UNIVERSITÀ
DI SIENA
1240

QUADERNI DEL DIPARTIMENTO

DI ECONOMIA POLITICA E STATISTICA

Federico Crudu
Giovanni Mellace
Zsolt Sándor

Inference in instrumental variables models with heteroskedasticity and many instruments

n. 821 – Febbraio 2020
versione aggiornata del Quaderno n.761 novembre 2017

Inference in instrumental variables models with heteroskedasticity and many instruments*

Federico Crudu†
Università di Siena and CRENoS

Giovanni Mellace‡
University of Southern Denmark

Zsolt Sándor§
Sapientia Hungarian University of Transylvania

November 2019

Abstract

This paper proposes novel inference procedures for instrumental variable models in the presence of many, potentially weak instruments that are robust to the presence of heteroskedasticity. First, we provide an Anderson-Rubin-type test for the entire parameter vector that is valid under assumptions weaker than previously proposed Anderson-Rubin-type tests. Second, we consider the case of testing a subset of parameters under the assumption that a consistent estimator for the parameters not under test exists. We show that under the null the proposed statistics have Gaussian limiting distributions and derive alternative chi square approximations. An extensive simulation study shows the competitive finite sample properties in terms of size and power of our procedures. Finally, we provide an empirical application using college proximity instruments to estimate the returns to education.

Key words: Instrumental variables, heteroskedasticity, many instruments, jackknife, inference.
JEL classification: C12, C13, C23.

*We are grateful to Stanislav Anatolyev, Samuele Centorrino, and Neil Davies for valuable help. F. Crudu thanks financial support from the Chilean government through CONICYT's grant FONDECYT Iniciacion n. 11140433. Z. Sándor thanks financial support from grant PN-II-ID-PCE-2012-4-0066 of the Romanian Ministry of National Education, CNCS-UEFISCDI.

†Department of Economics and Statistics, Piazza San Francesco 7/8, 53100 Siena, Italy, federico.crudu@unisi.it

‡Department of Business and Economics, Campusvej 55, 5230 Odense M, Denmark, giome@sam.sdu.dk

§Department of Business Sciences, Piata Libertăii 1, 530104 Miercurea Ciuc, Romania, sandorzsolt@cs.sapientia.ro

1 Introduction

The performance of test statistics based on instrumental variable (IV) models crucially depends on the quality and quantity of said IVs. In the presence of weak instruments, standard test statistics tend to deliver unreliable results. It is also well known that the number of instruments used in the construction of such tests plays a key role (see, e.g., Kleibergen, 2002, and references therein).

The Anderson-Rubin test (Anderson and Rubin, 1949, henceforth AR) is one of the most widely used statistics in the context of IV. Notoriously, this approach has the advantage of being robust to the presence of weak instruments. However, when the number of instruments grows larger than the number of parameters, the performance of the AR test starts deteriorating (e.g., Anatolyev and Gospodinov, 2011).The presence of heteroskedasticity may exacerbate the problem.

Over the years a number of improvements on the basic formulation of the AR test have been introduced (see, e.g., Staiger and Stock, 1997; Wang and Zivot, 1998; Zivot *et al.*, 1998; Kleibergen, 2002; Stock *et al.*, 2002; Andrews *et al.*, 2006; Moreira, 2009; Andrews *et al.*, 2019). However, those tests do not consider the framework when the number of instruments grows with the sample size.

Anatolyev and Gospodinov (2011) study the limiting behavior of the AR and Sargan statistic under Bekker's many instruments framework (Bekker, 1994). Under conditional homoskedasticity they find that their test statistics are asymptotically normal and that the resulting limiting distributions depend on $\lambda = \lim_{n\to\infty} k/n$, $0 < \lambda < 1$ where k is the number of instruments and n is the sample size. Since the tests may display some size distortion when λ is close to either zero or one, the authors propose a suitable chi square approximation. Donald *et al.* (2003) and Andrews and Stock (2007) obtain similar results where the instruments are allowed to grow at slower rates.

Probably, the paper closest to ours is that of Chao *et al.* (2014), where the authors propose an overidentification test for many (weak) instruments and heteroskedasticity that

exploits the properties of the jackknife IV estimator (see Hausman *et al.*, 2012). The framework in Chao *et al.* (2014) is sufficiently general to include the Bekker's many instruments case, the many weak instruments case of Chao and Swanson (2005) and instruments that are either weak or strong. Furthermore, the ratio k/n is bounded and the number of instruments cannot grow faster than the square of the concentration parameter.

Newey and Windmeijer (2009) study generalized empirical likelihood and generalized method of moments methods in a model with moment restrictions and show that the tests of Guggenberger and Smith (2005) and Kleibergen (2005) have canonical chi square limits even when the number of instruments goes to infinity. However, the rate of growth of the instruments is slower than that in Chao *et al.* (2014). In a recent paper Bun *et al.* (2018) propose a general version of the AR test based on an Edgeworth expansion argument both for k fixed and, in the homoskedastic linear model case, for $k \to \infty$. We are not aware of any other studies that generalize the AR test to the case of many instrumental variables and heteroskedasticity.

The objective of this paper is to construct test statistics for the parameter vector of a linear IV model in the presence of many, potentially weak instruments and heteroskedasticity. The starting point of our work is the paper by Bekker and Crudu (2015). The analysis is closely related to the papers by Hausman *et al.* (2012) and Chao *et al.* (2014).

First of all, we show that the many-instrument results in Anatolyev and Gospodinov (2011) are no longer valid under heteroskedasticity. Then, we propose a test statistic to test null hypotheses on the full vector of parameters associated to both endogenous and exogenous variables and a test statistic to test null hypotheses on a subset of the parameters of the model (see e.g., Guggenberger *et al.*, 2012). In the latter case we assume the existence of a plug-in estimator that is consistent under the null hypothesis. We also allow for heteroskedasticity of unknown form. In this sense, our test statistics may be seen as generalizations of the AR test. The first statistic we introduce refers to the whole parameter vector and, under the null, does not explicitly depend on the convergence properties of the concentration parameter. On the other hand, the second test statistic is

built to test a subset of the parameter vector and relies on a consistent plug-in estimator. In this case, when the plug-in is an IV estimator, the concentration parameter plays a role in the limiting properties of the test. The assumptions on the concentration parameter match those in Bekker and Crudu (2015) and are rather similar to those in Hausman *et al.* (2012). To the best of our knowledge there is no other test on only a subset of the parameter vector in IV models with many, potentially weak instruments that allow for the presence of heteroskedasticity (for the fixed instruments model see e.g., Guggenberger *et al.*, 2012,0).

The plan of the paper is as follows. Section 2 introduces the model, Section 3 describes the test statistics, the main asymptotic results and the associated assumptions. Section 4 and Section 5 contain the simulation results and an empirical application using the college proximity instruments of Card (1995), respectively. Section 6 concludes the paper. Proofs, auxiliary results and figures are relegated to the Appendix and some additional material is available in an online Supplemental Appendix.

2 The IV model

Let us consider the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Pi} + \boldsymbol{U} \tag{2}$$

where $\boldsymbol{y}$ is a vector of dimension n and $\boldsymbol{X}$ is a $n \times g$ matrix. Throughout the paper it is assumed that the $n \times k$ matrix of instruments $\boldsymbol{Z}$ is nonstochastic and $\mathrm{E}[\boldsymbol{X}] = \boldsymbol{Z}\boldsymbol{\Pi}$, where the components of $\boldsymbol{\Pi}$ are allowed to vary with the sample size n. Such assumptions are made for convenience and may be generalized.[1] The rows of the disturbance couple $(\boldsymbol{\varepsilon}, \boldsymbol{U})$,

[1]We may, for example, consider $\boldsymbol{Z}$ to be stochastic and in this case $\mathrm{E}[\boldsymbol{X}]$ should be interpreted as a conditional expectation with respect to $\boldsymbol{Z}$. The linearity of $\mathrm{E}[\boldsymbol{X}]$ may also be relaxed as suggested in, e.g., Bekker (1994) and Chao *et al.* (2014).

say $(\varepsilon_i, \boldsymbol{U}_i')$ $i = 1, \ldots, n$, are independent with zero mean and covariance matrices

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} \sigma_i^2 & \boldsymbol{\sigma}_{i12} \\ \boldsymbol{\sigma}_{i21} & \boldsymbol{\Sigma}_{i22} \end{pmatrix} \tag{3}$$

while the covariance matrix of the rows $(y_i, \boldsymbol{X}_i')$ are

$$\boldsymbol{\Omega}_i = \begin{pmatrix} 1 & \boldsymbol{\beta}' \\ \mathbf{0} & \boldsymbol{I}_g \end{pmatrix} \boldsymbol{\Sigma}_i \begin{pmatrix} 1 & \mathbf{0} \\ \boldsymbol{\beta} & \boldsymbol{I}_g \end{pmatrix}. \tag{4}$$

3 Asymptotic results

In this section we introduce a set of assumptions that are used to prove our asymptotic results. Furthermore, we generalize a result due to Anatolyev and Gospodinov (2011) to the heteroskedastic case and we introduce our main results. In addition to that, we compare our assumptions with those introduced in other related papers and we comment on the behavior of the proposed tests when some critical assumptions are violated.

The assumptions we use are similar to those in Bekker and Crudu (2015). Additional assumptions are included to generalize some results due to Anatolyev and Gospodinov (2011). In what follows it is understood that the generic positive constant c_u may be different in different situations.

Assumption 1. *The generic diagonal element P_{ii} of the projection matrix $\boldsymbol{P} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$ satisfies $\max_i P_{ii} \leq 1 - 1/c_u$, with $1 < c_u < \infty$. In addition, $k \to \infty$ as $n \to \infty$.*

Assumption 2. *The variances satisfy $\sigma_i^2 \geq \underline{\sigma}^2$ with $0 < \underline{\sigma}^2 < \infty$, for any i.*

Assumption 3. $\mathrm{E}[\varepsilon_i^4] \leq c_u$ *and* $\mathrm{E}\left[\|\boldsymbol{U}_i\|^4\right] \leq c_u$ *with $0 < c_u < \infty$, for any i.*

Assumption 1 is a technical condition on the projection matrix $\boldsymbol{P}$. It requires the main diagonal elements of $\boldsymbol{P}$ to be bounded away from 1. This assumption is rather standard in the literature (e.g., Hausman *et al.*, 2012; Bekker and Crudu, 2015) and is strictly

weaker than the so called asymptotic balanced design (see Anatolyev, 2018) imposed, for example, in Anatolyev and Gospodinov (2011) and Bun *et al.* (2018) according to which all the diagonal elements of the projection matrix converge to the same constant. The assumption $k \to \infty$ as $n \to \infty$ formalizes the many instruments idea in a way that is known as Bekker asymptotics. Assumption 2 and Assumption 3 are standard regularity conditions; the former bounds variances of the disturbances away from zero, while the latter bounds the fourth moments of the errors.

3.1 The AR test under heteroskedasticity

In this section we study the limiting distribution of the AR test statistics in the presence of heteroskedasticity. In addition, our derivation implies that the test statistics we propose in Section 3.2 are also valid under homoskedasticity.

