_2 + z_1^2 + z_2^2 + z_1^2 z_2^2 - 3}{\sqrt{26}},$$
$$\Pi_3(z) = \frac{1}{n^a}\frac{z_1^2 - 1}{\sqrt{3}}, \qquad \Pi_4(z) = \frac{1}{n^a}\frac{z_1 + z_2^2 - 1}{\sqrt{4}}$$

- 3 instrument strengths: strong $a = 0$; semi-strong $a = 1/4$; weak $a = 1/2$
- 2 data types: homoskedastic; heteroskedastic

$$\Omega = \begin{pmatrix} 1 & 0.81 \\ 0.81 & 1 \end{pmatrix}, \qquad \Omega(z) = \frac{1 + z_1^2}{2}\begin{pmatrix} 1 & 0.81 \\ 0.81 & 1 \end{pmatrix}$$

Keeping instrument strength constant the data $(y_i, x_i, z_i)_{i=1}^n$ has the same mean and variance whatever the setup. Note that for this reason, controlling instruments strength a is equivalent to controlling the value of the (nonlinear) concentration parameter $\mu^2 = \frac{\Pi(z)'\Pi(z)}{\sqrt{\text{Var}(v_i)}}$ as in papers which define instruments' strength by the value of the concentration parameter, see Stock and Yogo (2005), Staiger and Stock' (1997). In these simulations if $a = 1/2$ then $\mu^2 \approx 1$, if $a = 1/4$ then $\mu^2 \approx \sqrt{n}$, if $a = 0$ then $\mu^2 \approx n$.

Competing methods I consider 6 competing procedures for building confidence sets: the AR, the LM, the CLR, the ICM, the CICM and the Wald, the later is simply the confidence set built using the 2SLS estimator and using its traditional confidence interval based on the t-test Gaussian asymptotics. Note that if the first stage is polar

polynomial as specified above then the best linear projection of y_i and x_i on z_i is 0, and that if the first stage is semi-polar polynomial then the best linear projection of y_i and x_i on z_{2i} is 0. So even in case of strong instruments I expect the AR, LM and CLR tests to perform very badly in terms of power compared to the KICM (and ICM and CICM). The number of simulations required to build the CI of the CLR, ICM and CICM vary between $m = 200$ and $m = 500$. Finally to create comparable heteroskedasticity robust versions of the AR, LM, CLR, ICM, KICM and CICM I use the nonparametric estimator of $\Omega(\cdot)$ of Yin et al. (2010), and consider the Eicker-White estimator of the covariance matrix of the 2SLS estimator for the Wald test.

Empirical size First, the coverage of the confidence intervals built with KICM is of special interest, it is the probability that the true β, which equals 0 in this setting, is in said interval (which is random). The empirical coverage is equal to the empirical size when tests are inverted. Hence, when comparing sizes the best procedure is the one for which the empirical size is closest to nominal size which I set to 10%. I report the empirical sizes of the AR, LM, CLR, ICM KICM, CICM and Wald test for the different settups in table 1 in the weak instruments case, in table 2 in the semi-strong instruments case, and in table 3 in the strong instruments case constructed over 5000 simulations in appendix E.1.1.

From the tables, the AR, LM, CLR, ICM, KICM and CICM are all robust to weak instruments thus in terms of size there is little difference between them when the strength of instruments changes. This is not the case for the Wald test built from 2SLS, it is not robust to weak instruments nor non-linearities, in fact even with semi-strong instruments it is very oversized. More precisely in all settings it seems that the empirical size of KICM is closer to nominal size 10% than both ICM and CICM which require a larger number of simulations in order to be competitive, especially in the linear case $\Pi_1(\cdot)$. In addition, KICM has better size than the AR, LM and CLR in all settings expect in the linear one which is within expectations.

Power curves Second, to assess how KICM rejects wrong values of β_0 I plot its power curve and the power curves of other competing tests as is done in most of the literature on testing. A power curve is drawn by measuring the empirical probability of rejecting the null $H_0 : \beta = \beta_0$ for many different β_0 in a grid. This implies that at the grid point $\beta_0 = \beta$ it is the empirical size of the test that is computed. Power curves

are a useful tool as they tell if one test will reject false values of β more than another, consequently if a test's power curve dominates another's then its confidence intervals will be systematically tighter than its competitor's.

In appendix E.1.2 below are the power curves of the AR, LM, CLR, ICM, KICM, CICM and Wald test built from 5000 replications for a test of nominal size 10% for the linear, non-linear and polar polynomial first stages. Sample size is $n = 400$ and I used $m = 500$ simulations in order to compute the critical values of the CLR, ICM and CICM tests. Figure 1, figure 2, figure 3 are the power curves of the 7 tests for strong, semi-strong and weak instruments respectively with homoskedastic data. Figure 4, figure 5, figure 6 are the power curves for strong, semi-strong and weak instruments respectively with heteroskedastic data.

Several remarks should be made: Notice that the curves are almost similar whether data is homoskedastic or heteroskedastic. Next in the polar case and non linear case note that the tests which are non-robust to non-linearities (AR, LM CLR and Wald) experience a large loss of power even when instruments are strong. All tests have trivial power when instruments are weak which again is as expected. Additionally KICM has power overall similar to ICM and CICM except on one side, this is due to the fact that inverting the KICM test is equivalent to solving a quartic inequality, LM has similar properties, see Mikusheva (2010). Thus when choosing between KICM, ICM and CICM there may be a tradeoff between coverage and power. KICM always seem to have better coverage but CICM seem to have better power in some cases.

Average p-value curves To have an idea of the "average length" of the CI interval for each test procedure, I define an "average" confidence interval built by inverting a test over many simulations. I could consider taking the average bounds of the confidences intervals built over many simulations, however bounds may not exists when instruments are weak, so instead I find the average 90% coverage confidence interval by using average p-value curves: For any candidate β_0 for any test for any setting, I check if the average p-value when testing $H_0 : \beta = \beta_0$ is above 10%. Then all the β_0 for which it is true will constitute the average 90% confidence interval.

Figure 7, 8 and 9 in appendix E.1.3 are plots of the average p-value curves for strong, semi-strong and weak instruments respectively, with heteroskedastic data, $n = 400$, $m = 500$.

The sets are the whole real line in case of weak instruments for all tests except the

Wald test which gives a finite interval which has very low coverage. The tests which are non robust to non-linearities (AR, LM and CLR) have higher average p-values hence the average confidence intervals built from them are much large than the sets built from ICM, KICM and CICM in case of non-linear first stage, and infinite in case of polar polynomial first stage unlike sets built from ICM, KICM and CICM. There is little to differentiate the sets built with KICM from the ones built from ICM and CICM in the strong or weak instruments case. In case of semi-strong instruments however the set built from KICM is the union of 2 finites sets in the linear and non-linear first stage case, one big set which is common the ICM and CICM and one much smaller set. Again there seems to be a tradeoff between having the right coverage and higher power when choosing between KICM and ICM and CICM.

Empirical size with a higher number of instruments Finally I consider a first stage with a higher number of instruments. It is well known that the LM test fares much better when k starts to grpw compared to both the AR and CLR in the linear case. The same holds true for KICM compared to ICM and CICM. I consider the linear first stage

$$\Pi_5(z) = \frac{1}{n^a}\frac{z_1 + z_2 + z_3 + z_4}{\sqrt{4}}$$

Then the empirical coverage of the 7 tests over 5000 simulations for homoskedastic data, sample size $n = 100$ and $m = 200$ simulations of the distributions of CLR, ICM and CICM are in table 4.

Clearly all the tests are oversized except the LM and KICM, hence a confidence interval built from will have a lower coverage than the nominal 90% for weak, semi-strong or strong instruments.

6 Application: Returns to schooling

In this final section I provide inference via KICM for the causal effect of the number of years of schooling on the logarithm of wage using quarters of birth as instruments using the data from Angrist and Krueger (1991). The authors estimate the causal effect of the number of years spent in school on wage by using the exogenous variation of schooling due to difference in quarters of births: Teenagers born early in the year leave school earlier because they reach the age at which they can work earlier, this creates

a difference in the total number of years of schooling between children born early in the year and children born later. In addition the authors try different specifications by interacting these instruments with time and location dummies in order to increase the fit of the first stage with the belief that it increases the strength of the instruments. Data is from the US where they have access to different cohorts and I focus on cohort 20-29 with $247,199$ observations in the sample.

This paper is well-known for the fact that the instruments used are quite weak- see Stock and Yogo (2005) for a comprehensive look at thresholds which determine if instruments are weak, across all specifications and all cohorts the F-statistic vary between 1 and 15. Because of the large sample size ICM and CICM cannot be used. At the same time considering that the first stage is non-linear allows quarter of births to act as types, the first stage is equal to a different non-linear function of the exogenous regressors for each type. This degree of flexibility is enormous compared to a linear first stage with only a few interactions being considered. Thus one can expect the confidence intervals built from KICM to be small compared to the competition if there is a sufficient number of covariates.

Formally I consider the model in table IV of Angrist and Krueger (1991) which focuses on cohort 20-29 which writes

$$\begin{aligned} log(wage)_i &= \beta\ schooling_i + FE_y + FE_r + x_i'\gamma + u_i \\ schooling_i &= \sum_{j,t} \alpha_j\ 1_{QB_i=j,YB_i=t} + FE_y + FE_r + x_i'\zeta + v_i \end{aligned}$$

where FE_y are year of birth fixed effects, FE_r are region of residence fixed effects, x_i some covariates. Using KICM I allow the first stage to be completely non-linear in the instruments and covariates thus consider

$$schooling_i = \sum_{j,t,r} 1_{QB_i=j,YB_i=t,RR_i=r}\Pi_{jtr}(x_i) + v_i$$

In table 5 of appendix E.2 below I provide estimates of the 90% coverage heteroskedasticity robust confidence interval of β by inverting AR, LM, CLR, and KICM and for reference I also provide the OLS, 2SLS, LIML and Fuller estimates and their t-test Gaussian based confidence intervals for 4 different specifications.

In the simplest specification (1) there are no covariates only year fixed effects therefore KICM considers the same first stage as the other tests. At the same time without

covariates, schooling is very likely to still be endogenous even after being projected on the instruments. Assuming exogeneity however, observe that even in the simplest specification the KICM confidence set for returns to schooling is positive and has smaller length than the set built with the AR test, the CLR set however is significantly smaller. For specification (2) and (3) which add other covariates all 4 test procedures which are robust to weak instruments give the whole real line as the confidence interval. Finally in specification (4) which also includes region of residence fixed effects, while other tests give the whole real line as confidence sets, the KICM confidence interval is small and positive. In figure 10 of appendix E.2 is the p-value curve for KICM and Wald tests built using the different estimators over a grid of potential null $H_0 : \beta = \beta_0$ in specification (4). This result is not so surprising as KICM becomes more powerful with more covariates as it considers all non-linearities including interactions. This set is also very different from the OLS and 2SLS set but is included in the LIML and Fuller set which is not far-fetched because LIML and Fuller are known to perform better than 2SLS when there are weak possibly many instruments.

These results imply that estimates of the returns to schooling may not be as small as OLS and 2SLS and not as large as Fuller and LIML have indicated until now, from specification (4) an increase of 1 year of schooling yields an increase in wages between 13.8% and 24%.

7 Conclusion

On the one hand in the current literature on weak instruments, most inference procedures do not take into account non-linearities or interactions in the first stage. This leads to an important loss of relevance, or a total loss of relevance of the instruments, both in simulations and in applications. On the other hand estimating the first stage non-linearly is difficult and leads to a situation of having too many weak instruments, see Dieterle and Snell (2016). Thus like ICM and CICM from Antoine and Lavergne (2019), KICM relies on an integrated conditional moment in order to consider the non-linearities present in the first stage.

KICM has some advantages over the ICM and CICM. First, its size is closer to nominal size in practice hence sets built with KICM have coverages which are closer to nominal coverage compared to ones built with ICM or CICM. Second, it is more robust to many instruments than ICM and CICM just like how the LM is more robust

to many instruments than the AR and CLR. Third, it is pivotal hence it does not require simulations in order to obtain its asymptotic distribution, ICM and CICM require simulations and are thus unimplementable when samples get large as in the application of this paper. This implies that in larger samples or with more instruments KICM is very simple to implement and reliable. All in all KICM is an off-the shelf procedure to easily compute confidence sets which are robust to weak instruments and which consider non-linearities in the first stage, regardless of sample size or the number of instruments, normality or non-normality of the data, homoskedasticity or heteroskedasticity of the data.

Consequently linear IV models with a single instrument and few covariates and / or fixed effects should expect significant improvements in the quality of confidence sets at no cost when using KICM compared to alternative procedures as it will automatically consider both interactions and non-linearities in the first stage while maintaining pivotality of the test. The pivotality of KICM is very useful in applied micro settings with large samples, but this property is likely to be lost when data is clustered or auto-correlated without imposing a lot of structure, this requires more investigation. There may also exists other tests for weak-identification robust inference which consider a non-linear first stage with possibly better power properties than ICM, CICM or KICM, and which are also pivotal. As in the work of Moreira and Ridder (2017) or Andrews and Mikusheva (2016a), it may be possible to prove that ICM, CICM and KICM are or are not optimal.

Bibliography

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

Andrews, D. W. (1995): "Nonparametric Kernel Estimation for Semiparametric Models," *Econometric Theory*, 11, 560–586.

Andrews, D. W., M. Moreira, and J. Stock (2004): "Optimal Invariant Similar Tests for Instrumental Variables Regression," Tech. rep., National Bureau of Economic Research.

Andrews, D. W. and J. H. Stock (2007): "Testing with many weak instruments," *Journal of Econometrics*, 138, 24–46.

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 J. Stock (2005): "Inference with Weak Instruments," *NBER Technical Working Paper*.

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.

Angrist, J. D. and A. B. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earning?" *Quaterly Journal of Economics*, 106 (4), 979–1014.

Antoine, B. and P. Lavergne (2019): "Identification-Robust Nonparametric Inference in a Linear IV Model," *Working Paper*.

Bierens, H. J. (1982): "Consistent model specification tests," *Journal of Econometrics*, 20, 105–134.

Bound, J., D. A. Jaeger, and R. M. Baker (1995): "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak," *Journal of the American Statistical Association*, 90, 443–450.

Chernozhukov, V. and C. Hansen (2008): "The reduced form: A simple approach to inference with weak instruments," *Economics Letters*, 100, 68–71.

Darolles, S., J.-P. Florens, and E. Renault (2011): "Nonparametric Instrumental Regression," *Econometrica*, 79, 1541–1565.

Dieterle, S. G. and A. Snell (2016): "A simple diagnostic to investigate instrument validity and heterogeneous effects when using a single instrument," *Labour Economics*, 42, 76–86.

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/Revue Canadienne d'Economique*, 36, 767–808.

Dufour, J.-M. and M. Taamouti (2007): "Further results on projection-based inference in IV regressions with weak, collinear or missing instruments," *Journal of Econometrics*, 139, 133–153.

Guggenberger, P., F. Kleibergen, and S. Mavroeidis (2019): "A more powerful subvector Anderson Rubin test in linear instrumental variables regression," *Quantitative Economics*, 10, 487–526.

Guggenberger, P., F. Kleibergen, S. Mavroeidis, and L. Chen (2012): "On the Asymptotic Sizes of Subset Anderson-Rubin and Lagrange Multiplier Tests in Linear Instrumental Variables Regression," *Econometrica*, 80, 2649–2666.

Guggenberger, P. and R. J. Smith (2005): "Generalized Empirical Likelihood Estimators and Tests Under Partial, Weak, and Strong Identification," *Econometric Theory*, 21.