The AR statistic is a popular choice to test a null hypothesis defined as $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$. The statistic is defined as

$$AR = (n-k)\,\frac{\boldsymbol{\varepsilon}_0'\boldsymbol{P}\boldsymbol{\varepsilon}_0}{\boldsymbol{\varepsilon}_0'\left(\boldsymbol{I}_n-\boldsymbol{P}\right)\boldsymbol{\varepsilon}_0} \tag{5}$$

and, under certain assumptions, it is asymptotically chi square distributed with k degrees of freedom. In the many instruments context and in the presence of homoskedasticity, the behavior of the AR test has been studied by Andrews and Stock (2007) and Anatolyev and Gospodinov (2011), among others. The following result generalizes the results in Lemma 1 of Anatolyev and Gospodinov (2011) to the heteroskedastic case. Let us define $\overline{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2$ and $W_n = \frac{2}{k}\sum_{i\neq j} P_{ij}^2\sigma_i^2\sigma_j^2$.

Proposition 1. *Suppose that Assumption 3 is satisfied,* $\lambda = \lim_{n\to\infty}\frac{k}{n} < 1$ *exists and* $\frac{1}{k}\sum_{i=1}^n(P_{ii}-\frac{k}{n})^2 \to 0$, $\frac{1}{\sqrt{k}}\sum_{i=1}^n(P_{ii}-\frac{k}{n})\sigma_i{}^2 \to 0$ *hold.*[2] *In addition, assume that* $\lim_{n\to\infty}\overline{\sigma}_n^2 =$

[2]The assumption $\frac{1}{\sqrt{k}}\sum_{i=1}^n(P_{ii}-\frac{k}{n})\sigma_i{}^2 \to 0$ is needed here in order for the expected value of the statistic to converge to 0 because this does not always hold. In Example B.1 in the Supplemental Appendix we provide an instance when this property is violated in the context of indicator instruments (Bekker and Van der Ploeg, 2005).

σ_0^2 *and* $\lim_{n\to\infty} W_n = W_0$ *exist. Then the statistic* $AR_{AG} = \sqrt{k}\left(\frac{AR}{k} - 1\right)$ *proposed by Anatolyev and Gospodinov (2011) has the limit* [3]

$$AR_{AG} \xrightarrow{d} \mathcal{N}\left(0, \frac{W_0}{\sigma_0^4 (1-\lambda)^2}\right).$$

Remark 1. *The asymptotic distribution result in Proposition 1 has two important implications. First, the asymptotic size of this test is*

$$\begin{aligned}\Pr\left(AR_{AG} > \Phi^{-1}(1-\alpha)\right) &= \Pr\left(\frac{\sigma_0^2(1-\lambda)}{\sqrt{W_0}} AR_{AG} < \frac{\sigma_0^2(1-\lambda)}{\sqrt{W_0}} \Phi^{-1}(\alpha)\right) \\ &\to \Phi\left(\frac{\sigma_0^2(1-\lambda)}{\sqrt{W_0}} \Phi^{-1}(\alpha)\right).\end{aligned}$$

Second, the test statistic T_1 *proposed in Section 3.2 has broader applicability than that proposed by Anatolyev and Gospodinov even under homoskedasticity. This is because its asymptotic distribution requires the assumption that the main diagonal elements* P_{ii}, $i = 1, ..., n$, *of the projection matrix* $\boldsymbol{P}$ *should be bounded away from* 1. *The test statistic proposed by Anatolyev and Gospodinov (2011) requires the stronger assumption that the main diagonal elements of* $\boldsymbol{P}$ *converge to* λ. *This difference in the assumptions comes from the fact that the former test statistic does not involve the diagonal elements of* $\boldsymbol{P}$ *while the latter statistic does. The following example clarifies this concept.*

Example 1. *Consider indicator instruments with unequal group sizes (Bekker and Van der Ploeg, 2005). Anatolyev and Yaskov (2017, Section 5.1) show that in this case the main diagonal elements of* $\boldsymbol{P}$ *do not converge to* λ. *In the Supplemental Appendix we show that under homoskedasticity the convergence in distribution* $\sqrt{k}\left(\frac{AR}{k} - 1\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{2}{1-\lambda}\right)$ *is violated.*

[3] We note that under homoskedasticity

$$\frac{\overline{\sigma}_n^2}{\sqrt{W_n}} \to \frac{\sigma^2}{\sqrt{2(1-\lambda)\sigma^2}} = \frac{1}{\sqrt{2(1-\lambda)}}, \text{ so } \sqrt{k}\left(\frac{AR}{k} - 1\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{2}{1-\lambda}\right),$$

which is exactly as in Lemma 1 of Anatolyev and Gospodinov (2011).

3.2 Inference with heteroskedasticity and many instruments

In this section we present the main results. First, we present our test on the entire parameter vector. Then, we consider the more challenging case where we test the null on a subset of the coefficients; in this case we assume that a consistent plug-in estimator exists for the parameters not under test. Furthermore, we study our tests when the number of instruments is fixed. Finally, we briefly discuss the behavior of our "subset" test for some commonly encountered specific plug-in estimators and in some pathological situations.

The test statistics proposed in this paper are related to the symmetric jackknife instrumental variable estimator (SJIVE) proposed by Bekker and Crudu (2015). The SJIVE estimates consistently, in the many (weak) instruments sense, the parameter vector $\boldsymbol{\beta}$ and it is defined as

$$\widehat{\boldsymbol{\beta}}_{SJIVE} = \arg\min_{\boldsymbol{\beta}} Q_{SJIVE}(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} \frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{C}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{B}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})} \tag{6}$$

and, given the projection matrix $\boldsymbol{P}$ and the diagonal matrix $\boldsymbol{D}$ containing the diagonal elements of $\boldsymbol{P}$,

$$\boldsymbol{C} = \boldsymbol{A} - \boldsymbol{B}, \qquad \boldsymbol{A} = \boldsymbol{P} + \boldsymbol{\Delta}, \qquad \boldsymbol{B} = (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{D}(\boldsymbol{I}_n - \boldsymbol{D})^{-1}(\boldsymbol{I}_n - \boldsymbol{P}),$$
$$\boldsymbol{\Delta} = \boldsymbol{P}\boldsymbol{D}(\boldsymbol{I}_n - \boldsymbol{D})^{-1}\boldsymbol{P} - \frac{1}{2}\boldsymbol{P}\boldsymbol{D}(\boldsymbol{I}_n - \boldsymbol{D})^{-1} - \frac{1}{2}\boldsymbol{D}(\boldsymbol{I}_n - \boldsymbol{D})^{-1}\boldsymbol{P}.$$

Consider now testing the null hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, where $\boldsymbol{\beta}$ is the true parameter vector.

The test statistic we propose is based on the numerator of the objective function in equation (6), namely,

$$Q(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{C}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{7}$$

and it is defined as

$$T_1 = \frac{1}{\sqrt{k}} \frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0)' \boldsymbol{C} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0)}{\sqrt{\widehat{V}(\boldsymbol{\beta}_0)}}, \quad \widehat{V}(\boldsymbol{\beta}_0) = \frac{2}{k} \boldsymbol{\varepsilon}_0^{(2)'} \boldsymbol{C}^{(2)} \boldsymbol{\varepsilon}_0^{(2)} \tag{8}$$

where $\boldsymbol{\varepsilon}_0 = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0$ and the superscript "$^{(2)}$" indicates the elementwise product of two conformable matrices or vectors. The following theorem provides the asymptotic distribution of the T_1 test statistic. [4]

Theorem 1. *If Assumptions 1, 2, 3 are satisfied, then under* $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ *we have* $T_1 \xrightarrow{d} \mathcal{N}(0, 1)$.

Let us now consider a nominal level α and let z_α be the α-th quantile of the normal distribution. Then, the null hypothesis is rejected if $T_1 \geq z_{1-\alpha}$.

Sometimes one is interested only in performing inference on a subset of parameters. In particular, we would like to test the coefficients associated to the endogenous variables. Let us now define the parameter vector as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ and suppose we want to test the following null hypothesis

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_{10} \tag{9}$$

in the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

where the dimensions of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are $n \times g_1$ and $n \times g_2$ respectively with $g = g_1 + g_2$. Let $\boldsymbol{y}_0 = \boldsymbol{y} - \boldsymbol{X}_1\boldsymbol{\beta}_1$ so that under the null hypothesis the model becomes

$$\boldsymbol{y}_0 = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}.$$

[4] We could apply the same type of analysis by replacing $\boldsymbol{C}$ with $\boldsymbol{P} - \boldsymbol{D}$ as in Chao *et al.* (2014). We do not pursue that avenue since, as suggested in Bekker and Crudu (2015), $\boldsymbol{C}$ allows us to retain the whole signal matrix.

Accordingly, the reduced form model corresponding to $\boldsymbol{X}_2$ is

$$\boldsymbol{X}_2 = \boldsymbol{Z}\boldsymbol{\Pi}_2 + \boldsymbol{U}_2,$$

where $\boldsymbol{\Pi}_2$ and $\boldsymbol{U}_2$ both have g_2 columns. Further, let

$$\boldsymbol{H} = \boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Pi} = \begin{pmatrix} \boldsymbol{H}_{11} & \boldsymbol{H}_{12} \\ \boldsymbol{H}_{12}' & \boldsymbol{H}_{22} \end{pmatrix},$$

denote the signal matrix and let $\boldsymbol{H}_{22} = \boldsymbol{\Pi}_2'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Pi}_2$, which has dimension $g_2 \times g_2$.

We assume that a consistent estimator for $\boldsymbol{\beta}_2$, say $\widetilde{\boldsymbol{\beta}}_2$, exists. If the variables associated to $\boldsymbol{\beta}_2$ are exogenous, the OLS estimator is a valid candidate. However, if this is not the case, we need a suitable IV estimator. Under the null, a consistent estimator is, for example, the SJIVE. For the null hypothesis $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_{10}$ consider $\widetilde{\boldsymbol{\beta}} = \left(\boldsymbol{\beta}_{10}', \widetilde{\boldsymbol{\beta}}_2'\right)'$, $\widetilde{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}$ and let the modified test statistic, denoted as T_2, be

$$T_2 = \frac{1}{\sqrt{k}} \frac{\widetilde{\boldsymbol{\varepsilon}}'\boldsymbol{C}\widetilde{\boldsymbol{\varepsilon}}}{\sqrt{\widehat{V}(\widetilde{\boldsymbol{\beta}})}}, \quad \text{where } \widehat{V}(\widetilde{\boldsymbol{\beta}}) = \frac{2}{k}\widetilde{\boldsymbol{\varepsilon}}^{(2)\prime}\boldsymbol{C}^{(2)}\widetilde{\boldsymbol{\varepsilon}}^{(2)}. \tag{10}$$

Let now $r_{\min} = \lambda_{\min}(\boldsymbol{H}_{22})$ and $r_{\max} = \lambda_{\max}(\boldsymbol{H}_{22})$ be the smallest eigenvalue and the largest eigenvalue of $\boldsymbol{H}_{22}$, respectively. Moreover, let us define a generic constant κ such that $0 \leq \kappa < \infty$.

Assumption 4. *$k/r_{\min} \to \kappa$, $r_{\max}/k \to \kappa$ when $n \to \infty$.*

Assumption 5. *$r_{\max}/k \to \kappa$, $r_{\min}/k \to 0$, $\sqrt{k}/r_{\min} \to 0$ when $n \to \infty$.*

We have two remarks on these assumptions. First, Assumptions 4 and 5 are used in conjunction with Assumption 1 (specifically, $k \to \infty$ as $n \to \infty$), and therefore, either of them implies that $r_{\min} \to \infty$ and $r_{\max} \to \infty$ as $n \to \infty$. Second, Assumptions 4 and 5 regulate the convergence of the plug-in IV estimator. When the growth rates of $r_{\min}$ and $r_{\max}$ are the same, we are either in the many instruments framework of Bekker (1994) or

in the many weak instruments framework of Chao and Swanson (2005). As in Chao *et al.* (2014), the growth rates of $r_{\min}$ and $r_{\max}$ are allowed to vary.

The following theorem provides the asymptotic distribution of the T_2 test.

Theorem 2. *If Assumptions 1, 2, 3 and either 4 (many strong instruments case) or 5 (many weak instruments case) are satisfied, then* $T_2 \xrightarrow{d} \mathcal{N}(0,1)$.

Analogously to the T_1 case, the null hypothesis is rejected if $T_2 \geq z_{1-\alpha}$.

It is important to derive the limiting distribution of our tests in case the number of instruments does not grow with the sample size. The following theorem provides the limiting distribution of T_1 and T_2 under the assumption that k is fixed and the error couple $(\boldsymbol{\varepsilon}, \boldsymbol{U})$ is homoskedastic.

Theorem 3. *Let the disturbance couple* $(\boldsymbol{\varepsilon}, \boldsymbol{U})$ *be zero mean and homoskedastic and let Assumption 3 hold. Furthermore, assume (i)* k *fixed and* $n \to \infty$*, (ii) as* $n \to \infty$*,* $\frac{\boldsymbol{Z}'\boldsymbol{Z}}{n} \to \boldsymbol{\Sigma}_{ZZ}$ *a full rank non stochastic matrix, (iii) as* $n \to \infty$*,* $\frac{\boldsymbol{X}'\boldsymbol{Z}}{n} \to_p \boldsymbol{\Sigma}_{XZ}$ *a non stochastic matrix with* $rank(\boldsymbol{\Sigma}_{XZ}) = g$*, (iv) as* $n \to \infty$*,* $\frac{\boldsymbol{Z}'\boldsymbol{\varepsilon}}{\sqrt{n}} \to_d \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{\Sigma}_{ZZ})$*. Then* $\sqrt{2k}T_1 + k \to_d \chi^2_k$*. If* $\widetilde{\boldsymbol{\beta}}_2$ *is the two-stage least squares estimator, then* $\sqrt{2k}T_2 + k \to_d \chi^2_{k-g_2}$.

While the Gaussian approximation may work well in finite samples, it does not allow us to control for the number of instruments. This, as stressed in Anatolyev and Gospodinov (2011), may be an important issue. The following corollary shows how to obtain a chi square approximation for T_1 and T_2.

Corollary 1. *If the assumptions of Theorem 1 hold true, then (i)* $\sqrt{k}T_1 + k \to_d \chi^2_k$*. If the assumptions of Theorem 2 hold true, then (ii)* $\sqrt{k}T_2 + k \to_d \chi^2_k$*, (iii)* $\sqrt{k}T_2 + k \to_d \chi^2_{k-g_2}$ *or (iv)* $\sqrt{k-g_2}T_2 + k - g_2 \to_d \chi^2_{k-g_2}$.

Corollary 1 shows that there are different possible chi square approximations for T_2. While approximation (iii) seems to be a natural candidate, also because it matches the result in Theorem 3, it may not deliver the best results in finite samples. We expect, for

example, that, when k is small, approximations (ii) and (iv) enjoy better finite sample properties.[5]

The convergence properties of T_1 are determined by the behavior of the diagonal elements of $\boldsymbol{P}$ and by the properties of the disturbances. The T_2 test also depends on the properties of the plug-in estimator of the parameters not under test. When the OLS estimator $\widetilde{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{y}_0$ is a consistent plug-in estimator, that is, $\boldsymbol{X}_2$ is exogenous, it is easy to show that T_2 converges to a standard normal basically under the same assumptions as those imposed in Theorem 1. No further assumptions on the strength of the instruments need to be imposed (see Theorem B.1 in the Supplemental Appendix for a formal treatment). In the Supplemental Appendix (see Theorem B.2) we derive the limiting distribution of T_2 for the case when $\boldsymbol{X}_2$ is endogenous and the JIV1 estimator is used as plug-in. The convenient expression of the JIV1 estimator allows us to explain why underrejection of the null hypothesis occurs in most cases (see Remark B.1 in the Supplemental Appendix) and to better link the weak instrument assumption $\sqrt{k}/r_{\min} \to 0$ to the asymptotic distribution of T_2 (see Remark B.2 in the Supplemental Appendix). Our derivation suggests that the assumption $\sqrt{k}/r_{\min} \to 0$ is likely to be necessary for the asymptotic standard normality of the statistic T_2 (see Remark B.2). In Section 4.2 below we further discuss the behavior of T_2 in relation with the limiting behavior of $\sqrt{k}/r_{\min}$. Specifically, we illustrate that for relatively large $\sqrt{k}/r_{\min}$ the histogram of T_2 differs substantially from the standard normal density. Finally, in the case when the plug-in estimator converges slowly to the true value we find that the density of T_2 is shifted to the right causing the test to overreject (see Example B.2 in the Supplemental Appendix).

3.3 Comparison with other tests

In this section we compare our set of assumptions with those used in some closely related papers. Some papers provide a broad range of results and a certain degree of variation in the specification of the assumptions. Therefore, for ease of presentation, some assumptions

[5]See Section C in the Supplemental Appendix for some Monte Carlo evidence.

considered here are stronger than necessary (e.g., Newey and Windmeijer, 2009). In Table 1, we report the different assumptions imposed on the rate of convergence of the number of instruments and the concentration parameter, and whether or not they are robust to heteroskedasticity. Moreover, we distinguish between test statistics that consider null hypotheses on the full set of parameters or on a subset. We also report whether they allow for instruments with unbalanced design. Finally, we only consider the case where the model contains one endogenous variable. Thus, $r_{\min} = r_{\max} = r$ and $r = \boldsymbol{\pi}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\pi}$ where r is the scalar version of the signal matrix $\boldsymbol{H}$ and is proportional to the concentration parameter.

There is a certain degree of heterogeneity in the type of assumptions that we show in Table 1. For example, Anatolyev and Gospodinov (2011) and Bun *et al.* (2018) use Bekker's framework. Our assumptions, on the other hand, are more in line with those in Chao *et al.* (2014), with the difference that in our case r is bounded by k, while in Chao *et al.* (2014) it is bounded by n.[6] The assumptions in Andrews and Stock (2007) and Newey and Windmeijer (2009) are to some extent similar to ours but generally their rates tend to be slower. We also notice that only Newey and Windmeijer (2009) consider AR-type tests that are robust to heteroskedasticity. Moreover, no test other than T_2 seems to explicitly consider the subset null hypothesis presented in Equation (9).

4 Monte Carlo simulations

In this Section we study the finite sample properties of the T_1 and T_2 tests in terms of size and power (see Figures 1 and 2 for the results on size and Figures 3 to 6 for the results on power).[7] Further Monte Carlo results may be found in the Supplemental Appendix. We make inference on the full parameter vector and on the sole parameter associated to the endogenous variable. The proposed tests are compared to the version of the AR test proposed by Anatolyev and Gospodinov (2011), denoted as AR_{AG}, and the AR test

[6] They assume either $\frac{n}{r} \to \kappa$ or $\frac{n}{r} \to 0$ and $\frac{\sqrt{k}}{r} \to 0$.

[7] The size properties of T_1 and T_2 are investigated by means of PP-plots as described in Davidson and MacKinnon (1998).

Table 1: Comparison of assumptions in the many instruments framework.

	Subset	k/n	r	Heteroskedasticity	Unbalanced instruments
Anatolyev and Gospodinov (2011)	No	$\frac{k}{n} \to \lambda, 0 < \lambda < 1$	$\frac{r}{n} \to \kappa$, $\kappa \in (0, \infty)$	No	No
Andrews and Stock (2007)	No	$\frac{k^3}{n} \to 0$	$\frac{r}{k^\zeta} \to \kappa_\zeta$, $\kappa_\zeta \in [0, \infty)$ $\zeta \in (0, \infty)$	No	Yes
Bun *et al.* (2018)	No	$\frac{k}{n} \to \lambda, 0 < \lambda < 1$	–	No	No
Newey and Windmeijer (2009)	No	$\frac{k^2}{n} \to 0$ or $\frac{k^3}{n} \to 0$	$\frac{n}{r} \to \kappa$ or $\frac{r}{n} \to 0$ $\frac{k}{r}$ bounded	Yes	Yes
T_1	No	$\frac{k}{n}$ bounded	–	Yes	Yes
T_2	Yes	$\frac{k}{n}$ bounded	$\frac{k}{r} \to \kappa$ or $\frac{r}{k} \to 0$, $\frac{\sqrt{k}}{r} \to 0$	Yes	Yes

Notes: For simplicity we refer to the single endogenous variable case where $r = \boldsymbol{\pi}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\pi} = r_{\min} = r_{\max}$ and restrict ourselves to tests that use $k \to \infty$. Bun *et al.* (2018) also propose tests for the fixed k case that are robust to heteroskedasticity. Andrews and Stock (2007) and Newey and Windmeijer (2009) impose different set of assumptions depending on the problem considered and the ones reported here might be stronger than necessary in some cases.

introduced in Bun *et al.* (2018) and defined as

$$\widetilde{AR}_{df} = n\widehat{g}(\boldsymbol{\beta})'\widetilde{\boldsymbol{\Omega}}_{df}(\boldsymbol{\beta})^{-1}\widehat{g}(\boldsymbol{\beta}) \tag{11}$$

where $\widetilde{\boldsymbol{\Omega}}_{df}(\boldsymbol{\beta}) = \frac{n}{n-k}\widetilde{\boldsymbol{\Omega}}(\boldsymbol{\beta})$, $\widetilde{\boldsymbol{\Omega}}(\boldsymbol{\beta}) = \widehat{\boldsymbol{\Omega}}(\boldsymbol{\beta}) - \widehat{g}(\boldsymbol{\beta})\widehat{g}(\boldsymbol{\beta})'$ and $\widehat{\boldsymbol{\Omega}}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{\beta})\boldsymbol{g}(\boldsymbol{\beta})'$. The moment condition model is defined as $\boldsymbol{g}_i(\boldsymbol{\beta}) = \boldsymbol{Z}_i(y_i - \boldsymbol{X}_i'\boldsymbol{\beta})$ and $\widehat{g}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}_i(\boldsymbol{\beta})$. In the case of T_1, T_2 and AR_{AG}, we use the corresponding chi square asymptotic distribution.[8]

This comparison is interesting for a number of reasons. First, we get a clearer idea on how much we gain by using our tests in a heteroskedastic context. Second, we get some important insights on how the considered test statistics work in the extreme cases where $\frac{k}{n} \approx 0$ and $\frac{k}{n} \approx 1$. A priori, we may expect the AR_{AG} to work well under homoskedasticity and for moderately large values of $\frac{k}{n}$, while it is probable that $\widetilde{AR}_{df}$ performs well also in the heteroskedastic case.