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

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

Johnson, N. L., S. Kotz, and N. Balakrishnan (1995): *Continuous univariate distributions 2 2*, New York [u.a.: Wiley, oCLC: 1068188012.

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, New York, NY: Springer New York.

Li, Q. and J. S. Racine (2006): *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

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:1705.00231 [math, stat]*, arXiv: 1705.00231.

Newey, W. K. and J. L. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578.

Seifert, B., T. Gasser, and A. Wolf (1993): "Nonparametric Estimation of Residual Variance Revisited," *Biometrika*, 80 (2), 373–383.

Staiger, D. and J. H. Stock' (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65 (3), 557–586.

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 & Economic Statistics*, 20, 518–529.

Stock, J. H. and M. Yogo (2005): "Testing for Weak Instruments in Linear IV Regression," *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg.*

Vaart, A. W. v. d. (2007): *Asymptotic statistics*, Cambridge series in statistical and probabilistic mathematics, Cambridge: Cambridge Univ. Press, oCLC: 838749444.

Vaart, A. W. v. d. 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.

A Proof of Propositions

A.1 Proof of Proposition 3.1

The implication from H_0 to H_0^1 and from H_0^1 to H_0^2 are obvious so I prove the reverse implications. From H_0^1 to H_0 notice that

$$\mathrm{E}(y_i - x_i\beta_0|z_i) = \Pi(z_i)'(\beta - \beta_0)$$

Thus if $l = 1$ and $\mathrm{P}(\Pi(z_i) = 0) = 0$ then $\mathrm{E}(y_i - x_i\beta_0|z_i)$ implies $\beta = \beta_0$.

From H_0^2 to H_0^1 first let $\mathrm{E}(y_i - x_i'\beta_0|z_i) = r(z_i) = max\{r(z_i); 0\} - max\{-r(z_i); 0\} = r_1(z_i) - r_2(z_i)$. Then

$$\mathrm{E}((y_i - x_i'\beta_0)e^{it'z_i}) = \mathrm{E}(r_1(z_i)e^{it'z_i}) - \mathrm{E}(r_2(z_i)e^{it'z_i}) = \int r_1(s)dP_z(s) - \int r_2(s)dP_z(s)$$

Next for $j = 1, 2$ define the probability measures $P_j(B) = \int_B r_j(s)dP_z(s)/\mathrm{E}(r_j(z))\ \forall B \in \mathcal{B}(\mathbb{R}^k)$. This gives

$$\begin{aligned}
\forall t,\ \mathrm{E}((y_i - x_i'\beta_0)e^{it'z}) = \mathrm{E}(r_1(z_i)) \int e^{it's}dP_1(s) - \mathrm{E}(r_2(z_i)) \int e^{it's}dP_2(s) = 0 \\
&\Rightarrow \mathrm{E}(r_1(z_i)) = \mathrm{E}(r_2(z_i)) \text{ by taking } t = 0 \\
&\Leftrightarrow \int e^{it's}dP_1(s) - \int e^{it's}dP_2(s) = 0\ \forall t \\
&\Leftrightarrow \forall B \in \mathcal{B}(\mathbb{R}^k),\ P_1(B) - P_2(B) = 0 \\
&\Leftrightarrow \forall B \in \mathcal{B}(\mathbb{R}^k),\ \int_B r(s)P_z(s) = 0
\end{aligned}$$

Taking the event $B^+ = \{s : r(s) > 0\}$ and $B^- = \{s : r(s) < 0\}$ gives $\int_{B^+} r(s)dP_z(s) = \mathrm{E}(r(z_i)1_{r(z_i)>0}) = 0 \Leftrightarrow \mathrm{P}(r(z_i) > 0) = 0$ and $\int_{B^-} r(s)dP_z(s) = 0 \Leftrightarrow \mathrm{P}(r(z_i) < 0) = 0$, which imply that $\mathrm{P}(r(z_i) = 0) = 1 \Leftrightarrow r(z_i) = \mathrm{E}(y_i - x_i\beta_0|z_i) = 0$ a.s. hence H_0^2 implies H_0^1.

Equivalences between H_0^2, H_0^3 and H_0^4 are easily established using properties of positive functions and integrals.

A.2 Proof of Proposition 4.1

To prove that KICM is χ_l^2 conditionally distributed first use the eigendecomposition of P_{WT}

$$P_{WT} = H \begin{pmatrix} I_l & 0_{l\times(n-l)} \\ 0_{(n-l)\times l} & 0_{(n-l)\times(n-l)} \end{pmatrix} H'$$

where $H = (\ WT(T'W^2T)^{-1/2} \quad M_{WT}A(A'M_{WT}A)^{-1/2}\ )$, and $A = \begin{pmatrix} I_{n-l} \\ 0_{l\times(n-l)} \end{pmatrix}$, H is the orthogonal eigenvector matrix of P_{WT}. This allows to rewrite KICM as the sum of l components

$$\text{KICM} = \sum_{i=1}^{l} (S'H)_i^2$$

Second I prove that $S'H \sim \mathcal{N}(0, I_n)$ conditionally on z. Under the null $H_0 : \beta = \beta_0$, $\mathrm{E}(S_i|z_i) = 0$ and $\mathrm{Var}(S_i|z_i) = 1$ so that

$$\mathrm{Cov}(S_i, T_i|z_i) = \mathrm{E}(S_iT_i|z_i) = \frac{b_0' \,\mathrm{E}(Y_iY_i')\Omega^{-1}A_0(A_0'\Omega^{-1}A_0)^{-1/2}}{\sqrt{b_0'\Omega b_0}} = 0_l$$

Thus conditionally on z_i $(S_i, T_i) \sim \mathcal{N}((0, \mathrm{E}(T_i)), I_{l+1})$. Consequently under the null and conditionally on z_i $S_i \perp\!\!\!\perp T_i$ which implies that $S \perp\!\!\!\perp T$. In turn P is a function of z through W and of T therefore

$$\begin{aligned}
&\mathrm{E}(S'H|z, T) = \mathrm{E}(S|z)'H = 0_n \\
\Rightarrow &\mathrm{Cov}((S'H)_i, (S'H)_j|z, T) = \mathrm{E}(e_i'HSS'He_j|z) = e_iH'\,\mathrm{E}(SS'|z)He_j = 1_{i=j} \\
\Rightarrow &\mathrm{Var}(S'H|z, T) = H'\,\mathrm{E}(SS'|z)H = H'H = I_n \\
\Rightarrow &S'H \sim \mathcal{N}(0, I_n)|z, T \\
\Rightarrow &S'H \sim \mathcal{N}(0, I_n)
\end{aligned}$$

where e_j denotes a vector of size n equal to zero in all elements expect in coordinate j where it is equal to 1 and $1_{1=j}$ is an indicator function which equals one only when $i = j$. Consequently KICM is equal to a sum of l independent squared standard normal so KICM follows a χ^2_l under $H_0 : \beta = \beta_0$.

B Proof of theorem 4.3

Throughout the proof assumptions A(ii), A(iii), B(i), B(iii), B(iv), C(i), C(ii), D(ii), and D(iii) are maintained. The proof is divided in five parts, the first introduces notations and some useful matrix results, the second presents the decomposition of the KICM statistic used in the proof, the third proves the convergence of the random processes which compose KICM, the fourth characterizes the limits of these random processes, and the fifth proves validity by contradiction.

B.1 Notations and matrix results

Denote by $o_{\mathrm{P}}(1)$ and $O_{\mathrm{P}}(1)$ the small o in probability and big O in probability notations for degenerate in probability and bounded in probability: If $X = o_{\mathrm{P}}(1)$ then $\forall \varepsilon >$

0 $\mathrm{P}(|X| > \varepsilon) \to 0$. If $X = O_{\mathrm{P}}(1)$ then $\forall \varepsilon > 0\ \exists M > 0 : \mathrm{P}(|X| > M) \leqslant \varepsilon$. Denote by $o_{\underline{\mathrm{P}}}(1)$ and $O_{\bar{\mathrm{P}}}(1)$ the uniform counterparts of $o_{\mathrm{P}}(1)$ and $O_{\mathrm{P}}(1)$: If $X = O_{\bar{\mathrm{P}}}(1)$ then X is uniformly bounded in probability and $\forall \epsilon > 0\ \exists M : \underset{P \in \mathcal{P}}{sup}\ \mathrm{P}(|X| > M) \leqslant \epsilon$. If $X = o_{\underline{\mathrm{P}}}(1)$ then X is uniformly degenerate in probability and $\forall \varepsilon > 0,\ \underset{P \in \mathcal{P}}{sup}\ \mathrm{P}(|X| > \varepsilon) \to 0$. For any random object X then denote by $X \neq 0$ the condition $\mathrm{P}(X = 0) < 1$ or equivalently $\mathrm{P}(X \neq 0) > 0$. In addition for some square matrix X denote by $\lambda_{min}(X)$ and $\lambda_{max}(X)$ its smallest and biggest eigenvalue respectively.

Before starting the proof here are four useful matrix results which are used many times over: For any invertible matrices X and Y, for any full rank matrix A

$$X^{-1} - Y^{-1} = X^{-1}(Y - X)Y^{-1} \tag{B.6}$$

$$X - Y = \frac{1}{2}(X^{1/2} + Y^{1/2})(X^{1/2} - Y^{1/2}) + \frac{1}{2}(X^{1/2} - Y^{1/2})(X^{1/2} + Y^{1/2}) \tag{B.7}$$

$$(A'XA)^{-1} \leqslant (A'A)^{-1}\ \lambda_{min}(X)^{-1} \tag{B.8}$$

$$((A'XA)^{-1/2} + (A'YA)^{-1/2})^{-1} \leqslant (A'A)^{1/2}(\lambda_{max}(X)^{-1/2} + \lambda_{max}(Y)^{-1/2})^{-1}\} \tag{B.9}$$

$$((A'X^{-1}A)^{-1/2} + (A'Y^{-1}A)^{-1/2}) \geqslant (A'A)^{-1/2}(\lambda_{min}(X)^{1/2} + \lambda_{min}(Y)^{1/2})\} \tag{B.10}$$

B.2 Preliminary decomposition

As defined in section 3.3

$$T_{ih} = (A_0'\Omega_i^{-1}A_0)^{-1/2}A_0'\Omega_i^{-1}Y_i$$

Thus each component j of T_{ih} is a weighted sum of the components of Y_i. Thus without loss of generality

$$T_{ih,j} = \sum_{l'=1}^{l} w_{ij,l'}x_{il'} + w_{ij,y}y_i = \sum_{l'=1}^{l}(w_{ij,l'} + w_{ij,y}\beta_{0l'})x_{il'} + w_{ij,y}u_i \equiv \sum_{l'=1}^{l} w^*_{ij,l'}x_{il'} + w_{ij,y}u_i$$

where $T_{ih,j}$ is the j-th coordinate of T_{ih}, $w_{ij,l'} \in \mathbb{R}$ is the weight associated to $x_{il'}$ through x_i, $w_{ij,y} \in \mathbb{R}$ the weight associated to y_i, and ($w^*_{ij,l'} \in \mathbb{R}$ is the true weight associated to $x_{il'}$. These weights are functions of z_i and of the null β_0 and note that $\forall j\ T_{ih,j} \neq 0$ because $\mathrm{Var}(T_i) = I_n$, thus $\forall j\ (w^*_{j1i}, \ldots, w^*_{jli}, w_{ij,y}) \neq 0_{l+1}$. Therefore define $(a^*_j)_{j=1}^{l}$ where

$$a^*_j = \begin{cases} a & \text{if} \quad (w^*_{ij,l'})_{l'=1}^{l} \neq 0_l \\ 1 & \text{o.w} \end{cases}$$

a^*_j does not represent identification strength of β_j but instead characterize the limit (which exists) of $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} T_{ih,j}$. Additionally a^*_j does not only depend on instruments'

strength a, but also on the conditional covariance between x_{ij} and $(x_{ij'})_{j'=1}^{l}$, and on β_0. To get an heuristic idea of why a_j^* is introduced, notice that because $\Pi(\cdot) = n^{-a}C(\cdot)$ if $a_j^* > 1/2$ then $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} T_{ih,j}$ will converge towards a centered Normal distribution, if $a_j^* = 1/2$ then $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} T_{ih,j}$ will converge towards a non-centered distribution, and if $a_j^* < 1/2$ then $\frac{n^{a_j^*-1/2}}{\sqrt{n}}\sum_{i=1}^{n} T_{ih,j}$ will converge in probability towards a certain expectation. Hence KICM rewrites as follows

$$\text{KICM}_h = S_h'WT_h(T_h'W^2T_h)^{-1}T_h'WS_h = S_h'WT_hN_C^*(N_C^*T_h'W^2T_hN_C^*)^{-1}N_C^*T_h'WS_h$$

$$\text{where } N_C^* = \begin{pmatrix} n^{a_1^*-1/2}1_{a_1^*<1/2} + 1_{a_1^*\geqslant 1/2} & 0 & \cdots & 0 \\ 0 & n^{a_2^*-1/2}1_{a_2^*<1/2} + 1_{a_2^*\geqslant 1/2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & n^{a_l^*-1/2}1_{a_l^*<1/2} + 1_{a_l^*\geqslant 1/2} \end{pmatrix}$$

N_C^* is a $l \times l$ diagonal matrix, it is a theoretical tool which will allow to properly characterize the limit of KICM, given any instruments' strength, any covariance structure $\Omega(\cdot)$ and any null β_0.

I can then rewrite the components of KICM_h. $S_h'WT_hN_C^*$ is the integral of the product of the sums of $(S_{ih})_{i=1}^n$ and of $(T_{ih})_{i=1}^n$

$$\begin{aligned} S_h'WT_hN_C^* &= \frac{1}{n}\sum_{i,j} S_{ih}T_{jh}'N_C^*w(z_i - z_j) = \int \frac{1}{n}\sum_{i,j} S_{ih}T_{jh}'N_C^*e^{it'(z_i-z_j)}\mu(t) \\ &= \int \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_{ih}e^{it'z_i}\right)\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n} T_{jh}'N_C^*e^{-it'z_j}\right)\mu(t) \end{aligned}$$

$N_C^*T_h'W^2T_hN_C^*$ also rewrites. I reformulate the elements of the matrix W^2 first

$$\begin{aligned} (W^2)_{ij} &= \frac{1}{n^2}\sum_{m=1}^{n}\int_{\mathbb{R}^k} e^{it'(z_i-z_m)}d\mu(t)\int_{\mathbb{R}^k} e^{is'(z_m-z_j)}d\mu(s) \\ &= \frac{1}{n^2}\int\int e^{it'z_i}e^{-is'z_j}\left(\sum_{m=1}^{n} e^{i(s-t)z_m}\right)d\mu(t)d\mu(s) \end{aligned}$$

$$\Rightarrow N_C^*T_h'W^2T_hN_C^* = \sum_{i,j} N_C^*T_{ih}T_{jh}'N_C^*(W^2)_{ij}$$

$$N_C^*T_h'W^2T_hN_C^* = \int\int \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} N_C^*T_{ih}e^{it'z_i}\right)\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n} T_{jh}'N_C^*e^{-is'z_j}\right)\left(\frac{1}{n}\sum_{m=1}^{n} e^{i(s-t)z_m}\right)d\mu(t)d\mu(s)$$

Thereforedefine the processes which enter the integrals and characterize KICM

$$G_S(t,\Omega) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(b_0'\Omega(z_i)b_0)^{-1/2}b_0'Y_i e^{it'z_i} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_{ih}e^{it'z_i}$$
$$G_T(t,\Omega) = \frac{N_C^*}{\sqrt{n}}\sum_{i=1}^{n}(A_0'\Omega^{-1}(z_i)A_0)^{-1/2}A_0'\Omega^{-1}(z_i)Y_i e^{-it'z_i} = \frac{N_C^*}{\sqrt{n}}\sum_{i=1}^{n} T_{ih}e^{-it'z_i}$$
$$G_z(t) = \frac{1}{n}\sum_{i=1}^{n} e^{it'z_i}$$

So KICM_h rewrites as

$$\int_{\mathbb{R}^k} G_S(t,\Omega)G_T'(t,\Omega)d\mu(t)\left(\int_{\mathbb{R}^k} G_T(-t,\Omega)G_T'(s,\Omega)G_z(s-t)d\mu(t)\right)^{-1}\int_{\mathbb{R}^k} G_S(t,\Omega)G_T(t,\Omega)d\mu(t)$$

Lastly, define the conditionally normal version of G_S and G_T

$$G_S^*(t,\Omega) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(b_0'\Omega(z_i)b_0)^{-1/2}b_0'Y_i^* e^{it'z_i}, \quad Y_i^*|z_i \overset{iid}{\sim} \mathcal{N}(\mathrm{E}(Y_i|z_i),\Omega(z_i))$$
$$G_T^*(t,\Omega) = \frac{N_C^*}{\sqrt{n}}\sum_{i=1}^{n}(A_0'\Omega^{-1}(z_i)A_0)^{-1/2}A_0'\Omega^{-1}(z_i)Y_i^* e^{-it'z_i}$$

B.3 Donsker property and asymptotic equicontinuity

From the above decomposition in order for KICM to converge, the processes G_S, G_T, G_S^*, G_T^* must uniformly converge to Gaussian processes, ie to be Donsker, and G_z must uniformly converge to a function, ie to be Glivenko-Cantelli. In addition in order to make the difference between the feasible KICM test statistic and the KICM test statistic the processes G_S, G_S^*, G_T and G_T^* also need to be asymptotically equicontinuous in Ω.