Furthermore, we introduce a two parameter model with two endogenous regressors; this model is used to study the role played by the boundary condition $\sqrt{k}/r_{\min} \to 0$ and by the

[8]Due to the results in Figure C.2 in the Supplementary Appendix, for T_2 we use approximation (ii) in Corollary 1.

plug-in estimator in determining the behavior of T_2 (see Figure 7 and Figure 8 in Appendix B).

4.1 Data generating processes

Let us consider the Monte Carlo set up of Hausman *et al.* (2012). One of the features of this experiment is that the sum of the diagonal elements of $\boldsymbol{P}$ does not converge to $\lambda = \lim \frac{k}{n}$, as shown in Anatolyev and Yaskov (2017). The DGP is given by

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{\iota}\gamma + \boldsymbol{x}\beta + \boldsymbol{\varepsilon} \\ \boldsymbol{x} &= \boldsymbol{z}\pi + \boldsymbol{v} \end{aligned} \tag{12}$$

where $\gamma = \beta = 1$, while $\pi = 0.1$ in the analysis of size and $\pi \in \{0.1, 1\}$ in the analysis of power. The sample size is $n = 800$, $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0},\ \boldsymbol{I}_n)$ and independently $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}, 0.1^2 \times \boldsymbol{I}_n)$. The disturbances vector $\boldsymbol{\varepsilon}$ is generated as

$$\boldsymbol{\varepsilon} = \rho\boldsymbol{v} + \sqrt{\frac{1-\rho^2}{\phi^2 + \psi^4}}(\phi\boldsymbol{w}_1 + \psi\boldsymbol{w}_2), \tag{13}$$

where $\rho = 0.3$, $\psi = 0.86$ and conditional on $\boldsymbol{z}$, independent of $\boldsymbol{v}$, $\boldsymbol{w}_1 \sim \mathcal{N}(\boldsymbol{0},\ \mathrm{Diag}(\boldsymbol{z})^2)$ where $\mathrm{Diag}(\boldsymbol{z})$ is a diagonal matrix where the diagonal elements are the elements of $\boldsymbol{z}$ and $\boldsymbol{w}_2 \sim \mathcal{N}(\boldsymbol{0},\ \psi^2\boldsymbol{I}_n)$. Notice that, $\phi = 1.38072$ implies heteroskedasticity, while $\phi = 0$ corresponds to the homoskedastic case. The instrument matrix $\boldsymbol{Z}$ is given by matrices with rows $(1, z_i, z_i^2, z_i^3, z_i^4)$ and $(1, z_i, z_i^2, z_i^3, z_i^4, z_i b_{1i}, \ldots, z_i b_{\ell i})$, $\ell = 95, 695$, where, independent of other random variables, the elements $b_{1i}, \ldots, b_{\ell i}$ are i.i.d. Bernoulli distributed with $p = 1/2$.[9] We consider also two rather extreme situations: $k = 2$ and $k = 700$. We replicate our experiments 5000 times. When using the T_1 test and the T_2 test we consider $H_0 : (\gamma, \beta)' = (1, 1)'$ and $H_0 : \beta = 1$ respectively.[10]

[9] The same set of instruments is used throughout the various repetitions.

[10] We computed results also for $\ell = 5, 15, 35, 55, 75$ and we noticed that the p-value curves would converge from the p-value curve associated to $k = 5$ to the p-value curve with $k = 100$. This result replicates in all cases, including the power curves.

The following DGP is used to explore the properties of the T_2 test when the boundary condition $\sqrt{k}/r_{\min} \to 0$ is violated and when the plug-in estimator is inconsistent. Let us consider the following model

$$\begin{aligned} &\boldsymbol{y} = \boldsymbol{x}\beta + \boldsymbol{w}\gamma + \boldsymbol{\varepsilon} \\ &\boldsymbol{x} = \boldsymbol{Z}\boldsymbol{\pi}_x + \boldsymbol{u}_x, \;\; \boldsymbol{w} = \boldsymbol{Z}\boldsymbol{\pi}_w + \boldsymbol{u}_w. \end{aligned} \tag{14}$$

Let us now suppose we want to test the null $H_0 : \beta = \beta_0$, define $\boldsymbol{\eta}_i = (\varepsilon_i, u_{xi}, u_{wi})'$ and assume that

$$\boldsymbol{\eta}_i \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho & \xi \\ \rho & 1 & 0 \\ \xi & 0 & 1 \end{pmatrix}\right), \;\; \boldsymbol{Z}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k), i = 1, \ldots, n. \tag{15}$$

We consider two cases.

1. In the first case we consider $n = 400$. In addition to that, we set $\rho = 0.2$, $\xi = 0.3$, $\boldsymbol{\pi}_x = \boldsymbol{\pi}_w = \sqrt{\frac{R^2}{k(1-R^2)}}\boldsymbol{\iota}_k$, $k = n/2$ and R^2 is chosen in such a way that $\frac{\sqrt{k}}{n\boldsymbol{\pi}_w'\boldsymbol{\pi}_w} = 0.1$ and $\frac{\sqrt{k}}{n\boldsymbol{\pi}_w'\boldsymbol{\pi}_w} = 1$.[11] Finally, as a plug-in estimator we use the JIV1 estimator. The number of replications is 5000.

2. The sample size is set to $n = 400$. Moreover, $\rho = 0.2$, $\xi \in \{0, 0.1, 0.2, 0.3\}$, $k \in \{2, 20, 200\}$, $\boldsymbol{\pi}_x = \boldsymbol{\pi}_w = \sqrt{\frac{R^2}{k(1-R^2)}}\boldsymbol{\iota}_k$ with $R^2 = 0.2$. Finally, as a plug-in estimator we use the OLS estimator. The number of replications is 5000.

4.2 Simulation results

We first provide some interpretation of the simulations by separately analyzing the results on size and power. Then we discuss the behavior of T_2 when an inconsistent plug-in is

[11] The condition $\frac{\sqrt{k}}{n\boldsymbol{\pi}_w'\boldsymbol{\pi}_w} = 1$ replicates the idea that the boundary condition $\sqrt{k}/r_{\min} \to 0$ is violated. We did run simulations also for $\frac{\sqrt{k}}{n\boldsymbol{\pi}_w'\boldsymbol{\pi}_w} = 10$ and $n = 100, 200$ finding similar results.

used.

Size. Analyzing Figure 1 and Figure 2 we observe that, in general, T_1 and T_2 work well in all the considered cases.[12] The $\widetilde{AR}_{df}$ test, on the other hand, works well for most of the cases but tends overreject when k is large. Finally, as expected, the AR_{AG} test overrejects for any value of k and its performance deteriorates as k increases.

Power. The power properties of the various test statistics display some interesting patterns. When $k = 2, 5$, the T_1 and the T_2 tests along with the $\widetilde{AR}_{df}$ test are able to discriminate among alternatives (Figure 3 to Figure 6 panels (a) and (b)). To some extent the same could be said about the case where $k = 100$ (Figure 3 to Figure 6 panel (c)). Finally, when $k = 700$, $\pi = 0.1$, the T_1 and the T_2 tests are unable to discriminate among alternatives. More precisely, no test statistic among those considered seems to work well in this case. However, when $\pi = 1$, the T_1 and T_2 tests tend to outperform their competitors (Figure 3 to Figure 6 panel (d)). It is interesting to notice that the properties of T_1 and T_2 are affected by a trade off between size and power with respect to k: as k grows the empirical size approaches the nominal size, but the power curves tend to get wider. This may be a problem when the instruments are weak as the tests may eventually have no power for k large. When the instruments are stronger the effect of such a trade-off is less severe and our tests work well even in the extreme case with $k = 700$.

Over/underrejection. The comparison of the histograms and QQ–plots in Figure 7 displays how the T_2 test behaves when the boundary condition $\sqrt{k}/r_{\min} \to 0$ is violated. In particular, we notice that the (empirical) density tends to be more leptokurtic with respect to its asymptotic counterpart. This feature induces the test to underreject. On the other hand, the plots in Figure 8 show the behavior of the T_2 statistic when a slow plug-in estimator is used. We notice that the use of OLS instead of a more appropriate IV estimator causes T_2 to overreject. In particular, T_2 overrejects more as ξ increases. Furthermore, we notice that the overrejection tendency is mitigated by the increased number of instruments.

[12] It is worth noticing that, in general, for the hypothesis $H_0 : \beta_1 = \beta_{10}$ all the tests tend to underreject for small values of k.

Violation of the boundary condition. Figure 7 illustrates the behavior of the T_2 test with a small $\sqrt{k}/r_{\min}$ (Figure 7(a)) and a large $\sqrt{k}/r_{\min}$ (Figure 7(b)). In the latter case, the histogram of T_2 differs substantially from the one of a standard normal density, suggesting that $\sqrt{k}/r_{\min} \to 0$ is important for the asymptotic normality of our test.

5 Empirical application

In this section we apply our methods to the data from the National Longitudinal Survey of Young Men (NLSYM) used by Card (1995) to estimate the returns to education. The data set includes 3010 observations and 35 variables.[13]

We consider two different models to estimate the returns of education. Both models assume that the log of wages (*wage*) is a linear function of education measured in years of schooling (*school*) and a set of exogenous variables $\boldsymbol{x}$, namely

$$\log(wage_i) = \beta school_i + \boldsymbol{x}_i'\boldsymbol{\gamma} + \varepsilon_i.$$

Similar to Kleibergen (2004), $\boldsymbol{x}$ includes a constant and binary variables for race, residence in a metropolitan area, and residence in the south of the United States as well as IQ test score. As experience is measured simply as $age - school - 6$ in this data, we do not use it as a control variable in our models. [14] For the instruments, following once again Kleibergen (2004), in our first specifications we use age and age square and two variables that indicate college proximity. The exogeneity of the college proximity instruments is somewhat questionable for several reasons. For example, areas with a high prevalence of people with high unobserved ability may be more likely to have a college nearby. Card (1995) argues that including other observable characteristics, as we do, should mitigate this issue. However, we cannot completely exclude the potential endogeneity of our instruments.

[13] The data are from the `R` package `ivmodel` of Jiang *et al.* (2016).

[14] Another reason not to control for experience when estimating returns of education, at least in this data, is that experience is mechanically an outcome of education and it is therefore a bad control as discussed for example in Angrist and Pischke (2008).

In our second specification, we generate additional excluded instruments by interacting age, age squared, and the two college proximity variables with the geographical indicators and race. In the first specification, the instrument set includes four variables, while in the second it includes fourteen variables.

It is very likely that the variance of the error depends on the exogenous variables which motivates the use of our T_2 test for inference. For example, it appears very plausible that the conditional variance of the unobservables driving wages differ by college proximity, location as well as race. We run our T_2 statistic, using both the chi square (T_2) and Gaussian (T_2^{gauss}) approximations, the AR_{AG} statistic of Anatolyev and Gospodinov (2011), the $\widetilde{AR}_{df}$ statistics of Bun *et al.* (2018), and the standard AR statistic to test 301 equidistant values in the interval $[0, 3]$ for the coefficient of education β. The results for the model with four instruments are reported in Figure 9. With only 4 excluded instruments all the tests give very similar results, in particular they are not able to reject values in the (approximate) interval $[.5, 1.5]$, at the 5% significance level, and in the (approximate) interval $[.5, .1]$, at the 10% significance level. The implied effects of education on wages are much higher than the one found in Card's study who, however, includes experience (which is arguably a "bad control") in his model. However, the large effects implied by our models are in line with the one found in Imbens and Rubin (1997).

The results with fourteen instruments are reported in Figure 10. Probably due to the presence of heteroskedasticity, adding instruments deteriorates the performances of the AR, AR_{AG}, and $\widehat{AR}_{df}$ tests, which reject every single value of β at the 10% significance level. On the other hand, increasing the number of instruments does not seem to have a big impact on our T_2 and T_2^{gauss} tests. These results are in line with what we find in our simulation study where, with strong heteroskedasticity, the performances of the AR_{AG} deteriorates when we increase the number of instruments and the $\widehat{AR}_{df}$ tends to overreject with many instruments.

6 Conclusion

This paper introduces two test statistics for the parameters of a linear model in the presence of endogeneity, heteroskedasticity and many, potentially weak, instruments. The tests are easy to build as they are based on the numerator of the SJIVE estimator proposed by Bekker and Crudu (2015). We prove that, after appropriate rescaling, the limiting distribution of the test statistics are standard normal. Moreover, simulation evidence shows that, in finite samples, the proposed tests generally outperform their competitors, such as the AR tests proposed in Anatolyev and Gospodinov (2011) and in Bun *et al.* (2018).

In our empirical application, the standard AR test and its modification by Anatolyev and Gospodinov (2011), probably due to the presence of heteroskedasticity, reject every single value chosen for the null when we increase the number of instruments from four to fourteen. On the other hand, our proposed statistic provides similar results independently of the number of instruments used.

The tests we propose can be applied broadly to any linear overidentified IV model and they are particularly appealing for the growing literature using genetic markers as instruments, see for example Von Hinke *et al.* (2016). In this literature, the number of instruments is potentially very large and the instruments are typically weak, a framework where our tests potentially outperform existing methods. Another potential field of application for our tests is the framework of Kang *et al.* (2016) and Windmeijer *et al.* (2017) where inference is carried out after a potentially large set of valid instruments is selected via LASSO.

Appendix

A Proofs of main results

This section contains the proofs of the main theorems and some auxiliary results. In what follows it is understood that $\boldsymbol{O}$ is a conformable matrix of zeros and that the abbreviations LLN, CLT and IID stand for law of large numbers, central limit theorem and independently and identically distributed respectively. In addition to that, $\sum_{i\neq j}$ is a double sum for $i,j=1,\ldots,n$ that excludes the same index elements and $\sum_{i,j,k,\ell}$ replaces the quadruple sum $\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{\ell=1}^{n}$. Triple sums are defined similarly.

Proof of Proposition 1. Under $H_0:\boldsymbol{\beta}=\boldsymbol{\beta}_0$ we have

$$\sqrt{k}\left(\frac{AR}{k}-1\right)=\frac{\frac{1}{\sqrt{k}}\left(\frac{n-k}{k}\boldsymbol{\varepsilon}'\boldsymbol{P}\boldsymbol{\varepsilon}-\boldsymbol{\varepsilon}'\left(\boldsymbol{I}-\boldsymbol{P}\right)\boldsymbol{\varepsilon}\right)}{\frac{1}{k}\boldsymbol{\varepsilon}'\left(\boldsymbol{I}-\boldsymbol{P}\right)\boldsymbol{\varepsilon}}=\frac{n}{k}\frac{\frac{1}{\sqrt{k}}\left(\boldsymbol{\varepsilon}'\boldsymbol{P}\boldsymbol{\varepsilon}-\frac{k}{n}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}\right)}{\frac{1}{k}\boldsymbol{\varepsilon}'\left(\boldsymbol{I}-\boldsymbol{P}\right)\boldsymbol{\varepsilon}}. \tag{16}$$

Note that

$$\frac{1}{\sqrt{k}}\left(\boldsymbol{\varepsilon}'\boldsymbol{P}\boldsymbol{\varepsilon}-\frac{k}{n}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}\right)=\frac{1}{\sqrt{k}}\sum_{i\neq j}P_{ij}\varepsilon_i\varepsilon_j+\frac{1}{\sqrt{k}}\sum_{i=1}^{n}\left(P_{ii}-\frac{k}{n}\right)\varepsilon_i^2\equiv E_1+E_2. \tag{17}$$

We can apply the CLT from (Chao *et al.*, 2012, Lemma A2) to the quadratic form

$$R=\sum_{i\neq j}P_{ij}\varepsilon_i\varepsilon_j$$

involved in E_1. We obtain that

$$\frac{R}{\sqrt{kW_n}}\xrightarrow{d}\mathcal{N}\left(0,1\right),$$

where

$$W_n=\frac{\mathrm{Var}[R]}{k}=\frac{2}{k}\sum_{i\neq j}P_{ij}^2\sigma_i^2\sigma_j^2$$

with the property that

$$\frac{1}{k}\operatorname{Var}[R] = \frac{2}{k}\sum_{i\neq j} P_{ij}^2\sigma_i^2\sigma_j^2 \geq \frac{2\underline{\sigma}^4}{k}\sum_{i\neq j} P_{ij}^2 \geq \frac{2\underline{\sigma}^4}{c_u},$$

(the latter inequality comes from (22)), which is bounded away from 0. Consequently, W_n is bounded between two positive numbers. We obtain that $E_1/\sqrt{W_n} \xrightarrow{d} \mathcal{N}(0,1)$. Regarding E_2, by the assumption $\frac{1}{\sqrt{k}}\sum_{i=1}^n (P_{ii} - \frac{k}{n})\sigma_i{}^2 \to 0$ we have

$$\mathrm{E}[E_2] = \frac{1}{\sqrt{k}}\sum_{i=1}^n \left(P_{ii} - \frac{k}{n}\right)\sigma_i^2 \to 0.$$

Further, by Assumption 3

$$\operatorname{Var}[E_2] = \frac{1}{k}\sum_i \left(P_{ii} - \frac{k}{n}\right)^2 \operatorname{Var}\left[\varepsilon_i^2\right] \leq \frac{c_u}{k}\sum_{i=1}^n \left(P_{ii} - \frac{k}{n}\right)^2.$$

Using the assumption $\frac{1}{k}\sum_{i=1}^n \left(P_{ii} - \frac{k}{n}\right)^2 \to 0$, we obtain that $\operatorname{Var}[E_2] = o(1)$. Then by Chebyshev's inequality $E_2 = o_p(1)$. Therefore,

$$\frac{E_1 + E_2}{\sqrt{W_n}} \xrightarrow{d} \mathcal{N}(0,1). \tag{18}$$

Regarding the denominator involved in (16) we observe that

$$\frac{1}{k}\boldsymbol{\varepsilon}'(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\varepsilon} = \frac{1}{k}\left(1 - \frac{k}{n}\right)\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} - \frac{1}{k}\boldsymbol{\varepsilon}'\left(\boldsymbol{P} - \frac{k}{n}\boldsymbol{I}\right)\boldsymbol{\varepsilon}.$$

The second term is just the expression from (17) divided by $\sqrt{k}$, that is,

$$\frac{1}{k}\boldsymbol{\varepsilon}'(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\varepsilon} = \frac{1}{k}\left(1 - \frac{k}{n}\right)\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} - \frac{1}{\sqrt{k}}(E_1 + E_2) = \frac{1}{k}\left(1 - \frac{k}{n}\right)\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} + O_p\left(\frac{1}{\sqrt{k}}\right).$$

Using Assumption 3 and the LLN, using the notation

$$\overline{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2$$

we have that

$$\frac{1}{n}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} - \overline{\sigma}_n^2 = O_p\left(\frac{1}{\sqrt{k}}\right). \tag{19}$$

Consequently,

$$\frac{1}{k}\boldsymbol{\varepsilon}'\left(\boldsymbol{I}-\boldsymbol{P}\right)\boldsymbol{\varepsilon} = \frac{n}{k}\left(1-\frac{k}{n}\right)\overline{\sigma}_n^2 + O_p\left(\frac{1}{\sqrt{k}}\right).$$

Now, from equation (16) and the fact that $\frac{n}{k}\left(1-\frac{k}{n}\right)\overline{\sigma}_n^2$ is bounded between two positive numbers, we have

$$\begin{aligned}\sqrt{k}\left(\frac{AR}{k}-1\right) &= \frac{n}{k}\frac{\frac{1}{\sqrt{k}}\left(\boldsymbol{\varepsilon}'\boldsymbol{P}\boldsymbol{\varepsilon}-\frac{k}{n}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}\right)}{\frac{n}{k}\left(1-\frac{k}{n}\right)\overline{\sigma}_n^2} + \frac{n}{k}\frac{\frac{1}{\sqrt{k}}\left(\boldsymbol{\varepsilon}'\boldsymbol{P}\boldsymbol{\varepsilon}-\frac{k}{n}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}\right)}{\frac{n}{k}\left(1-\frac{k}{n}\right)\overline{\sigma}_n^2}\left(\frac{\frac{n}{k}\left(1-\frac{k}{n}\right)\overline{\sigma}_n^2}{\frac{1}{k}\boldsymbol{\varepsilon}'\left(\boldsymbol{I}-\boldsymbol{P}\right)\boldsymbol{\varepsilon}}-1\right)\\ &= \frac{E_1+E_2}{\left(1-\frac{k}{n}\right)\overline{\sigma}_n^2} + o_p\left(1\right).\end{aligned}$$

Therefore, collecting the above results we obtain that

$$\left(1-\frac{k}{n}\right)\frac{\overline{\sigma}_n^2}{\sqrt{W_n}}\sqrt{k}\left(\frac{AR}{k}-1\right) = \frac{E_1+E_2}{\sqrt{W_n}} + o_p\left(1\right),$$

which by (18) implies that

$$\left(1-\frac{k}{n}\right)\frac{\overline{\sigma}_n^2}{\sqrt{W_n}}\sqrt{k}\left(\frac{AR}{k}-1\right) \xrightarrow{d} \mathcal{N}\left(0,1\right). \tag{20}$$

Since we assume that $\lim_{n\to\infty}\overline{\sigma}_n^2 = \sigma_0^2$ and $\lim_{n\to\infty} W_n = W_0$ exist, we obtain the result. □

In the proof of Theorem 1 we use the following CLT, which, as argued by Bekker and Crudu (2015, Appendix A.4) can be proved in a way similar to Lemma A2 from Chao *et al.*

(2012).