To obtain these results the following families of functions

$$\mathcal{F}_S = \{(c,d) \in \mathbb{R}^k \times \mathbb{R}^{l+1} \mapsto (b_0'\Omega b_0)^{-1/2}b_0' d\ e^{it'c}, t \in \mathbb{R}^k, \Omega \in \mathcal{O}\}$$
$$\mathcal{F}_T = \{(c,d) \in \mathbb{R}^k \times \mathbb{R}^{l+1} \mapsto (A_0'\Omega^{-1}A_0)^{-1/2}A_0'\Omega^{-1} d\ e^{it'c}, t \in \mathbb{R}^k, \Omega \in \mathcal{O}\}$$
$$\mathcal{F}_z = \{c \in \mathbb{R}^k \mapsto e^{it'c}, t \in \mathbb{R}^k\}$$

need an envelope F which satisfies $\mathrm{E}(F^2(y_i,x_i,z_i)1_{F(y_i,x_i,z_i)\geqslant M}) \underset{M\to+\infty}{\to} 0$ and a bounded uniform entropy integral (BUEI). If that's the case then by theorem 2.8.3 in Vaart and Wellner (2000) $\mathcal{F}_S$, $\mathcal{F}_T$ and $\mathcal{F}_z$ are uniformly Donsker and pre-Gaussian and thus by

theorem 2.8.2 from Vaart and Wellner (2000) processes G_S, G_T, G_S^* and G_T^* weakly converge to Gaussian processes and are uniformly asymptotically equicontinuous, and G_z weakly converge to a function. Asymptotic equicontinuity implies that for any $\Omega \in \mathcal{O}$ with consistent estimator $\hat{\Omega}$ as per assumption D the difference between the feasible $G_S(\cdot, \Omega)$ and $G_T(\cdot, \hat{\Omega})$ vanishes uniformly in probability

$$\forall \varepsilon > 0, \ \sup_{\beta_0} \sup_{P \in \mathcal{P}: \beta = \beta_0} \mathrm{P}(|G_S(\cdot, \Omega) - G_S(\cdot, \hat{\Omega})| > \varepsilon) \to 0$$

$$\forall j = 1, \cdots, l, \ \forall \varepsilon > 0, \ \sup_{\beta_0} \sup_{P \in \mathcal{P}: \beta = \beta_0} \mathrm{P}(|G_{Tj}(\cdot, \Omega) - G_{Tj}(\cdot, \hat{\Omega})| > \varepsilon) \to 0$$

Here assumption D(iii) and D(iv) was used, specifically the fact that $\hat{\Omega}$ converge uniformly towards Ω in the L_2 sense and that $\hat{\Omega}$ is uniformly in $\mathcal{O}$ at the limit.

To show that $\mathcal{F}_S$, $\mathcal{F}_T$ and $\mathcal{F}_z$ have a "bounded" envelope and are BUEI, I first consider

$$\mathcal{F}_{cos,S} = \{(c, d) \in \mathbb{R}^k \times \mathbb{R}^{l+1} \mapsto (b_0' \Omega b_0)^{-1/2} b_0' d \ cos(t'c), t \in \mathbb{R}^k, \Omega \in \mathcal{O}\}$$
$$\mathcal{F}_{cos,T} = \{(c, d) \in \mathbb{R}^k \times \mathbb{R}^{l+1} \mapsto (A_0' \Omega^{-1} A_0)^{-1/2} A_0' \Omega^{-1} d \ cos(t'c), t \in \mathbb{R}^k, \Omega \in \mathcal{O}\}$$

Note that $|cos(.)| \leqslant 1$, and that from assumption D(ii) $\exists (\underline{\lambda}, \bar{\lambda}) : \underline{\lambda} I_{l+1} \leqslant \Omega(\cdot) \leqslant \bar{\lambda} I_{l+1}$ where $\underline{\lambda} > 0$ and $\bar{\lambda} < +\infty$. Thus for any $(t, \Omega) \in \mathbb{R}^k \times \mathcal{O}$, for any $A \in \mathbb{R}^{(l+1) \times l}$ full rank, for any $b \neq 0_{l+1}$, for any $K \in \mathbb{R}^{l+1}$

$$|(b_0' \Omega b_0)^{-1/2} b_0' d \ cos(t'c)| \leqslant \underline{\lambda}^{-1/2} ||b_0||_2^{-1/2} |b_0' d|$$
$$|(A_0' \Omega^{-1} A_0)^{-1/2} A_0' \Omega^{-1} d \ cos(t'c)|_1 \leqslant \bar{\lambda}^{1/2} \underline{\lambda}^{-1} |(A_0' A_0)^{-1/2} A_0 d|_1$$

Thus an envelope for $\mathcal{F}_{cos,S}$ is $F_{\mathcal{F}_{cos,S}} : d \mapsto \underline{\lambda}^{-1/2} ||b_0||_2^{-1/2} |b_0' d|$, $F_{\mathcal{F}_{cos,S}}(Y_i)$ is square integrable by A(iii), in addition $1_{F_{\mathcal{F}_{cos,S}}(Y_i) > M} \underset{M \to +\infty}{\to} 0$, thus by the dominated convergence theorem (DCT) $\mathrm{E}(F_{\mathcal{F}_{cos,S}}(Y_i)^2 1_{F \geqslant M}) \underset{M \to +\infty}{\to} 0$. Envelopes by coordinate for $\mathcal{F}_{cos,T}$ are $F_{l', \mathcal{F}_{cos,T}} : d \mapsto \bar{\lambda}^{1/2} \underline{\lambda}^{-1} |e_{j'}' (A_0' A_0)^{-1/2} A_0 d|$ with $l' = 1, \ldots, l$, using previous arguments by the DCT they also satisfy the condition. Using similar arguments $\mathcal{F}_S$, $\mathcal{F}_T$ and $\mathcal{F}_z$ also satisfy this condition and have a "bounded" envelope.

Next, by assumption D(ii) $\mathcal{O}$ has finite covering number so it is BUEI. The fact that the composite of Ω are BUEI still needs to be proven. To do that, I prove that the functions $\Omega \mapsto (A_0' \Omega^{-1} A_0)^{-1/2} A_0' \Omega^{-1} d$ and $\Omega \mapsto (b_0' \Omega b)^{-1/2} b_0' d$ are Lipschitz and bounded in Ω $\forall d \in \mathbb{R}^{l+1}$ so that by Lemma 9.14 in Kosorok (2008) the families $\{d \mapsto (b_0' \Omega b_0)^{-1/2} b_0' d, \Omega \in \mathcal{O}\}$ and $\{d \mapsto (A_0' \Omega^{-1} A_0)^{-1/2} A_0' \Omega^{-1} d, \Omega \in \mathcal{O}\}$ are BUEI.

Boundedness is trivial from the fact that any covariance function in $\mathcal{O}$ has eigenvalues in a bounded subset of $\mathbb{R}_*^+$. First I deal with Lipschitz continuity of the first function

$$\begin{aligned}\frac{|b_0'd(b_0'\Omega_1 b_0)^{-1/2} - b_0'd(b_0'\Omega_2 b_0)^{-1/2}|}{|(b_0'\Omega_1 b_0)^{-1} - (b_0'\Omega_2 b_0)^{-1}|} &= \frac{|b_0'd(b_0'\Omega_1 b_0)^{-1/2} - b_0'd(b_0'\Omega_2 b_0)^{-1/2}|}{|(b_0'\Omega_1 b_0)^{-1/2} - (b_0'\Omega_2 b_0)^{-1/2}|\ |(b_0'\Omega_1 b_0)^{-1/2} + (b_0'\Omega_2 b_0)^{-1/2}|} \\ &= |b_0'd|\ |(b_0'\Omega_1 b_0)^{-1/2} + (b_0'\Omega_2 b_0)^{-1/2}|^{-1} \\ &\leqslant |b_0'd|\ ||b||_2(\lambda_{max}(\Omega_1)^{-1/2} + \lambda_{max}(\Omega_2)^{-1/2})^{-1} \\ &\leqslant |b_0'd|\ ||b||_2\ \frac{\sqrt{\underline{\lambda}}}{2} \equiv \tilde{K}\end{aligned}$$

where the 3rd line is obtained by (B.9) and the last line by assumption D(ii), and due to the fact that eigenvalues of any $\Omega \in \mathcal{O}$ are bounded in $\mathbb{R}_*^+$. This implies Lipchitzness of $\Omega \mapsto (b_0'\Omega b_0)^{-1/2} b_0'd$ using (B.8)

$$\begin{aligned}|b_0'd(b_0'\Omega_1 b_0)^{-1/2} - b_0'd(b_0'\Omega_2 b_0)^{-1/2}| &\leqslant \tilde{K}\ |(b_0'\Omega_1 b_0)^{-1} - (b_0'\Omega_2 b_0)^{-1}| \\ &\leqslant \tilde{K}\ ||(b_0'\Omega_2 b_0)^{-1}(b_0'\Omega_2 b_0 - b_0'\Omega_1 b_0)(b_0'\Omega_1 b_0)^{-1}||_2 \\ &\leqslant \tilde{K}\ ||b||_2^{-2}\lambda_{min}(\Omega_1)^{-1}\lambda_{min}(\Omega_2)^{-1}||b_0'(\Omega_2 - \Omega_1)b||_2 \\ &\leqslant \tilde{K}\ ||b||_2^{-2}\underline{\lambda}^{-2}||b_0'(\Omega_2 - \Omega_1)b||_2 \\ &\leqslant \frac{\tilde{K}}{||b||_2\ \underline{\lambda}^2}||\Omega_1 - \Omega_2||_2 \\ &= \frac{|b_0'd|\sqrt{\underline{\lambda}}}{2\underline{\lambda}^2}||\Omega_1 - \Omega_2||_2\end{aligned}$$

Next without loss of generality assume that $\Omega_1 > \Omega_2$. This implies that

$$\Omega_1^{-1} < \Omega_2^{-1} \Rightarrow\ A_0'\Omega_1^{-1}A_0 < A_0'\Omega_2^{-1}A_0 \Rightarrow (A_0'\Omega_1^{-1}A_0)^{-1/2} > (A_0'\Omega_2^{-1}A_0)^{-1/2}$$

Then using (B.10) I obtain that

$$\frac{1}{2}((A_0'\Omega_1^{-1}A_0)^{-1/2} + (A_0'\Omega_2^{-1}A_0)^{-1/2}) \geqslant \frac{1}{2}(A_0'A_0)^{-1/2}(\lambda_{min}(\Omega_1)^{-1/2} + \lambda_{min}(\Omega_2)^{-1/2})^{-1} \geqslant (A_0'A_0)^{-1/2}\sqrt{\underline{\lambda}}$$

It then follows using (B.7)

$$\begin{aligned}||(A_0'\Omega_1^{-1}A_0)^{-1} - (A_0'\Omega_2^{-1}A_0)^{-1}||_2 &= ||\frac{1}{2}((A_0'\Omega_1^{-1}A_0)^{-1/2} + (A_0'\Omega_2^{-1}A_0)^{-1/2})((A_0'\Omega_1^{-1}A_0)^{-1/2} - (A_0'\Omega_2^{-1}A_0)^{-1/2}) \\ &\quad + \frac{1}{2}((A_0'\Omega_1^{-1}A_0)^{-1/2} - (A_0'\Omega_2^{-1}A_0)^{-1/2})((A_0'\Omega_1^{-1}A_0)^{-1/2} + (A_0'\Omega_2^{-1}A_0)^{-1/2})||_2 \\ &\geqslant \sqrt{\underline{\lambda}}\ ||(A_0'A_0)^{-1/2}((A_0'\Omega_1^{-1}A_0)^{-1/2} - (A_0'\Omega_2^{-1}A_0)^{-1/2}) \\ &\quad + ((A_0'\Omega_1^{-1}A_0)^{-1/2} - (A_0'\Omega_2^{-1}A_0)^{-1/2})(A_0'A_0)^{-1/2}||_2 \\ &= 2\sqrt{\frac{\underline{\lambda}}{\lambda_{max}(A_0'A_0)}}||(A_0'\Omega_1^{-1}A_0)^{-1/2} - (A_0'\Omega_2^{-1}A_0)^{-1/2}||_2 \\ &\equiv \tilde{\tilde{K}}||(A_0'\Omega_1^{-1}A_0)^{-1/2} - (A_0'\Omega_2^{-1}A_0)^{-1/2}||_2\end{aligned}$$

Using (B.6) the above inequality implies the following

$$\begin{aligned}||(A_0'\Omega_1^{-1}A_0)^{-1/2} - (A_0'\Omega_2^{-1}A_0)^{-1/2}||_2 &\leqslant \tilde{\tilde{K}}^{-1}\ ||(A_0'\Omega_1^{-1}A_0)^{-1} - (A_0'\Omega_2^{-1}A_0)^{-1}||_2 \\ &\leqslant \tilde{\tilde{K}}^{-1}\ ||(A_0'\Omega_1^{-1}A_0)^{-1}||_2\ ||(A_0'\Omega_2^{-1}A_0)^{-1}||_2\ ||A_0||_2^2\ ||\Omega_1 - \Omega_2||_2 \\ &\leqslant \tilde{\tilde{K}}^{-1}\ ||(A_0'A_0)^{-1}||_2^2\ \bar{\lambda}^2\ ||\Omega_1 - \Omega_2||_2\end{aligned}$$

Then using the triangular inequality and previous arguments it follows that $\Omega \mapsto (A_0'\Omega^{-1}A_0)^{-1/2}A_0'\Omega^{-1}d$ is Lipschitz $\forall d \in \mathbb{R}l + 1$ and for some strictly positive constant $\tilde{\tilde{\tilde{K}}}$

$$\begin{aligned}||(A_0'\Omega_1^{-1}A_0)^{-1/2}A_0'\Omega_1^{-1}K - (A_0'\Omega_2^{-1}A_0)^{-1/2}A_0'\Omega_2^{-1}K||_2 &\leqslant ||(A_0'\Omega_1^{-1}A_0)^{-1/2}A_0'\Omega_1^{-1}K - (A_0'\Omega_1^{-1}A_0)^{-1/2}A_0'\Omega_2^{-1}K||_2 \\ &\quad + ||(A_0'\Omega_1^{-1}A_0)^{-1/2}A_0'\Omega_2^{-1}K - (A_0'\Omega_2^{-1}A_0)^{-1/2}A_0'\Omega_2^{-1}K||_2 \\ &\leqslant |\tilde{\tilde{\tilde{K}}}|\ ||\Omega_1 - \Omega_2||_2\end{aligned}$$

As for the class $\{c \mapsto t'c, t \in \mathbb{R}^k\}$ it has Vapnick-Cervonenkis index $k+1$ hence by Sauer's Lemma (Vaart (2007)) it is BUEI. From there as the cosinus function is also bounded and Lipschitz continuous then the class $\{c \mapsto cos(t'c), t \in \mathbb{R}^k\}$ is BUEI by 9.14 in Kosorok (2008). The Lipschitz-continuity in t can be proven with the mean value theorem in the following way

$$\begin{aligned}\frac{|cos(t_1'c) - cos(t_2'c)|}{||t_1 - t_2||_2} &\leqslant \frac{|cos(t_1'c) - cos(t_2'c)|}{|(t_1 - t_2)'c|}\ \frac{|(t_1 - t_2)'c|}{||t_1 - t_2||_2} \\ &\leqslant ||c||_2 \frac{|cos(t_1'c) - cos(t_2'c)|}{|(t_1 - t_2)'c|} \leqslant ||c||_2\end{aligned}$$

Then by theorem 9.15 in Kosorok (2008) the "product" of 2 BUEI families is BUEI so that $\mathcal{F}_{cos,T}$ and $\mathcal{F}_{cos,S}$ are BUEI. Replacing the cosinus function by the sinus function keep these two families BUEI therefore by applying Lemma 9.14 from Kosorok (2008), the families $\mathcal{F}_T$, $\mathcal{F}_S$ and $\mathcal{F}_z$ are BUEI.

B.4 Limits of G_S, G_T and G_z

Next, from the previous results G_S and G_S^* converge uniformly towards Gaussian processes if demeaned whereas G_z converges uniformly towards a function with values in $\mathbb{C}$. As for G_T and G_T^*, depending on instruments' strength a it can be decomposed as the sum of different terms, some of which are uniformly degenerate, some of which converge uniformly towards functions with values in $\mathbb{C}$, and some of which converge uniformly towards Gaussian processes. Note that the families of functions over which G_S, G_T, G_ϵ, G_ζ, and G_z are defined such as $\mathcal{F}_T$ and $\mathcal{F}_S$ are Donsker thus also Glivenko-Cantelli. These limits are characterized in the following way:

1. Regarding G_S recall $G_S(t,\Omega) = \frac{1}{\sqrt{n}}\sum_{i=1}^n S_{ih}e^{it'z_i} = \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i' b_0 (b_0'\Omega_i b_0)^{-1/2} e^{it'z_i}$, therefore under the null the and $\forall t \in \mathbb{R}^k,\ \forall \Omega \in \mathcal{O}$ the asymptotic mean of G_S writes

$$\mathrm{E}(S_{ih}e^{it'z_i}) = \mathrm{E}(\mathrm{E}(y_i - x_i'\beta_0|z_i)(b_0'\Omega_i b_0)^{-1/2}e^{it'z_i}) = \mathrm{E}(\epsilon_i e^{it'z_i}) = 0$$

And $\forall (t,\tilde{t}),\ \forall \Omega \in \mathcal{O}$, the asymptotic covariance of G_S writes

$$\mathrm{Var}(S_{ih}e^{it'z_i}) = \mathrm{E}(\mathrm{E}((y_i - x_i'\beta_0)^2|z_i)(b_0'\Omega_i b_0)^{-1}e^{2it'z_i}) = \mathrm{Var}(u_i e^{it'z_i}) = \mathrm{E}(e^{2it'z_i})$$
$$\mathrm{Cov}(S_{ih}e^{it'z_i}, S_{ih}e^{i\tilde{t}'z_i}) = \mathrm{E}(e^{i(t+\tilde{t})'z_i})$$

Consequently $\forall \Omega \in \mathcal{O}$ the limit of $G_S(\cdot,\Omega)$ is a Gaussian process with, under the null, a constant mean equal to 0 and a covariance function $(t,\tilde{t}) \mapsto \mathrm{E}(e^{i(t+\tilde{t})'z_i})$. Note that under any alternative $H_1 : \beta \neq \beta_0$ the asymptotic equicontinuity result still holds, ie the difference between $G_S(\cdot,\Omega)$ and $G_S(\cdot,\hat{\Omega})$ vanishes, however G_S does not converge towards the aforementioned Gaussian process. Characterization of the limit of G_S in Ω is unecessary for the proof, G_S is considered a function of Ω only in order to prove convergence of the feasible process $G_S(\cdot,\hat{\Omega})$.

Regarding G_S^*, it has the same limit as G_S because the conditional normality doesn't play a role asymptotically, indeed $\mathrm{E}(Y_i) = \mathrm{E}(Y_i^*)$ and $\mathrm{Var}(Y_i) = \mathrm{Var}(Y_i^*)$. Thus for any $\Omega \in \mathcal{O}$ both $G_S(\cdot,\Omega)$ and $G_S^*(\cdot,\Omega)$ converge uniformly towards the same complex Gaussian process. This implies that the distance between the 2 shrinks asymptotically uniformly by definition of weak convergence.

2. Regarding G_z its limit is the characteristic function of z_i $t \mapsto \mathrm{E}(e^{it'z_i})$.

3. Regarding G_T, first recall

$$T_{ih,j} = \sum_{l'=1}^{l} w_{ij,l'}^* x_{il'} + w_{ij,y}u_i = \sum_{l'=1}^{l} w_{ij,l'}^*(\frac{C(z_i)_{l'}}{n^a} + v_{il'}) + w_{ij,y}u_i = n^{-a}w_{ij}^{*\prime}C(z_i) + w_{ij}^{*\prime}v_i + w_{ij,y}u_i$$

Therefore for some $j = 1,\dots,l$ G_{Tj} rewrites

$$G_{Tj}(t,\Omega) = \frac{n^{a_j^*-1/2}1_{a_j^*<1/2} + 1_{a_j^*\geqslant 1/2}}{\sqrt{n}}\sum_{i=1}^n T_{ih,j}e^{-it'z_i}$$

where $a_j^* = \begin{cases} a & \text{if } (w_{ij,l'}^*)_{l'=1}^l \neq 0_l \\ 1 & \text{o.w} \end{cases}$ so if $a_j^* \geqslant 1/2$ then G_{T_j} writes

$$G_{T_j}(t,\Omega) = \sum_{l'=1}^{l}\frac{1}{n^{1/2+a}}\sum_{i=1}^n w_{ij,l'}^* C(z_i)_{l'}e^{-it'z_i} + \sum_{l'=1}^{l}\frac{1}{\sqrt{n}}\sum_{i=1}^n w_{ij,l'}^* v_{il'}e^{-it'z_i}$$
$$+ \frac{1}{\sqrt{n}}\sum_{i=1}^n w_{ij,y}u_i e^{-it'z_i}$$

And if $a_j^* < 1/2$ it must be that $a_j^* = a$ so G_{T_j} writes

$$G_{T_j}(t,\Omega) = \sum_{l'=1}^{l} \frac{1}{n} \sum_{i=1}^{n} w_{ij,l'}^* C(z_i)_{l'} e^{-it'z_i} + \sum_{l'=1}^{l} \frac{n^{a-1/2}}{\sqrt{n}} \sum_{i=1}^{n} w_{ij,l'}^* v_{il'} e^{-it'z_i} + \frac{n^{a-1/2}}{\sqrt{n}} \sum_{i=1}^{n} w_{ij,y} u_i e^{-it'z_i}$$

It follows then that if $a_j^* > 1/2$ then, either $\forall l'\ w_{ij,l'} = 0$ and $a_j^* = 1$, either $\exists l' : \mathrm{E}(w_{ij,l'} C(z_i)_{l'}) \neq 0$ and $a_j^* = a > 1/2$. Thus if $a_j^* > 1/2$

$$\sum_{l'=1}^{l} \frac{1}{n^{1/2+a}} \sum_{i=1}^{n} w_{ij,l'}^* C(z_i)_{l'} e^{-it'z_i} = o_{\underline{\mathrm{P}}}(1)$$

This implies that $G_{T_j}(\cdot,\Omega)$ will converge to a Gaussian process with constant mean 0 and covariance function $(t,s) \mapsto \mathrm{E}(e^{-i(t+s)'z_i})$ because $\sum_{l'=1}^{l} \mathrm{E}(w_{ij,l'} v_{il'} e^{-t'z_i}) = \mathrm{E}(w_{ij,y} u_i e^{-it'z_i}) = 0$ by the law of iterated expectations and because T_{ih} has variance identity conditionally on z_i.

If $a_j^* = 1/2$ then $G_{T_j}(\cdot,\Omega)$ will converge towards a non-centered Gaussian process with mean $t \mapsto \sum_{l=1}^{l}{}' \mathrm{E}(w_{ij,l'}^* C(z_i) e^{-itz_i})$ and covariance function $(t,s) \mapsto \mathrm{E}(e^{-i(t+s)'z_i})$.

If $a_j^* < 1/2$ then $\exists l' : w_{ij,l'}^* \neq 0$ and $a_j^* = a$. Then

$$\sum_{l'=1}^{l} \frac{n^{a-1/2}}{\sqrt{n}} \sum_{i=1}^{n} w_{ij,l'}^* v_{il'} e^{-it'z_i} + \frac{n^{a-1/2}}{\sqrt{n}} \sum_{i=1}^{n} w_{jyi}^* u_i e^{-it'z_i} = o_{\underline{\mathrm{P}}}(1)$$

Therefore $G_{T_j}(\cdot,\Omega)$ will converge towards $t \mapsto \sum_{l'=1}^{l} \mathrm{E}(w_{ij,l'}^* C(z_i)_{l'} e^{-it'z_i})$ for any $\Omega(\cdot) \in \mathcal{O}$.

Regarding the whole vector T_i, because T_i has conditional variance identity (even under the alternative), $G_T(\cdot,\Omega) = (G_{T_1} \ \cdots \ G_{T_l})$ will converge to a vector of random processes whose elements are uncorrelated and characterized by the limits G_{T_j} for $j = 1,\ldots,l$ which are characterized by the vector $(a_j^*)_{j=1}^{l}$.

Now with regards to G_T^*, it has the same limit as G_T for the same reason that G_S^* and G_S have the same limit: the conditional normality does not play a role anymore asymptotically, asymptotically what matters are the means and the covariances of the limitting processes which in the case G_T and G_T^* are the same. In conclusion, asymptotically the distance between $G_T(\cdot,\Omega)$ and $G_T^*(\cdot,\Omega)$ vanishes for any $\Omega \in \mathcal{O}$ uniformly.

B.5 Lipschitzness of KICM

Define f and g for any triple of functions (A, B, C) with inputs in $\mathbb{R}^k$ and values in a bounded subset of $\mathbb{C}$, $\mathbb{C}^l \setminus 0_l$ and $\mathbb{C}_*^+$ respectively

$$f(A,B) = \int_{\mathbb{R}^k} A(t)B(t)\mu(t), \quad g(B,C) = \int_{\mathbb{R}^k}\int_{\mathbb{R}^k} B(t)B'(t)C(s-t)d\mu(s)d\mu(t)$$

Notice that $\text{KICM}_h = ||f(G_S, G_T)g(G_T, G_z)||_2^2$ so I shall prove that KICM_h is Lipschitz in G_S, G_T and G_z through f and g. In the following K is an unspecified constant.

- To prove Lipschitzness in A note that because (A, B, C) take values in bounded subset I can always find a supremum to $A(t)$ and $B(t)$ and an infimum to $\lambda_{min}(g(B, C))$

$$\begin{aligned}
| \, ||f(A_1,B)g(B,C)^{-1/2}||_2^2 - ||f(A_2,B)g(B,C)^{-1/2}||_2^2 \, | &= | \int (A_1 - A_2)(t)B(t)'d\mu(t)g(B,C)^{-1} \\
&\quad \int (A_1 + A_2)(t)B(t)d\mu(t) \, | \\
&\leqslant ||A_1 - A_2||_\infty \int ||B(t)||_2 d\mu(t) \lambda_{min}(g(B,C))^{-1} \\
&\quad \int |A_1(t) + A_2(t)| \, ||B(t)||_2 d\mu(t) \\
&\leqslant K \times \sup_t |A_1(t) - B_2(t)|
\end{aligned}$$

- To prove Lipschitzness in C I use result ((B.6))

$$\begin{aligned}
||g(B,C_1) - g(B,C_2)||_2 &= \leqslant \lambda_{min}(g(B,C_1))^{-1}\lambda_{min}(g(B,C_2))^{-1} \\
&\quad \times \int\int ||B(t)B(-s)'||_2 d\mu(t)d\mu(s)||C_1 - C_2||_\infty \\
&\leqslant K \, ||C_1 - C_2||_\infty
\end{aligned}$$

  Then Lipschitzness of $C \mapsto ||f(A,B)g(B,C)^{-1/2}||_2^2$ is established because it is a function of C only through g and because I can find an upper bound on $f(A, B)$.

- To prove Lipchitzness in B I express $| \, ||f(A,B_1)g(B_1,C)^{-1/2}||_2^2 - ||f(A,B_2)g(B_2,C)^{-1/2}||_2^2 \, |$ as a sum of 3 components which depend on $f(A, B_1) - f(A, B_2)$ and $g(B_1, C) - g(B_2, C)$ then by the triangular inequality and for some positive constants (K_1, K_2)

$$\begin{aligned}
&| \, ||f(A,B_1)g(B_1,C)^{-1/2}||_2^2 - ||f(A,B_2)g(B_2,C)^{-1/2}||_2^2 \, | \\
&= |f(A,B_1)'(g(B_1,C)^{-1} - g(B_2,C)^{-1})f(A,B_1) \\
&\quad + f(A,B_2)' \left(g(B_1,C)^{-1}(f(A,B_1) - f(A,B_2)) - (g(B_2,C)^{-1} - g(B_1,C)^{-1})f(A,B_2)\right) | \\
&\leqslant K_1 \, ||g(B_1,C)^{-1} - g(B_2,C)^{-1}||_2 + K_2 \, ||f(A,B_1) - f(A,B_2)||_2
\end{aligned}$$

  Reusing (B.6) and aforementioned arguments for the Lipchitzness in A and C of $(A, B, C) \mapsto\mapsto ||f(A,B)g(B,C)^{-1/2}||_2^2$ it is also Lipchitz in B.

The Lipschitzness of KICM_h in G_S and G_T allows for the difference between feasible KICM (KICM_{hf}) and unfeasible KICM (KICM_h) to vanish uniformly asymptotically.

B.6 KICM Validity by Contradiction

Let $q_{1-\alpha}$ be the $1-\alpha$ quantile of a χ^2_l then suppose that the theorem does not hold. The theorem not holding is equivalent to $\exists P$ such that $\beta = \beta_0$ and $\text{P}(\text{KICM}_h > q_{1-\alpha}) > \alpha$ which implies that $\exists \delta > 0 : \text{P}(\text{KICM}_h > q_{1-\alpha}) \geqslant \alpha + 2\delta$.