Lemma A.1. *Consider the quadratic form* $Q = \sum_{i \neq j} C_{ij}\varepsilon_i\varepsilon_j$, *where* C_{ij} *is the* (i,j) *element of matrix* $\boldsymbol{C}$ *that is symmetric and has zero main diagonal elements. Suppose that there is a matrix* $\boldsymbol{P}$ *that is symmetric, idempotent,* $P_{ii} \leq c_u < 1$, $|C_{ij}| \leq c_u |P_{ij}|$ *for any* $i \neq j$, *and* $rank(\boldsymbol{P}) = k$, *where* $k \to \infty$ *as* $n \to \infty$, *and the following properties hold:* (a) $\mathrm{E}[\varepsilon_i] = 0$ *and* $\varepsilon_1, ..., \varepsilon_n$ *are independent;* (b) $\mathrm{E}[\varepsilon_i^4] < \infty$; (c) $\frac{1}{k}\mathrm{Var}[Q] \geq c_u > 0$. *Then,*

$$\frac{Q}{\sqrt{\mathrm{Var}[Q]}} \xrightarrow{d} \mathcal{N}(0,1).$$

Lemma A.2. *Let* $\widehat{V}(\boldsymbol{\beta}_0) = \frac{2}{k}\boldsymbol{\varepsilon}_0^{(2)\prime}\boldsymbol{C}^{(2)}\boldsymbol{\varepsilon}_0^{(2)}$. *If Assumptions 1, 3 hold,* $\widehat{V}(\boldsymbol{\beta}_0) - V_n = O_p\left(\frac{1}{\sqrt{k}}\right)$; *consequently* $\widehat{V}(\boldsymbol{\beta}_0) - V_n \xrightarrow{p} 0$, *where*

$$V_n = \frac{2}{k}\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}^2\sigma_i^2\sigma_j^2.$$

(For a proof see Supplemental Appendix A.)

Proof of Theorem 1. Under the null hypothesis we have

$$\begin{aligned} \mathrm{E}[\boldsymbol{\varepsilon}_0'\boldsymbol{C}\boldsymbol{\varepsilon}_0] &= 0, \\ \mathrm{Var}[\boldsymbol{\varepsilon}_0'\boldsymbol{C}\boldsymbol{\varepsilon}_0] &= \mathrm{E}\left[(\boldsymbol{\varepsilon}_0'\boldsymbol{C}\boldsymbol{\varepsilon}_0)^2\right] = 2\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}^2\sigma_i^2\sigma_j^2 \equiv kV_n. \end{aligned}$$

We verify the conditions of the CLT stated in Lemma A.1 for $\boldsymbol{C}$ and $\boldsymbol{P}$ defined in Section 2. The properties of $\boldsymbol{C}$ and $\boldsymbol{P}$ hold by definition, Assumption 1 and the fact that $|C_{ij}| = \frac{|P_{ij}|}{2}\left(\frac{1}{1-P_{ii}} + \frac{1}{1-P_{jj}}\right) \leq c_u|P_{ij}|$ for any i, j, (see the proof of Lemma A.2 in the Supplemental Appendix).

Further, (a) is clearly satisfied; (b) is satisfied due to Assumption 3. Regarding (c) note that by Assumption 2

$$\frac{1}{k}\mathrm{Var}[Q] \equiv V_n = \frac{2}{k}\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}^2\sigma_i^2\sigma_j^2 \geq \frac{2\underline{\sigma}^4}{k}\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}^2,$$

where

$$\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}^2 = \sum_{i\neq j} \frac{P_{ij}^2}{4}\left(\frac{1}{1-P_{ii}}+\frac{1}{1-P_{jj}}\right)^2 \geq \sum_{i\neq j}\frac{P_{ij}^2}{4}(1+1)^2 = \sum_{i\neq j} P_{ij}^2$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{n} P_{ij}^2 - \sum_{i=1}^{n} P_{ii}^2 = tr(\boldsymbol{P}) - \sum_{i=1}^{n} P_{ii}^2 = k - \sum_{i=1}^{n} P_{ii}^2. \tag{21}$$

By Assumption 1

$$\sum_{i=1}^{n} P_{ii}^2 \leq \max P_{ii} \sum_{i=1}^{n} P_{ii} \leq (1-1/c_u)\, tr(\boldsymbol{P}) = (1-1/c_u)\, k. \tag{22}$$

So

$$\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}^2 \geq k/c_u,$$

therefore,

$$\frac{1}{k}\operatorname{Var}[Q] \geq \frac{2\sigma^4}{c_u},$$

which is bounded away from 0. In this case we can apply the CLT in Lemma A.1 and complete the proof by using Lemma A.2. □

For the proof of Theorem 2 we need the following results (for proofs see Supplemental Appendix A).

Lemma A.3. *Let* $\widehat{V}\left(\widetilde{\boldsymbol{\beta}}\right) = \frac{2}{k}\widetilde{\boldsymbol{\varepsilon}}^{(2)\prime}\boldsymbol{C}^{(2)}\widetilde{\boldsymbol{\varepsilon}}^{(2)}$. *If* $\widetilde{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ *and Assumptions 1, 3 hold, then* $\widehat{V}\left(\widetilde{\boldsymbol{\beta}}\right) - V_n \xrightarrow{p} 0$.

Lemma A.4. *Under Assumptions 1, 3*

1. $\mathrm{E}[\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2] = \boldsymbol{H}_{22},\ \operatorname{Var}[\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2] \leq c_u\boldsymbol{H}_{22} + c_u k\boldsymbol{I}_{g_2} + c_u r_{\max}\boldsymbol{I}_{g_2}$,

2. $\mathrm{E}[\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{\varepsilon}] = \boldsymbol{0},\ \operatorname{Var}[\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{\varepsilon}] \leq c_u\boldsymbol{H}_{22} + c_u k\boldsymbol{I}_{g_2}$.

Before proceeding to the proof of Theorem 2 we present some general facts that are used in the proofs of several results below. Consider a plug-in estimator $\widetilde{\boldsymbol{\beta}}_2$ of $\boldsymbol{\beta}_2$ and, as

above, let $\widetilde{\boldsymbol{\beta}} = \left(\boldsymbol{\beta}_1', \widetilde{\boldsymbol{\beta}}_2'\right)'$. Notice that under $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_{10}$ it holds that

$$\boldsymbol{X}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \boldsymbol{X}_1\left(\boldsymbol{\beta}_{10} - \boldsymbol{\beta}_1\right) + \boldsymbol{X}_2\left(\widetilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\right) = \boldsymbol{X}_2\left(\widetilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\right),$$

so the residual vector can be written as

$$\widetilde{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}} = \boldsymbol{\varepsilon} - \boldsymbol{X}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \boldsymbol{\varepsilon} - \boldsymbol{X}_2\left(\widetilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\right).$$

The statistic T_2 can be rewritten as

$$\begin{aligned} T_2 &= \frac{1}{\sqrt{k}} \frac{\boldsymbol{\varepsilon}'\boldsymbol{C}\boldsymbol{\varepsilon}}{\sqrt{\widehat{V}\left(\boldsymbol{\beta}\right)}} \left(\frac{\sqrt{\widehat{V}\left(\boldsymbol{\beta}\right)}}{\sqrt{\widehat{V}\left(\widetilde{\boldsymbol{\beta}}\right)}} - 1 \right) + \frac{1}{\sqrt{k}} \frac{\Delta}{\sqrt{\widehat{V}\left(\widetilde{\boldsymbol{\beta}}\right)}} + \frac{1}{\sqrt{k}} \frac{\boldsymbol{\varepsilon}'\boldsymbol{C}\boldsymbol{\varepsilon}}{\sqrt{\widehat{V}\left(\beta\right)}} \\ &\equiv B_1 + B_2 + B_3, \end{aligned} \tag{23}$$

where

$$\Delta = \left(\widetilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\right)' \boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2\left(\widetilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\right) - 2\left(\widetilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\right)' \boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{\varepsilon}. \tag{24}$$

The first term is equal to

$$B_1 = \frac{1}{\sqrt{k}} \frac{\boldsymbol{\varepsilon}'\boldsymbol{C}\boldsymbol{\varepsilon}}{\sqrt{\widehat{V}\left(\boldsymbol{\beta}\right)}} \frac{\sqrt{\widehat{V}\left(\boldsymbol{\beta}\right)} - \sqrt{\widehat{V}\left(\widetilde{\boldsymbol{\beta}}\right)}}{\sqrt{\widehat{V}\left(\widetilde{\boldsymbol{\beta}}\right)}},$$

where from Lemma A.3 and the consistency of the plug-in it follows that $\sqrt{\widehat{V}\left(\boldsymbol{\beta}\right)} - \sqrt{\widehat{V}\left(\widetilde{\boldsymbol{\beta}}\right)} = o_p\left(1\right)$, while since V_n is bounded away from 0 by Assumption 3, it follows that $1/\sqrt{\widehat{V}\left(\widetilde{\boldsymbol{\beta}}\right)} = O_p\left(1\right)$. Theorem 1 implies that $\frac{1}{\sqrt{k}} \frac{\boldsymbol{\varepsilon}'\boldsymbol{C}\boldsymbol{\varepsilon}}{\sqrt{\widehat{V}\left(\boldsymbol{\beta}\right)}} = O_p\left(1\right)$, so $B_1 = o_p\left(1\right)$. Regarding B_3, from Theorem 1 we have that $B_3 \to_d \mathcal{N}\left(0, 1\right)$.

Consequently, if the plug-in estimator $\widetilde{\boldsymbol{\beta}}$ is consistent then under Assumptions 1, 3 we

have $B_1 = o_p(1)$ and $B_3 \to_d \mathcal{N}(0,1)$. In order to derive the asymptotic distribution of T_2 we need to study the term B_2.

Proof of Theorem 2. Note that the first term from Δ in (24) is

$$\left(\widetilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\right)' \boldsymbol{X}_2' \boldsymbol{C} \boldsymbol{X}_2 \left(\widetilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\right) = \left(\widetilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\right)' \boldsymbol{H}_{22}^{1/2} \boldsymbol{H}_{22}^{-1/2} \boldsymbol{X}_2' \boldsymbol{C} \boldsymbol{X}_2 \boldsymbol{H}_{22}^{-1/2} \boldsymbol{H}_{22}^{1/2} \left(\widetilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\right). \tag{25}$$

First we show that $\boldsymbol{H}_{22}^{-1/2} \boldsymbol{X}_2' \boldsymbol{C} \boldsymbol{X}_2 \boldsymbol{H}_{22}^{-1/2} \xrightarrow{p} \boldsymbol{I}_{g_2}$. Lemma A.4 implies that

$$\mathrm{E}\left[\boldsymbol{H}_{22}^{-1/2} \boldsymbol{X}_2' \boldsymbol{C} \boldsymbol{X}_2 \boldsymbol{H}_{22}^{-1/2}\right] = \boldsymbol{I}_{g_2} \tag{26}$$

and

$$\mathrm{Var}\left[\boldsymbol{H}_{22}^{-1/2} \boldsymbol{X}_2' \boldsymbol{C} \boldsymbol{X}_2 \boldsymbol{H}_{22}^{-1/2}\right] \leq \frac{1}{r_{\min}} \boldsymbol{H}_{22}^{-1/2} \left(c_u \boldsymbol{H}_{22} + c_u k \boldsymbol{I}_{g_2} + c_u r_{\max} \boldsymbol{I}_{g_2}\right) \boldsymbol{H}_{22}^{-1/2}$$

$$= \frac{1}{r_{\min}} \left(c_u + c_u \frac{k}{r_{\min}} + c_u \frac{r_{\max}}{r_{\min}}\right) \boldsymbol{I}_{g_2},$$

where the inequality holds due to $\boldsymbol{H}_{22}^{-1} \leq \frac{1}{r_{\min}} \boldsymbol{I}_{g_2}$. Therefore, Assumption 4 (many strong instruments case) implies that $\mathrm{Var}\left[\boldsymbol{H}_{22}^{-1/2} \boldsymbol{X}_2' \boldsymbol{C} \boldsymbol{X}_2 \boldsymbol{H}_{22}^{-1/2}\right] = O\left(\frac{1}{k}\right)$ while Assumption 5 (many weak instruments case) implies that $\mathrm{Var}\left[\boldsymbol{H}_{22}^{-1/2} \boldsymbol{X}_2' \boldsymbol{C} \boldsymbol{X}_2 \boldsymbol{H}_{22}^{-1/2}\right] = o(1)$. In either case we obtain that $\mathrm{Var}\left[\boldsymbol{H}_{22}^{-1/2} \boldsymbol{X}_2' \boldsymbol{C} \boldsymbol{X}_2 \boldsymbol{H}_{22}^{-1/2}\right] \to 0$, which together with (26) implies that $\boldsymbol{H}_{22}^{-1/2} \boldsymbol{X}_2' \boldsymbol{C} \boldsymbol{X}_2 \boldsymbol{H}_{22}^{-1/2} \xrightarrow{p} \boldsymbol{I}_{g_2}$.