Next, I can find bounded subsets $C_S \subset \mathbb{C}_*$, $C_T \subset \mathbb{C}^l_*$ and $C_z \subset \mathbb{C}_*$ such that $\delta > \text{P}(G_S \notin C_S, G_T \notin C_T, G_z \notin C_z) = 1 - \text{P}(G_S \in C_S, G_T \in C_T, G_z \in C_z)$. Indeed, these C_S, C_T and C_z exist because G_S, G_T and G_z are uniformly bounded in probability and non-degenerate as shown in B.4. Note that, because the integral of a bounded random process with respect to a finite measure is bounded and the product of bounded random variables is bounded, the event $G_S \in C_S \cap G_T \in C_T \cap G_z \in C_z$ implies some event $\text{KICM}_h \leqslant C$ which is consistent with the fact that $\text{KICM}_h = O_{\bar{\text{P}}}(1) \neq o_{\underline{\text{P}}}(1)$.

Next, let K_h be KICM_h but with Y_i assumed normal conditionally on z_i. Then as I saw before in 4.1 $\text{K}_h \sim \chi^2_l | z, T \Rightarrow \text{K}_h \sim \chi^2_l$ hence the $1-\alpha$ quantile of K_h is $q_{1-\alpha}$ the quantile $1-\alpha$ of a χ^2_l. Then introduce $\text{KICM}_{hC} = \text{KICM}_h \, 1_{G_S \in C_S, G_T \in C_T, G_z \in C_z}$ and $\text{K}_{hC} = \text{K}_h \, 1_{G_\epsilon \in C_S, G_T \in C_T, G_z \in C_z}$. I define the quantile of K_{hC} as $q_{C,1-\alpha} = inf\{q : \text{P}(\text{K}_{hC} \leqslant q) \geqslant 1-\alpha\}$.

Thus because $\text{KICM}_h = \text{KICM}_{hC}$ if $G_S \in C_S, G_T \in C_T, G_z \in C_z$ it follows that:

$$\begin{aligned}
1_{\text{KICM}_h > x} &= 1_{\text{KICM}_h > x, G_S \in C_S, G_T \in C_T, G_z \in C_z} + 1_{\text{KICM}_h > x, \overline{G_S \in C_S, G_T \in C_T, G_z \in C_z}} \\
&= 1_{\text{KICM}_{hC} > x, G_S \in C_S, G_T \in C_T, G_z \in C_z} + 1_{\text{KICM}_h > x, \overline{G_S \in C_S, G_T \in C_T, G_z \in C_z}} \\
&\leqslant 1_{\text{KICM}_{hC} > x} + 1_{\overline{G_S \in C_S, G_T \in C_T, G_z \in C_z}}
\end{aligned}$$

By taking the mean it implies that,

$$\text{P}(\text{KICM}_h > x) \leqslant \text{P}(\text{KICM}_{hC} > x) + \text{P}(\overline{G_S \in C_S, G_T \in C_T, G_z \in C_z}) < \text{P}(\text{KICM}_{hC} > x) + \delta$$

Then using $\text{P}(\text{KICM}_h > q_{1-\alpha}) > \alpha + 2\delta$ I obtain:

$$\text{P}(\text{KICM}_{hC} > q_{1-\alpha}) > \alpha + \delta$$

Additionally $\text{KICM}_{hC} \leqslant \text{KICM}_h$ implies that $q_{C,1-\alpha} \leqslant q_{1-\alpha} \forall \alpha$ which leads to,

$$\text{P}(\text{KICM}_{hC} > q_{C,1-\alpha}) > \alpha + \delta$$

Going forward note that $x \in \mathbb{R}^+ \mapsto x1_{x<C}$ is bounded by C and Lipschitz. At the same time KICM_h is Lipschitz in G_S, G_T and G_z as long as these take values in bounded sets as saw in B.6 thus KICM_{hC} is Lipschitz in G_S, G_T and G_z. Now that the difference between G_S and G_S^* and between G_T and G_T^* vanishes uniformly by B.4, hence by the Portemanteau Lemma, as KICM_{hC} is Lipschitz and bounded, then KICM_{hC} and K_{hC} converge uniformly towards the same distribution which is the distribution of K_{hC}. It implies by definition of weak convergence that,

$$\sup_x |\text{P}(\text{KICM}_{hC} > x) - \text{P}(\text{K}_{hC} > x)| \to 0$$

Consequently asymptotically

$$\lim_{n\to\infty} \text{P}(\text{KICM}_{hC} > q_{C,-\alpha}) = \text{P}(\text{K}_{hC} > q_{C,1-\alpha}) = \alpha \geqslant \alpha + \delta \Leftrightarrow \delta \leqslant 0$$

Which is impossible because $\delta > 0$.

On a final note, the Lipschitzness and boundedness of KICM_{hC} and K_{hC} in G_S and G_T imply that the difference between the feasible and unfeasible statistics vanishes uniformly by asymptotic equicontinuity uniform of the processes G_S and G_T. Therefore the contradiction still holds even if KICM_{hf} is used instead of KICM_h, and K_{hf} instead of K_h (where K_{hf} denotes KICM_{hf} but with Y_i conditionally normal).

C Proof of corollary 4.4

For any null β_0, G_T is still bounded and bounded away from 0 as shown in B.4 uniformly over the $\text{P} \in \mathcal{P} : \beta \neq \beta_0$. G_T's limit is different under the alternative tough as it depends on the actual distribution of y_i. As for G_S and G_S^* the difference between the two will also vanish as per the arguments of B.4, even if G_S and G_S^* don't have a limit but explode. Before deriving the power of KICM under different types of identification let's first decompose G_S under the alternative $\beta \neq \beta_0$:

$$\begin{aligned} G_S(t,\Omega) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n S_{ih} e^{it'z_i} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{y_i - x_i'\beta_0}{\sqrt{b_0'\Omega_i b_0}} e^{it'z_i} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{y_i - x_i'\beta}{\sqrt{b_0'\Omega_i b_0}} e^{it'z_i} + \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{x_i'(\beta-\beta_0)}{\sqrt{b_0'\Omega_i b_0}} e^{it'z_i} \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n (u_i + v_i'(\beta-\beta_0)) \frac{e^{it'z_i}}{\sqrt{b_0'\Omega_i b_0}} + \frac{1}{n^{1/2+a}}\sum_{i=1}^n C(z_i)'(\beta-\beta_0)\frac{e^{it'z_i}}{\sqrt{b_0'\Omega_i b_0}} \end{aligned}$$

The term on the left will, reusing arguments from B.4, converge towards a complex Gaussian process with mean 0 and covariance $(s,t) \mapsto \text{E}(Me^{i(s+t)'z_i})$ with

$$M = (1 \ \beta' - \beta_0') \frac{\Omega_i}{b_0'\Omega_i b_0} (1 \ \beta' - \beta_0')'$$

As a consequence

$$G_S(t,\Omega) = O_{\bar{\mathrm{P}}}(1) + \frac{n^{1/2-a}}{n}\sum_{i=1}^{n} C(z_i)'(\beta-\beta_0)\frac{e^{it'z_i}}{\sqrt{b_0'\Omega_i b_0}} = O_{\bar{\mathrm{P}}}(1) + O_{\bar{P}}(n^{1/2-a})$$

And there are 3 possible behaviors of KICM:

1. If $a < 1/2$ then $\frac{G_S(t,\Omega)}{n^{1/2-a}} = O_{\bar{\mathrm{P}}}(1) \neq o_{\underline{\mathrm{P}}}(1)$. In other words $G_S(\cdot,\Omega)/n^{1/2-a}$ is uniformly bounded in probability and uniformly non-degenerate and converges towards the following function,

$$t \mapsto \mathrm{E}(C(z_i)'\frac{e^{it'z_i}}{\sqrt{b_0'\Omega_i b_0}})(\beta-\beta_0)$$

   Then because $G_S/n^{1/2-a}$, G_T and G_z are $O_{\bar{\mathrm{P}}}(1)$ then for any $\zeta > 0$ there exists bounded subsets $C_S \subset \mathbb{C}_*$, $C_T \subset \mathbb{C}_*^l$ and $C_z \in \mathbb{C}_*$ such that $P(G_S/n^{1/2-a} \notin C_S, G_T \notin C_T, G_z \notin C_z) < \zeta$. Also note that, conditionally on the event $C = (G_S/n^{1/2-a} \in C_S, G_T \in C_T, G_z \in C_z)$, $\mathrm{KICM}_h/n^{1-2a} > K$ surely where K is a strictly positive constant (by properties of integrals).

   Now assume that for some $x \in \mathbb{R}^+$, $\mathrm{P}(\mathrm{KICM}_h > x) \to 1$ doesn't hold. Then it implies that $\exists \eta \in [0;1)$ such that $\lim_{n\to\infty} \mathrm{P}(\mathrm{KICM}_h > x) < \eta$ or equivalently that $\exists(\eta,\zeta) \in [0;1) \times \mathbb{R}_*^+$ such that $\lim_{n\to\infty} \mathrm{P}(\mathrm{KICM}_h > x) \leqslant \eta - \zeta$. Additionally,

$$\begin{aligned}
\eta - \zeta \geqslant \mathrm{P}(\mathrm{KICM}_h > x) &= \mathrm{P}(\mathrm{KICM}_h > x|C)\,\mathrm{P}(C) + \mathrm{P}(\mathrm{KICM}_h > x|\bar{C})\,\mathrm{P}(\bar{C}) \\
&\geqslant \mathrm{P}(\mathrm{KICM}_h > x|C)\,\mathrm{P}(C) \\
&= \mathrm{P}(\mathrm{KICM}_h/n^{1-2a} > x/n^{1-2a}|C)(1-\zeta) \\
&\geqslant \mathrm{P}(K > x/n^{1-2a}|C)(1-\zeta) \\
&= \mathrm{P}(K > x/n^{1-2a})(1-\zeta) \\
&\to 1-\zeta
\end{aligned}$$

   Which is equivalent to $\eta - \zeta \geqslant 1 - \zeta$ asymptotically which is impossible for any $x \in \mathbb{R}^+$ thus the test is consistent. It trivially follows that this contradiction holds uniformly over the β_0 and the $P : \beta \neq \beta_0,\ a < 1/2$ per the results in B.3.

2. If $a = 1/2$ then G_S and G_S^* converges towards the same non-centered complex Gaussian process. This means that the distance between KICM_h under conditional normality and KICM_h still vanishes. Then KICM_h does not follow a χ_l^2 asymptotically anymore but a non centered $\chi_{l,\lambda}^2$ where $\lambda = \mathrm{E}\left(\frac{C(z_i)'(\beta-\beta_0)}{\sqrt{b_0'\Omega_i b_0}}\right)^2$. Indeed when Y_i is conditionally normal then even if $\mathrm{E}(S_{ih}|z_i) \neq 0$, $\mathrm{Var}(S_{ih}|z_i) = 1$

and $\text{Cov}(S_{ih}, T_{ih}) = 0_l$ which implies that S is independent from T. Consequently under conditional normality KICM_h is indeed the sum of l independent non-centered standard normal variable thus conditionally on z and T and assuming conditional normality $\text{KICM}_h \sim \chi^2_{l,\lambda}$. Furthermore the cdf of a non-central χ^2 evaluated at any point is strictly decreasing in its non-centrality parameter. Therefore,

$$\begin{aligned} &\text{P}(\text{KICM}_h > x|z,T) \to \text{P}(\chi^2_{l,\lambda} > x) > \text{P}(\chi^2_l > x) \\ \Rightarrow &\lim_{n\to\infty} \text{P}(\text{KICM}_h > q_{1-\alpha}) = \lim_{n\to\infty} \text{E}(\text{P}(\text{KICM}_h > q_{1-\alpha}|z,T)) > \alpha \end{aligned}$$

In other words, there is more than trivial power. As the convergence towards a $\chi^2_{l,\lambda}$ can be proven to be uniform over the β_0 and $\text{P} : \beta \neq \beta_0,\ a = 1/2$ per arguments in B.3

$$\lim_{n\to\infty} \underset{\beta_0}{sup} \underset{P\in\mathcal{P}:\beta\neq\beta_0,\ a=1/2}{sup} \text{P}(\text{KICM}_h > q_{1-\alpha}) \geqslant \alpha$$

3. If $a > 1/2$ then reusing arguments from B.4 the processes $G_S(\cdot,\Omega)$ and $G^*_S(\cdot,\Omega)$ will converge to the same centered complex Gaussian process as under the null because S_{ih} has mean 0 asymptotically. Therefore, following the proof of theorem 4.3, KICM_h will still behave as in the conditionally normal case. From there, because $\text{Var}(S) = I_n$ still, KICM_h will behave like the sum of l standard normal and follow a χ^2_l. Thus the probability to reject the null at nominal level α will be inferior or equal to α uniformly over the $P : \beta \neq \beta_0$ and the β_0. The omitted part of the proof is a copy of the proof of theorem 4.3.

D Non-Uniform Identification Strength

It is important to note that one could assume that the parameters do not have the same degree of identification as in assumption B(iii) instead of B(ii). Then the asymptotic behavior of the KICM test would remain relatively unchanged using a new a^*_j to characterize the asymptotics

$$a^*_j = \begin{cases} min\{a_j : \beta_j \neq \beta_{0j}\} & \text{if}(w^*_{ij,l'})^l_{l'=1} \neq 0_l \\ 1 & \text{o.w} \end{cases}$$

Non-uniformity of the degree of identification of β does not affect the size of KICM, but affects its power. Among the components j of β such that the null being tested $\beta_{0j} \neq \beta_j$ if at least one parameter is semi-strongly identified then the test is consistent, if at least one parameter is weakly identified then the test has more than trivial power, if all the parameters are very weakly identified then the test has trivial power. These results are summarized in the following corollary

Corollary 4.1 (Uniform consistency of KICM, non-uniform identification strength) *Denote by $q_{1-\alpha}$ the $1-\alpha$ quantile of the chi-square distribution with l degrees of freedom. Then under assumptions A(ii), A(iii), B(i), B(iii), B(iv), C(i), C(ii), D(ii), and D(iii),*

- $\lim_{n\to\infty} \inf_{\beta_0} \inf_{P\in P:\beta\neq\beta_0, min\{a_j:\beta_j\neq\beta_{0j}\}<1/2} \mathrm{P}(\mathrm{KICM}_{hf} > q_{1-\alpha}) = 1$*;*
  The test is consistent when at least one parameter is semi-strongly identified.

- $\lim_{n\to\infty} \inf_{\beta_0} \inf_{P\in\mathcal{P}:\beta\neq\beta_0, min\{a_j:\beta_j\neq\beta_{0j}\}=1/2} \mathrm{P}(\mathrm{KICM}_{hf} > q_{1-\alpha}) \in [\alpha; 1)$*;*
  The test has more than trivial power when at least one parameter is weakly identified.

- $\lim_{n\to\infty} \sup_{\beta_0} \sup_{P\in\mathcal{P}:\beta\neq\beta_0, min\{a_j:\beta_j\neq\beta_{0j}\}>1/2} \mathrm{P}(\mathrm{KICM}_{hf} > q_{1-\alpha}) \leqslant \alpha$*;*
  The test has trivial power when all parameters are very weakly identified.

The proof is extremely similar to that of corollary 4.4 in appendix C and is omitted.

E Plots and tables

E.1 Small sample simulations

E.1.1 Empirical size

Weak instruments	**AR**	**LM**	**CLR**	**ICM**	**KICM**	**CICM**	**W-2SLS**
Linear model							
n=100, m=200	0.1042	0.1042	0.1044	0.1414	0.1112	0.1398	0.1684
n=100, m=500	0.1042	0.1042	0.1028	0.1188	0.1112	0.1122	0.1684
n=400, m=200	0.0976	0.0976	0.0944	0.1024	0.1004	0.0982	0.1592
n=400, m=500	0.0976	0.0976	0.0974	0.1096	0.1004	0.1074	0.1592
n=400, m=500, Heteroskedasticity	0.1056	0.1056	0.1050	0.1164	0.1070	0.1124	0.1848
Non-linear model							
n=100, m=200	0.1170	0.1120	0.1174	0.1364	0.1072	0.1354	0.3964
n=100, m=500	0.1170	0.1120	0.1158	0.1176	0.1072	0.1180	0.3964
n=400, m=200	0.0980	0.0958	0.0942	0.1052	0.0946	0.0992	0.3878
n=400, m=500	0.0980	0.0958	0.0944	0.1122	0.0946	0.1056	0.3878
n=400, m=500, Heteroskedasticity	0.1044	0.1044	0.1044	0.1174	0.1014	0.1060	0.3856
Polar polynomial model							
n=100, m=200	0.1042	0.1042	0.1044	0.1414	0.1034	0.1378	0.2030
n=100, m=500	0.1042	0.1042	0.1028	0.1188	0.1034	0.1200	0.2030
n=400, m=200	0.0976	0.0976	0.0944	0.1024	0.0976	0.1006	0.1986
n=400, m=500	0.0976	0.0976	0.0974	0.1096	0.0976	0.1090	0.1986
n=400, m=500, Heteroskedasticity	0.1056	0.1056	0.1050	0.1164	0.1028	0.1174	0.1992
Semi-polar polynomial model							
n=100, m=200	0.1170	0.1092	0.1134	0.1364	0.1106	0.1366	0.3908
n=100, m=500	0.1170	0.1092	0.1108	0.1176	0.1106	0.1176	0.3908
n=400, m=200	0.0980	0.0958	0.0948	0.1052	0.0954	0.0998	0.3848
n=400, m=500	0.0980	0.0958	0.0934	0.1122	0.0954	0.1060	0.3848
n=400, m=500, Heteroskedasticity	0.1048	0.1010	0.1040	0.1186	0.0990	0.1084	0.3786

Table 1: Empirical size of the tests for nominal size 10%, weak instruments case

Semi-strong instruments	**AR**	**LM**	**CLR**	**ICM**	**KICM**	**CICM**	**W-2SLS**
Linear model							
n=100, m=200	0.1042	0.1042	0.1044	0.1414	0.1098	0.1238	0.1078
n=100, m=500	0.1042	0.1042	0.1028	0.1188	0.1098	0.1174	0.1078
n=400, m=200	0.0976	0.0976	0.0944	0.1024	0.0984	0.0998	0.0908
n=400, m=500	0.0976	0.0976	0.0974	0.1096	0.0984	0.0996	0.0908
n=400, m=500, Heteroskedasticity	0.1056	0.1056	0.1050	0.1164	0.1038	0.1060	0.1024
Non-linear model							
n=100, m=200	0.1170	0.1152	0.1128	0.1364	0.894	0.1228	0.3206
n=100, m=500	0.1170	0.1152	0.1134	0.1176	0.894	0.1106	0.3206
n=400, m=200	0.0980	0.1000	0.0948	0.1052	0.0902	0.0936	0.2844
n=400, m=500	0.0980	0.1000	0.0956	0.1122	0.0902	0.1014	0.2844
n=400, m=500, Heteroskedasticity	0.1048	0.1016	0.1014	0.1164	0.0980	0.1028	0.2940
Polar polynomial model							
n=100, m=200	0.1042	0.1042	0.1044	0.1414	0.0998	0.1350	0.1788
n=100, m=500	0.1042	0.1042	0.1028	0.1188	0.0998	0.1152	0.1788
n=400, m=200	0.0976	0.0976	0.0944	0.1024	0.0994	0.0990	0.1764
n=400, m=500	0.0976	0.0976	0.0974	0.1096	0.0994	0.1056	0.1764
n=400, m=500, Heteroskedasticity	0.1056	0.1056	0.1050	0.1150	0.1042	0.1086	0.1758
Semi-polar polynomial model							
n=100, m=200	0.1170	0.1102	0.1106	0.1364	0.0950	0.1200	0.2620
n=100, m=500	0.1170	0.1102	0.1088	0.1176	0.0950	0.1098	0.2620
n=400, m=200	0.0980	0.0936	0.0942	0.1052	0.0918	0.0976	0.1972
n=400, m=500	0.0980	0.0936	0.0958	0.1122	0.0918	0.1010	0.1972
n=400, m=500, Heteroskedasticity	0.1048	0.1042	0.1018	0.1186	0.0968	0.1016	0.2136

Table 2: Empirical size of the tests for nominal size 10%, semi-strong instruments case

Strong instruments	**AR**	**LM**	**CLR**	**ICM**	**KICM**	**CICM**	**W-2SLS**
Linear model							
n=100, m=200	0.1042	0.1042	0.1044	0.1414	0.1124	0.1128	0.1008
n=100, m=500	0.1042	0.1042	0.1028	0.1188	0.1124	0.1118	0.1008
n=400, m=200	0.0976	0.0976	0.0944	0.1024	0.0976	0.0938	0.0954
n=400, m=500	0.0976	0.0976	0.0974	0.1096	0.0976	0.0962	0.0954
n=400, m=500, Heteroskedasticity	0.1056	0.1056	0.1050	0.1164	0.1068	0.1086	0.0986
Non-linear model							
n=100, m=200	0.1070	0.1056	0.1064	0.1364	0.0914	0.1120	0.1478
n=100, m=500	0.1070	0.1056	0.1030	0.1176	0.0914	0.0964	0.1478
n=400, m=200	0.0980	0.0968	0.0962	0.1052	0.966	0.1012	0.1040
n=400, m=500	0.0980	0.0968	0.0944	0.1122	0.966	0.1018	0.1040
n=400, m=500, Heteroskedasticity	0.1048	0.0960	0.0948	0.1122	0.0964	0.1046	0.0934
Polar polynomial model							
n=100, m=200	0.1042	0.1042	0.1044	0.1414	0.0958	0.1114	0.0836
n=100, m=500	0.1042	0.1042	0.1028	0.1188	0.0958	0.1036	0.0836
n=400, m=200	0.0976	0.0976	0.0944	0.1024	0.1000	0.1004	0.0708
n=400, m=500	0.0976	0.0976	0.0974	0.1096	0.1000	0.1032	0.0708
n=400, m=500, Heteroskedasticity	0.1050	0.1050	0.1036	0.1074	0.0980	0.1052	0.0404
Semi-polar polynomial model							
n=100, m=200	0.1170	0.1096	0.1094	0.1364	0.0902	0.0996	0.1156
n=100, m=500	0.1170	0.1096	0.1074	0.1176	0.0902	0.0924	0.1156
n=400, m=200	0.0980	0.0974	0.0940	0.1052	0.0944	0.1006	0.0960
n=400, m=500	0.0980	0.0974	0.0962	0.1122	0.0944	0.0984	0.0960
n=400, m=500, Heteroskedasticity	0.08	0.10	0.10	0.07	0.10	0.10	0.06

Table 3: Empirical size of the tests for nominal size 10%, strong instruments case

Instruments Strength	**AR**	**LM**	**CLR**	**ICM**	**KICM**	**CICM**	**W-2SLS**
Weak	0.1298	0.1186	0.1270	0.1438	0.0900	0.1370	0.7052
Semi-Strong	0.1298	0.1178	0.1232	0.1438	0.0868	0.1362	0.6506
Strong	0.1298	0.1078	0.1110	0.1438	0.0984	0.1250	0.3704