Note that under Assumption 4 $\boldsymbol{H}_{22}^{1/2} \left(\widetilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\right) = O_p(1)$ while under Assumption 5 $\frac{1}{\sqrt{k}} \boldsymbol{H}_{22} \left(\widetilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\right) = O_p(1)$ (see Section 4 in Bekker and Crudu, 2015). Therefore, under either Assumption 4 or Assumption 5, from (25) we conclude that

$$\frac{1}{\sqrt{k}} \left(\widetilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\right)' \boldsymbol{X}_2' \boldsymbol{C} \boldsymbol{X}_2 \left(\widetilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\right) = o_p(1). \tag{27}$$

The second term from Δ in (24) involves

$$\left(\widetilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\right)' \boldsymbol{X}_2' \boldsymbol{C}\boldsymbol{\varepsilon} = \left(\widetilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\right)' \boldsymbol{H}_{22}\boldsymbol{H}_{22}^{-1}\boldsymbol{X}_2' \boldsymbol{C}\boldsymbol{\varepsilon}.$$

Lemma A.4 implies that

$$\mathrm{E}\left[\boldsymbol{X}_2' \boldsymbol{C}\boldsymbol{\varepsilon}\right] = \boldsymbol{0} \tag{28}$$

and

$$\mathrm{Var}\left[\boldsymbol{H}_{22}^{-1}\boldsymbol{X}_2' \boldsymbol{C}\boldsymbol{\varepsilon}\right] \leq c_u \left(\frac{1}{r_{\min}} + \frac{k}{r_{\min}^2}\right) \boldsymbol{I}_{g_2}, \tag{29}$$

where the latter inequality is due to $\boldsymbol{H}_{22}^{-1} \leq \frac{1}{r_{\min}} \boldsymbol{I}_{g_2}$. We also obtain that

$$\mathrm{Var}\left[\frac{1}{\sqrt{k}} \boldsymbol{H}_{22}^{-1/2}\boldsymbol{X}_2' \boldsymbol{C}\boldsymbol{\varepsilon}\right] \leq c_u \left(\frac{1}{k} + \frac{1}{r_{\min}}\right) \boldsymbol{I}_{g_2}. \tag{30}$$

Under Assumption 4 (many strong instruments case) we get $\mathrm{Var}\left[\frac{1}{\sqrt{k}} \boldsymbol{H}_{22}^{-1/2}\boldsymbol{X}_2' \boldsymbol{C}\boldsymbol{\varepsilon}\right] = O\left(\frac{1}{k}\right)$, which together with (28) implies that $\frac{1}{\sqrt{k}} \boldsymbol{H}_{22}^{-1/2}\boldsymbol{X}_2' \boldsymbol{C}\boldsymbol{\varepsilon} = o_p\left(1\right)$. Since $\boldsymbol{H}_{22}^{1/2}\left(\widetilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\right) = O_p\left(1\right)$ holds, we obtain

$$\frac{1}{\sqrt{k}} \left(\widetilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\right)' \boldsymbol{X}_2' \boldsymbol{C}\boldsymbol{\varepsilon} = o_p\left(1\right).$$

Under Assumption 5 (many weak instruments case) (29) implies $\mathrm{Var}\left[\boldsymbol{H}_{22}^{-1}\boldsymbol{X}_2' \boldsymbol{C}\boldsymbol{\varepsilon}\right] = o\left(1\right)$, which together with (28) implies that $\boldsymbol{H}_{22}^{-1}\boldsymbol{X}_2' \boldsymbol{C}\boldsymbol{\varepsilon} = o_p\left(1\right)$. Since $\frac{1}{\sqrt{k}} \boldsymbol{H}_{22}\left(\widetilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\right) = O_p\left(1\right)$ holds, we obtain

$$\frac{1}{\sqrt{k}} \left(\widetilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\right)' \boldsymbol{X}_2' \boldsymbol{C}\boldsymbol{\varepsilon} = o_p\left(1\right).$$

Regarding B_3, from Theorem 1 we have that $B_3 \xrightarrow{d} \mathcal{N}\left(0, 1\right)$. □

Proof of Theorem 3. Note that

$$\sum_{i\neq j} C_{ij}\varepsilon_i\varepsilon_j = \sum_{i\neq j} \frac{P_{ij}}{2}\left(\frac{1}{1-P_{ii}}+\frac{1}{1-P_{jj}}\right)\varepsilon_i\varepsilon_j = (1+o(1))\sum_{i\neq j} P_{ij}\varepsilon_i\varepsilon_j$$

as $\max_i P_{ii} \to 0$. Further,

$$\sum_{i\neq j} C_{ij}\varepsilon_i\varepsilon_j = (1+o(1))\boldsymbol{\varepsilon}'\boldsymbol{P}\boldsymbol{\varepsilon} - (1+o(1))\sum_{i=1}^{n} P_{ii}\varepsilon_i^2.$$

By assumptions (ii) and (iv), $\boldsymbol{\varepsilon}'\boldsymbol{P}\boldsymbol{\varepsilon} \to_d \sigma^2\chi_k^2$. Moreover, by independence of ε_i, Assumption 3 and the properties of P_{ii}

$$\mathrm{E}\left[\left(\sum_{i=1}^{n} P_{ii}\varepsilon_i^2 - k\sigma^2\right)^2\right] \to 0,$$

which implies $\sum_{i=1}^n P_{ii}\varepsilon_i^2 \to_p k\sigma^2$ (see Chao *et al.*, 2014). Consider now

$$V_n = \frac{2}{k}\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}^2\sigma_i^2\sigma_j^2 = \frac{2\sigma^4}{k}\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}^2 = \frac{2\sigma^4}{k}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{P_{ij}^2}{4}\left(\frac{1}{1-P_{ii}}+\frac{1}{1-P_{jj}}\right)^2.$$

Since $\sum_{i=1}^n\sum_{j=1}^n P_{ij}^2 = k$ is fixed and $\max_i P_{ii} \to 0$, $V_n \to 2\sigma^4$. Hence, by Lemma A.2, $\widehat{V}(\boldsymbol{\beta}_0) \to_p 2\sigma^4$. Finally,

$$T_1 = \frac{1}{\sqrt{k}}\frac{\boldsymbol{\varepsilon}_0'\boldsymbol{C}\boldsymbol{\varepsilon}_0}{\sqrt{\frac{2}{k}\boldsymbol{\varepsilon}_0^{(2)\prime}\boldsymbol{C}^{(2)}\boldsymbol{\varepsilon}_0^{(2)}}} \to_d \frac{\chi_k^2-k}{\sqrt{2k}}.$$

Thus, $\sqrt{2k}T_1 + k \to_d \chi_k^2$. Let us consider now the T_2 statistic. Notice that $\widetilde{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}_1\boldsymbol{\beta}_{10} - \boldsymbol{X}_2\widetilde{\boldsymbol{\beta}}_2$ where $\widetilde{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2'\boldsymbol{P}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{P}(\boldsymbol{y} - \boldsymbol{X}_1\boldsymbol{\beta}_{10})$, the two-stage least squares under the null. By standard manipulations, CLT and Slutsky's theorem we get

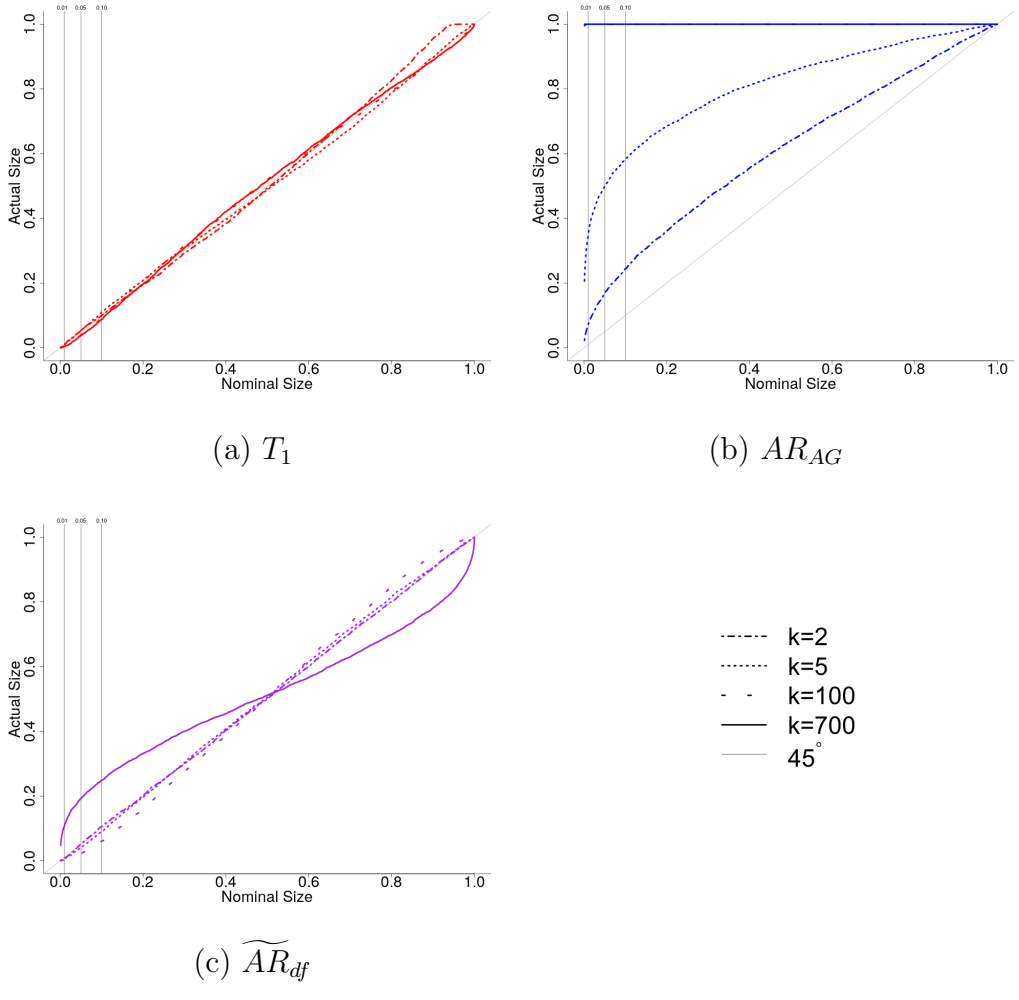
$$\widetilde{\boldsymbol{\varepsilon}}'\boldsymbol{P}\widetilde{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon}'\left(\boldsymbol{P} - \boldsymbol{P}\boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{P}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{P}\right)\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}'\boldsymbol{Z}^*\left(\boldsymbol{I}_k - \boldsymbol{P}_{X_2'Z^*}\right)\boldsymbol{Z}^{*\prime}\boldsymbol{\varepsilon} \to_d \sigma^2\chi_{k-g_2}^2$$

where $\boldsymbol{Z}^* = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1/2}$. Using the results in Lemma A.3 we get $\sum_{i=1}^n P_{ii}\widetilde{\varepsilon}_i^2 \to_p k\sigma^2$ and