Table 4: Empirical size for nominal size 10%, 4 Instruments

E.1.2 Power curves

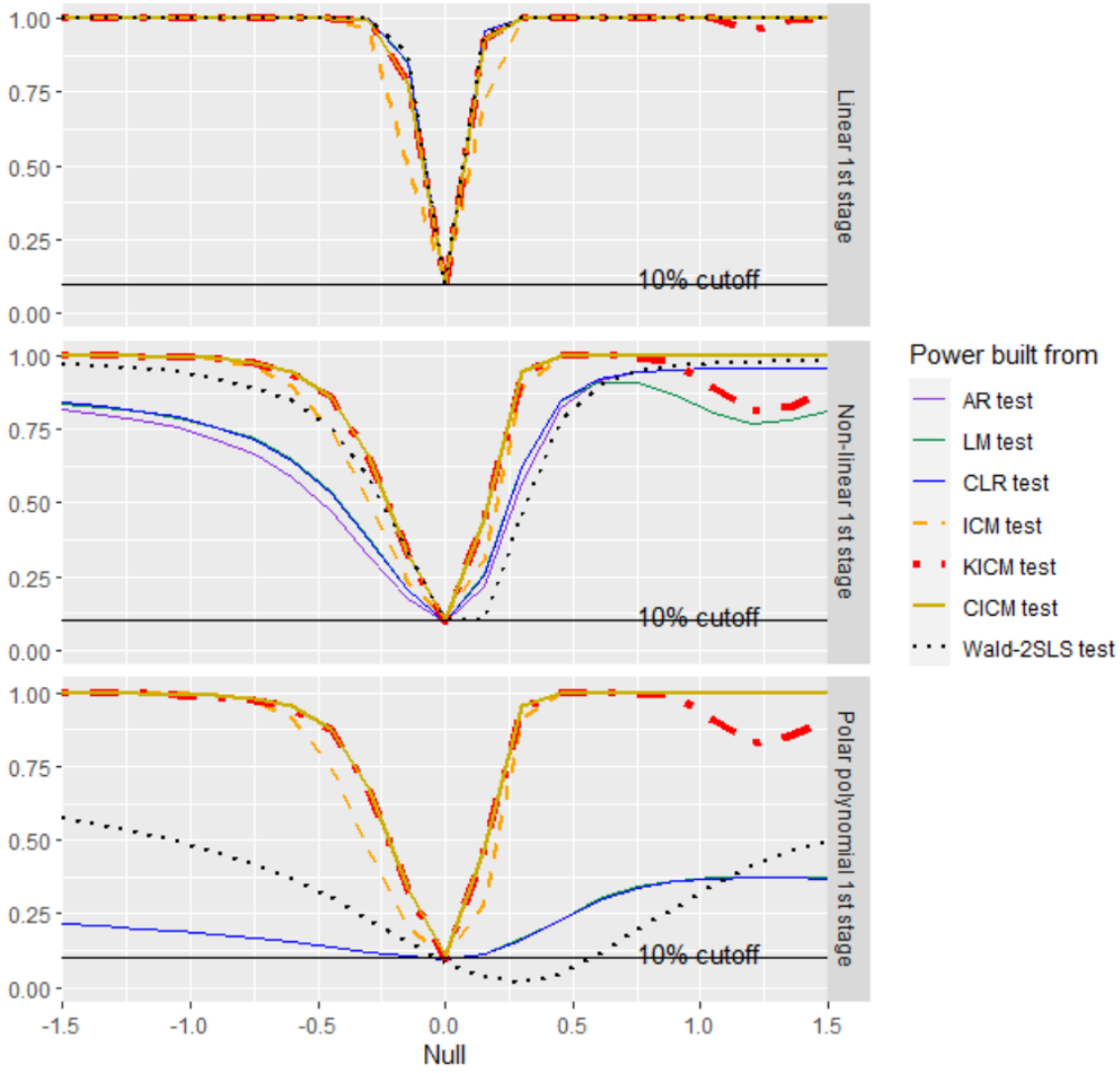

Figure 1: Power curves, strong instruments, homoskedastic Data

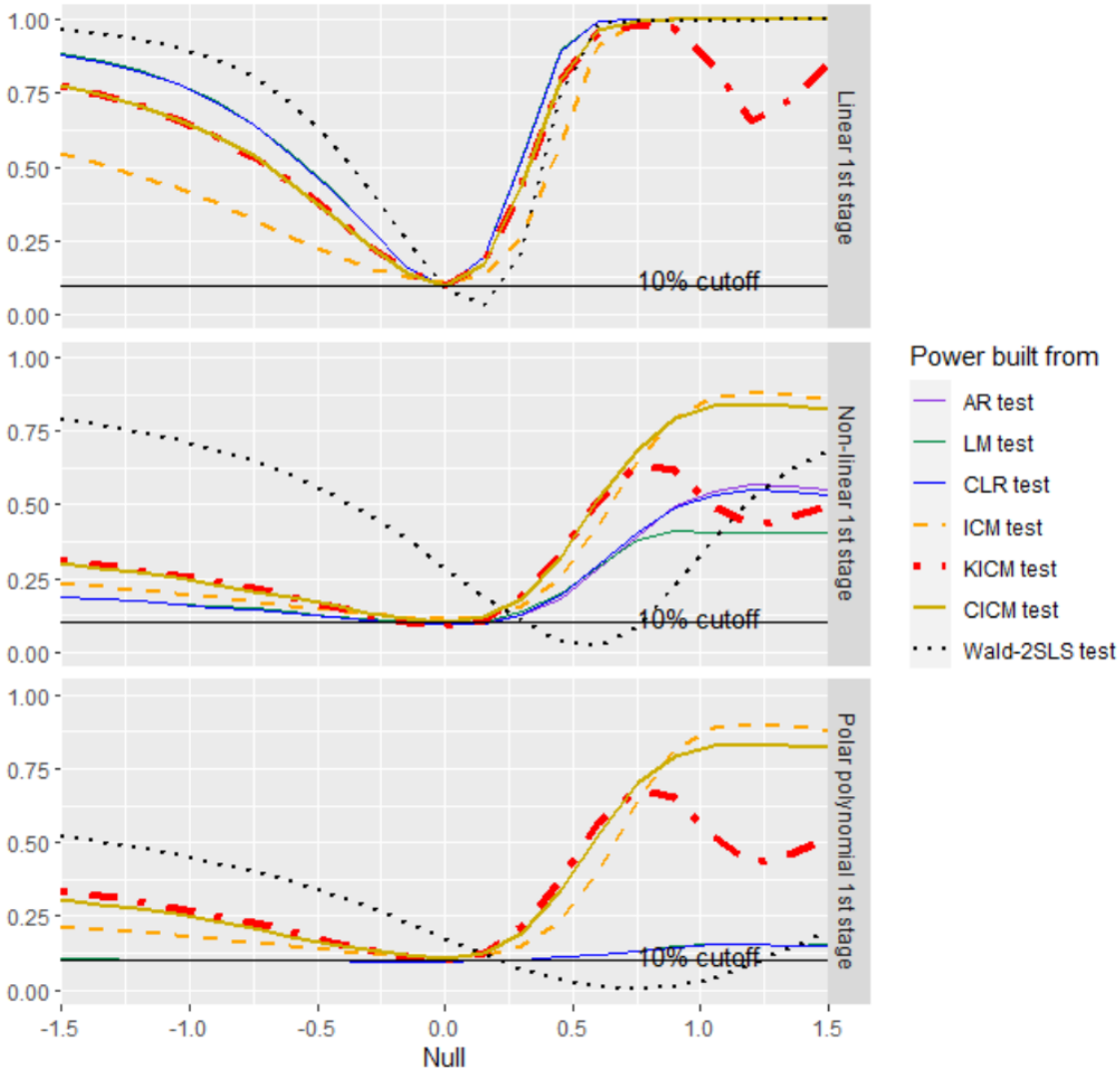

Figure 2: Power curves, semi-strong instruments, homoskedastic data

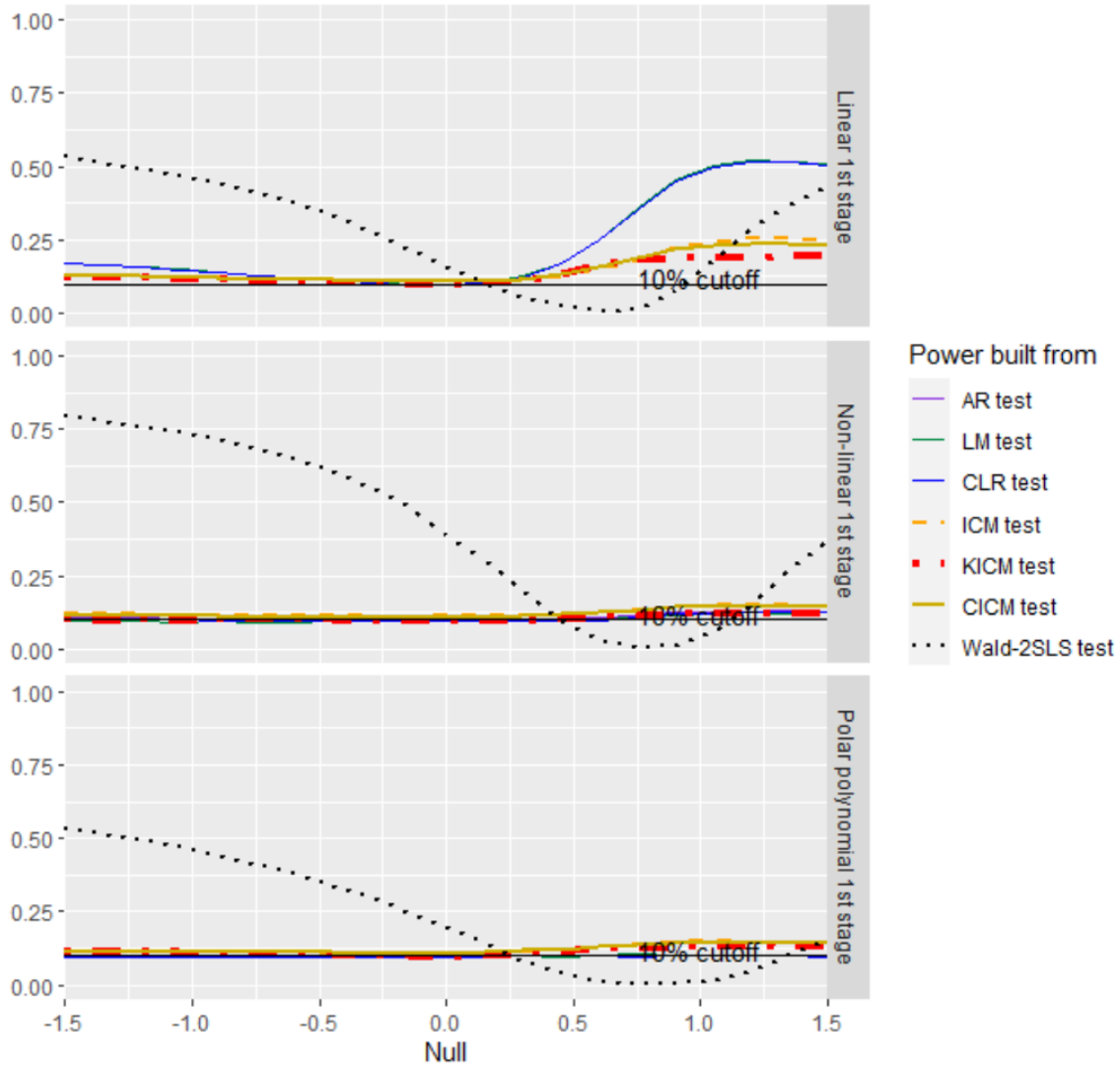

Figure 3: Power curves, weak instruments, homoskedastic data

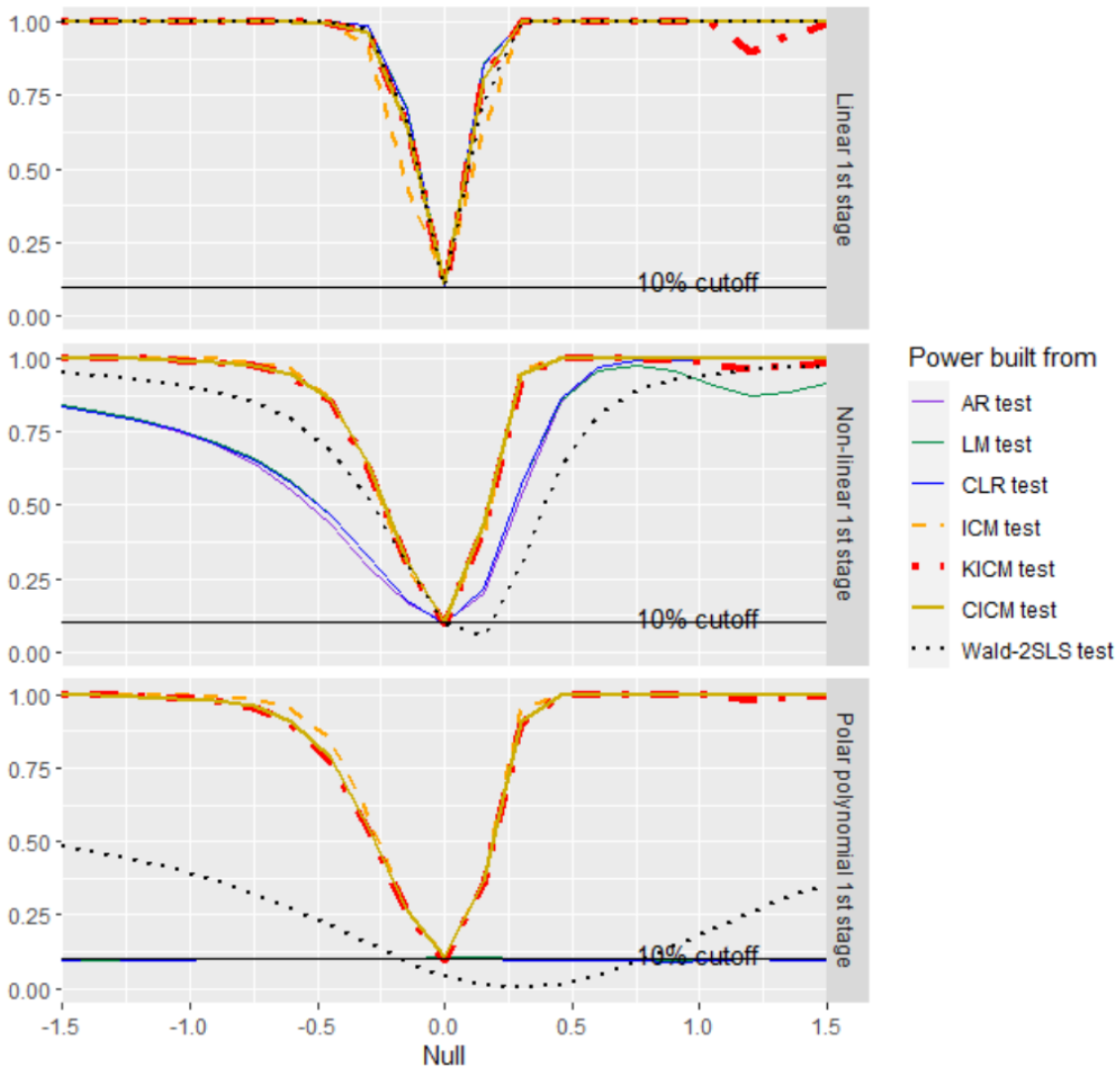

Figure 4: Power curves, strong instruments, heteroskedastic data

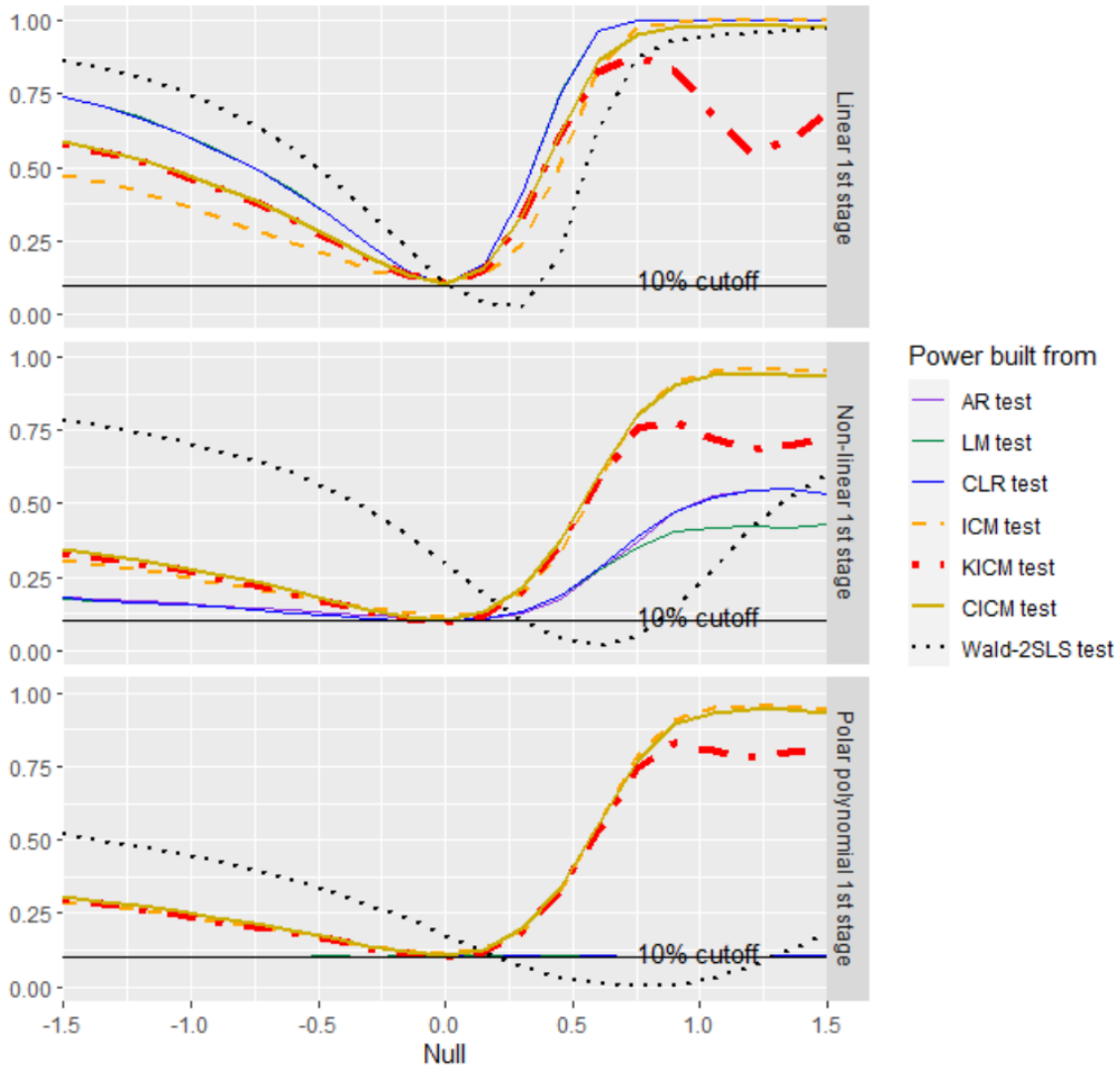

Figure 5: Power curves, semi-strong instruments, heteroskedastic data

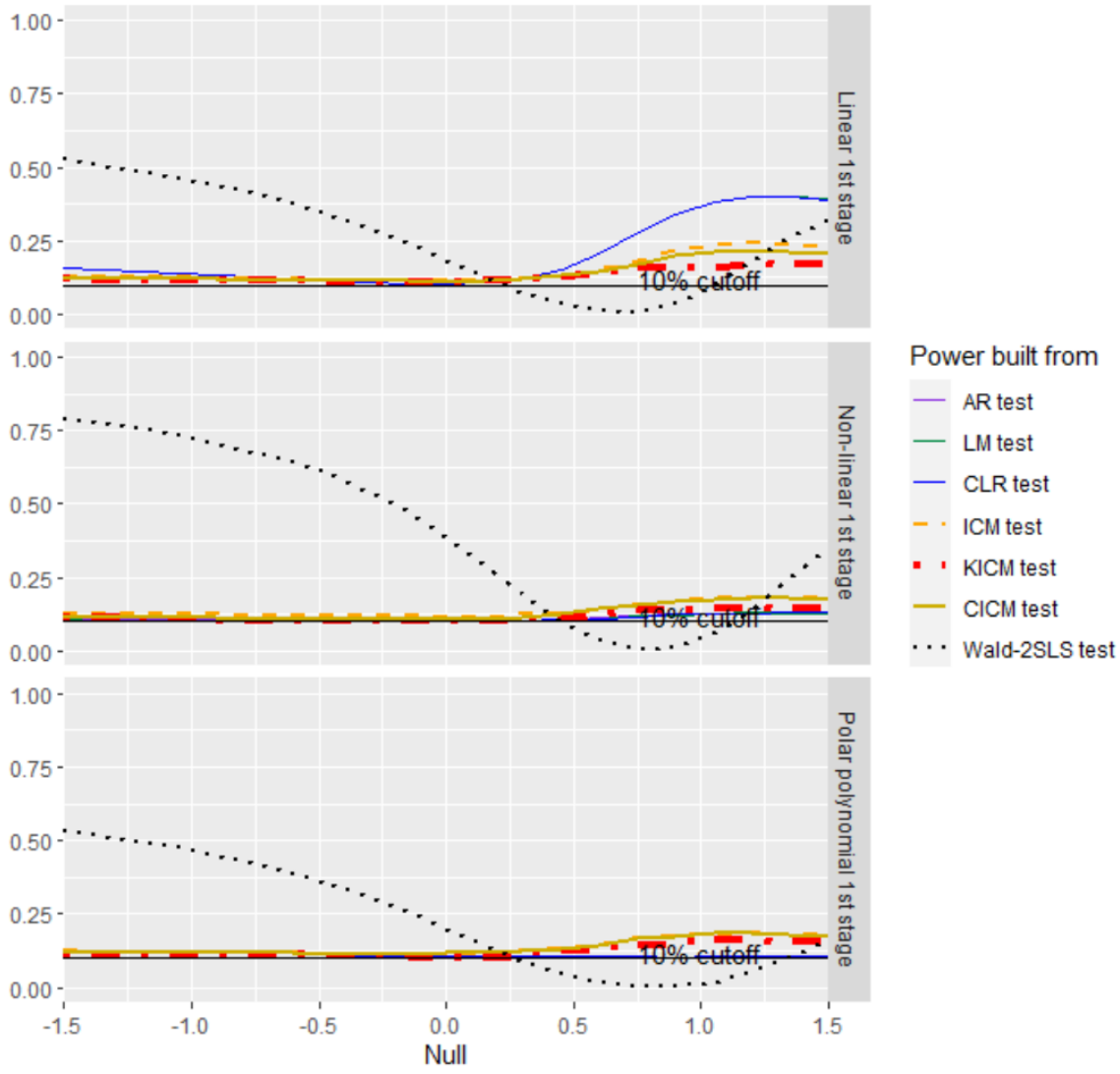

Figure 6: Power curves, weak instruments, heteroskedastic data

E.1.3 Average p-value curves

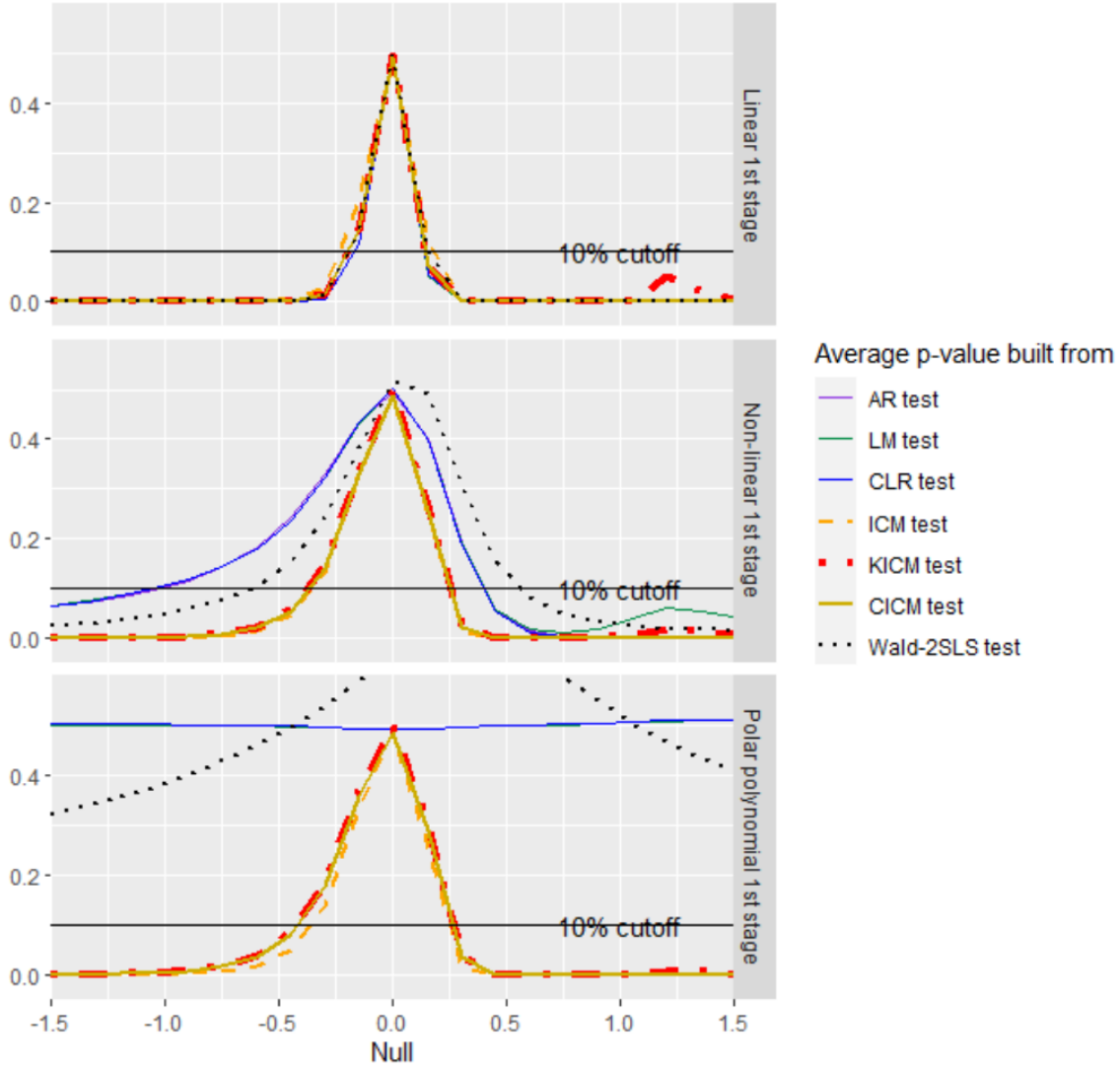

Figure 7: P-value curves, strong instruments, heteroskedastic data

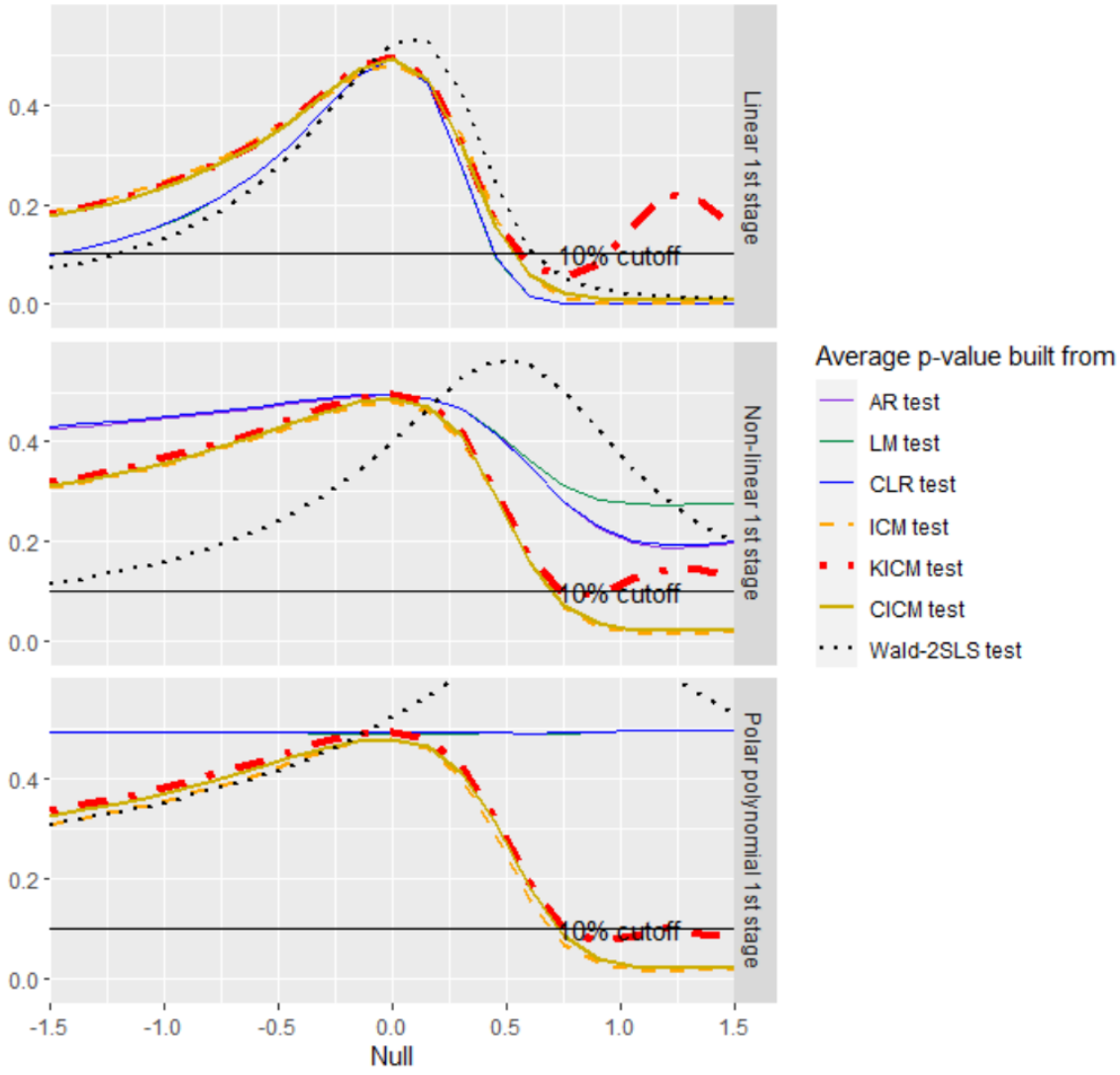

Figure 8: P-value curves, semi-strong instruments, heteroskedastic data

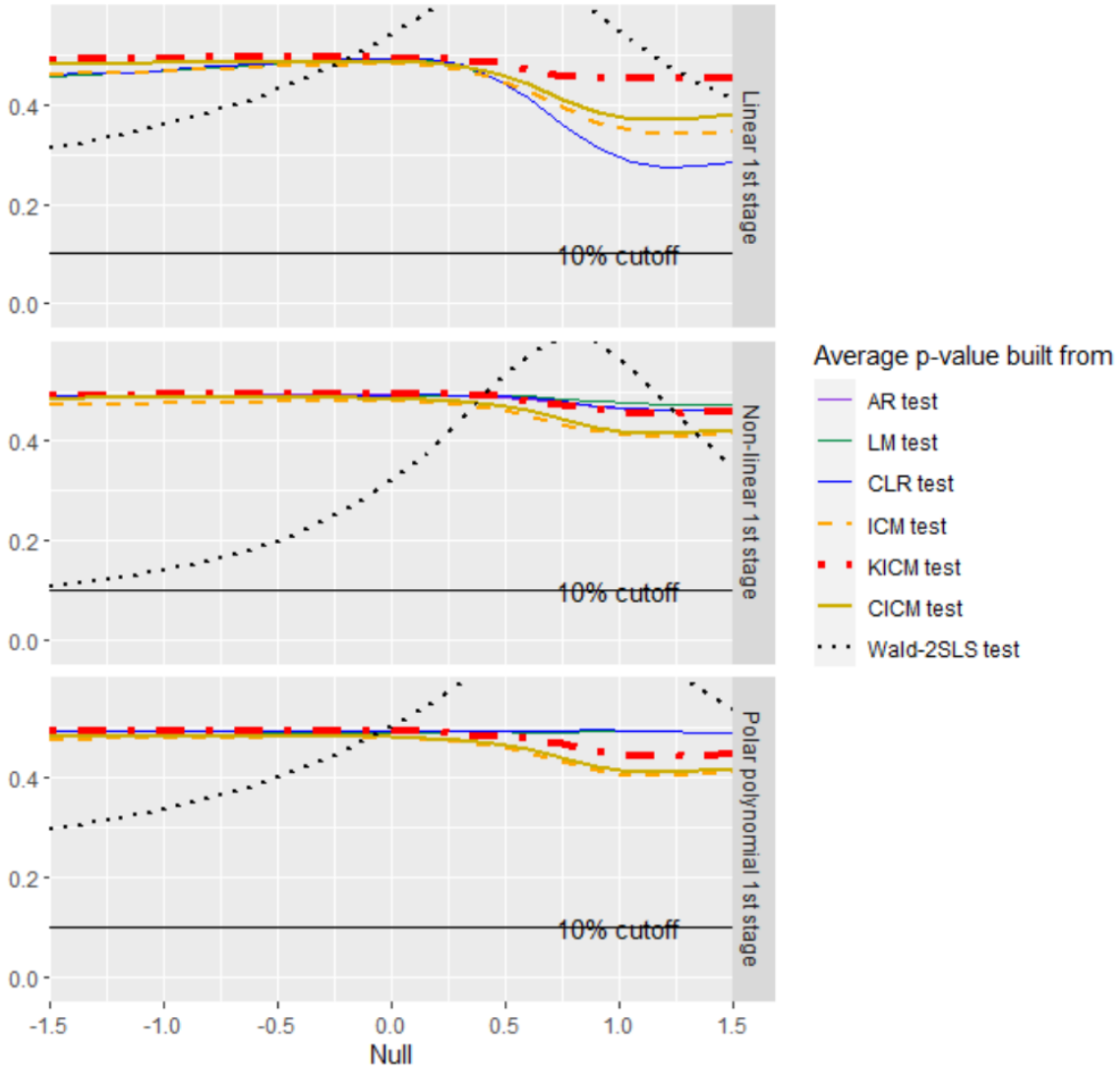

Figure 9: P-value curves, weak instruments, heteroskedastic data

E.2 Application

Specification	(1)	(2)	(3)	(4)
KICM	[0.009;0.132]	$\mathbb{R}$	$\mathbb{R}$	[0.138;0.240]
AR	[0.009;0.162]	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$
LM	[0.047;0.118]	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$
CLR	[0.042;0.110]	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$
OLS	0.080	0.080	0.072	0.070
	[0.080;0.081]	[0.080;0.081]	[0.071;0.072]	[0.070;0.071]
2SLS	0.077	0.131	0.106	0.101
	[0.052;0.102]	[0.076;0.186]	[0.050;0.163]	[0.046;0.156]
LIML	0.076	0.255	0.300	0.282
	[0.047;0.105]	[0.118;0.393]	[0.068;0.531]	[0.059;0.505]
FULLER	0.76	0.238	0.256	0.241
	[0.047;0.105]	[0.100;0.375]	[0.024;0.487]	[0.018;0.464]
Age and age square	-	Yes	Yes	Yes
Additional covariates	-	-	Yes	Yes
Region residence FE	-	-	-	Yes
First stage F test statistic	4.68	1.08	0.99	1.03

Table 5: 90% confidence intervals for returns to education, cohort 20-29

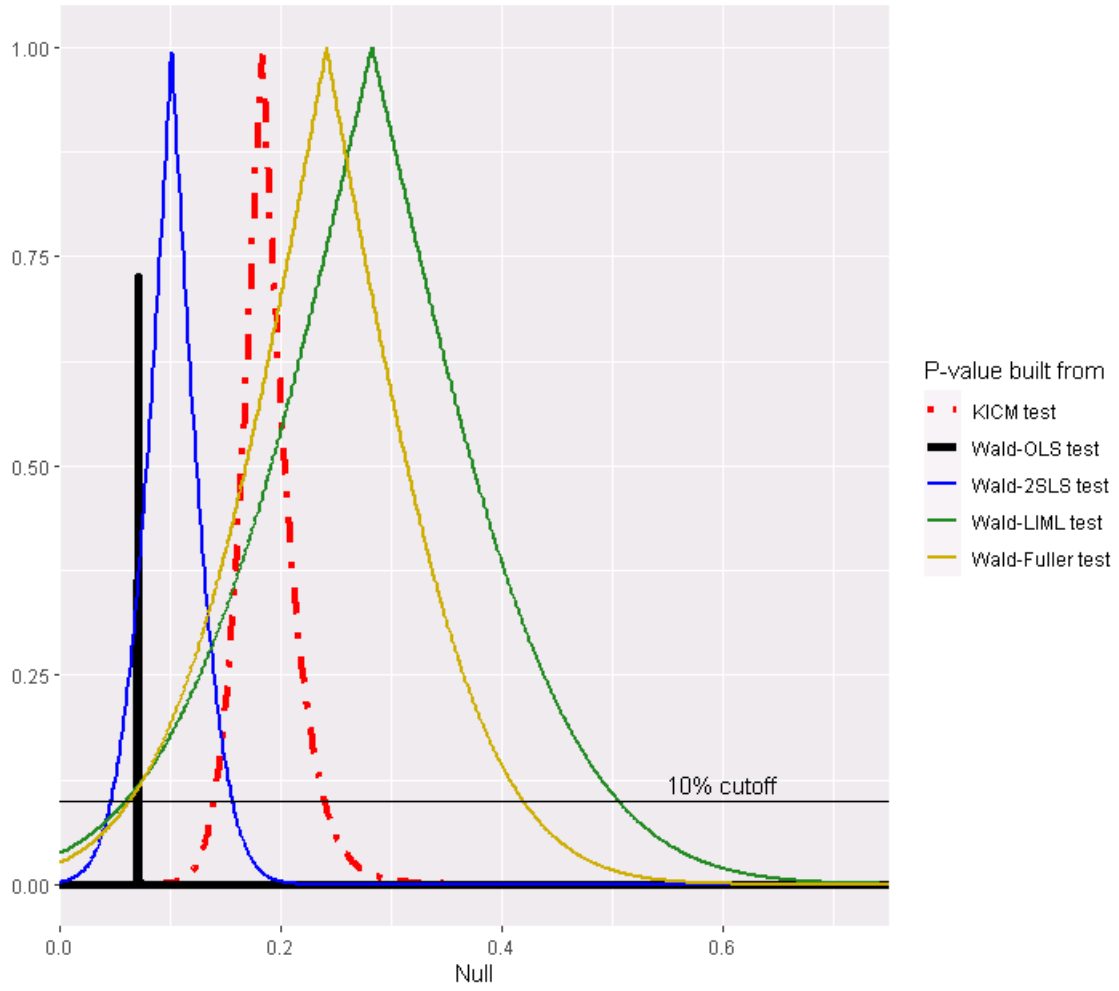

Figure 10: P-value curve of return to education, all covariates and fixed effects setting, cohort 20-29