$\widehat{V}(\widetilde{\boldsymbol{\beta}}) \to_p 2\sigma^4$. So, by the usual standard arguments $\sqrt{2k}T_2 + k \to_d \chi^2_{k-g_2}$. □

Proof of Corollary 1. The proof mimics that of Theorem 1 in Chao *et al.* (2014). Let $q_\alpha^{\chi^2_k}$ be the generic α-th quantile of the chi square distribution with k degrees of freedom. As $k \to \infty$, $\frac{q_\alpha^{\chi^2_k} - k}{\sqrt{2k}} \to z_\alpha$, where z_α is the generic α-th quantile of the standard normal distribution. This proves parts (i) and (ii). With respect to part (iii) and part (iv), notice that $\sqrt{\frac{k-g_2}{k}} \frac{q_\alpha^{\chi^2_{k-g_2}} - (k-g_2)}{\sqrt{2(k-g_2)}} - \frac{g_2}{\sqrt{2k}} \to z_\alpha$ as $k \to \infty$. □

B Figures

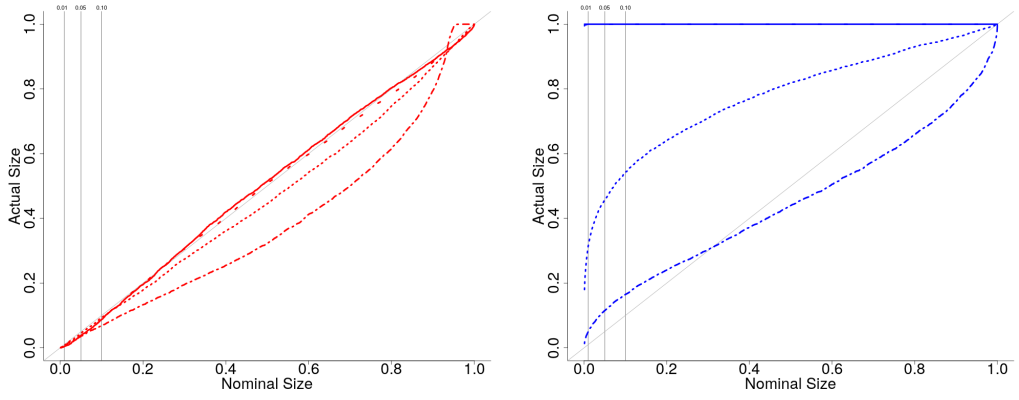


(a) T_1

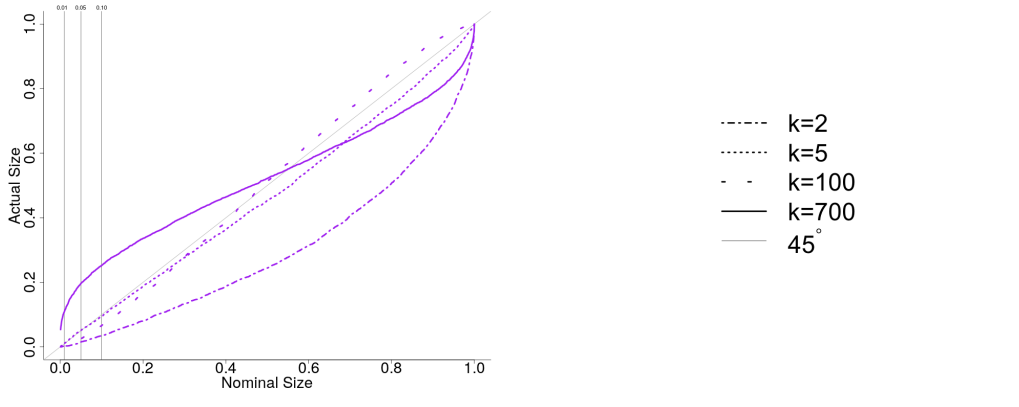
(b) AR_{AG}

(c) $\widetilde{AR}_{df}$

Figure 1: PP-plots with heteroskedasticity, $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$.

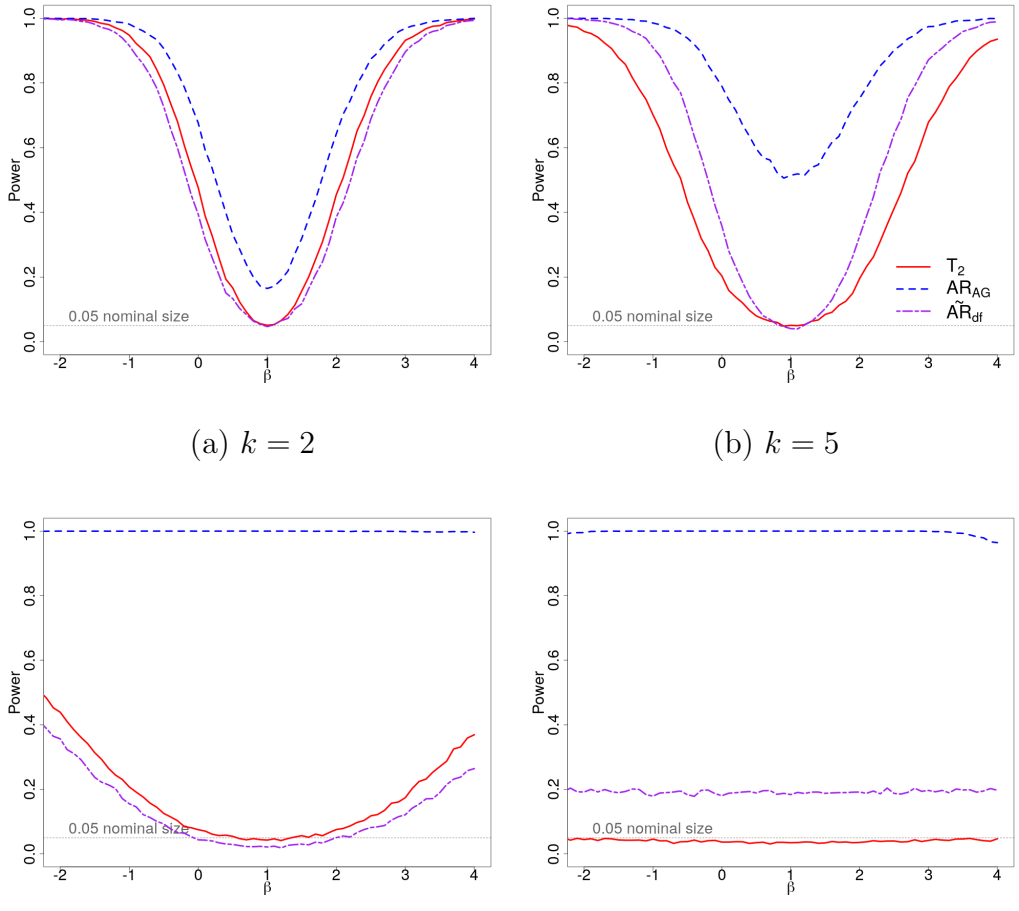


(a) T_2 (b) AR_{AG}



(c) $\widetilde{AR}_{df}$

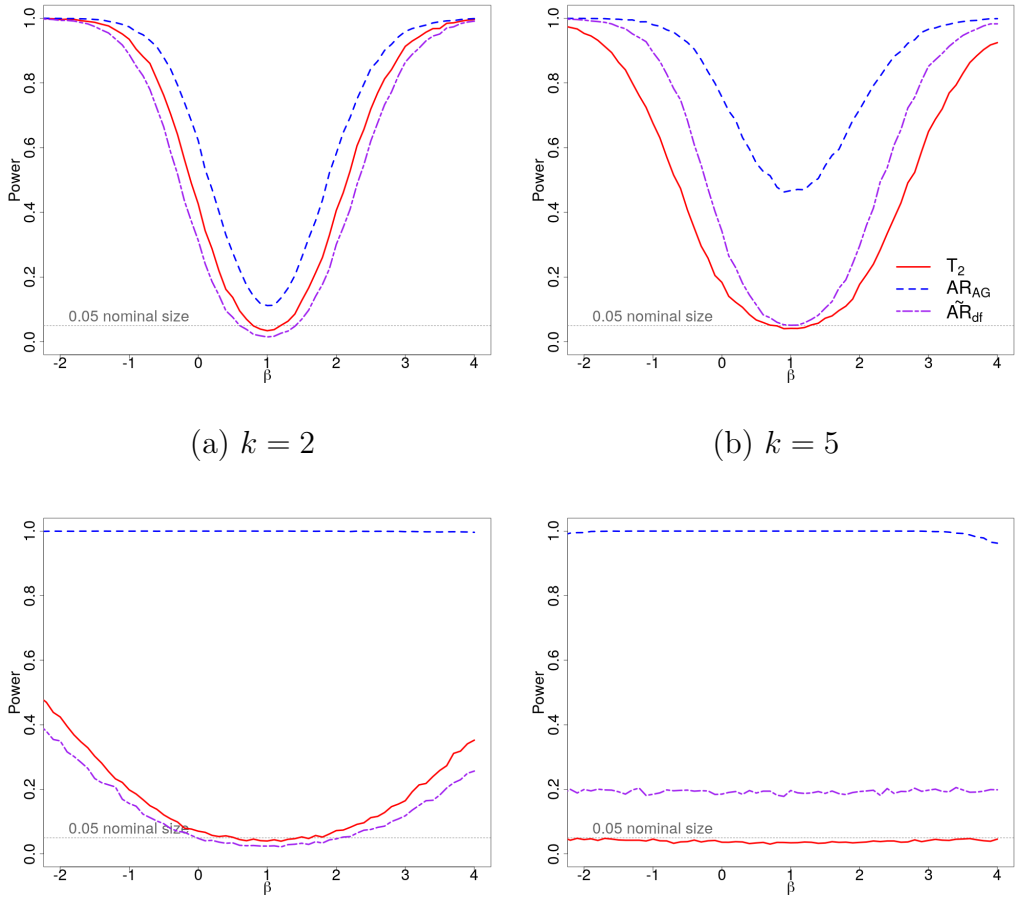
Figure 2: PP-plots with heteroskedasticity, $H_0 : \beta_1 = \beta_{10}$.



(a) $k = 2$ (b) $k = 5$

(c) $k = 100$ (d) $k = 700$

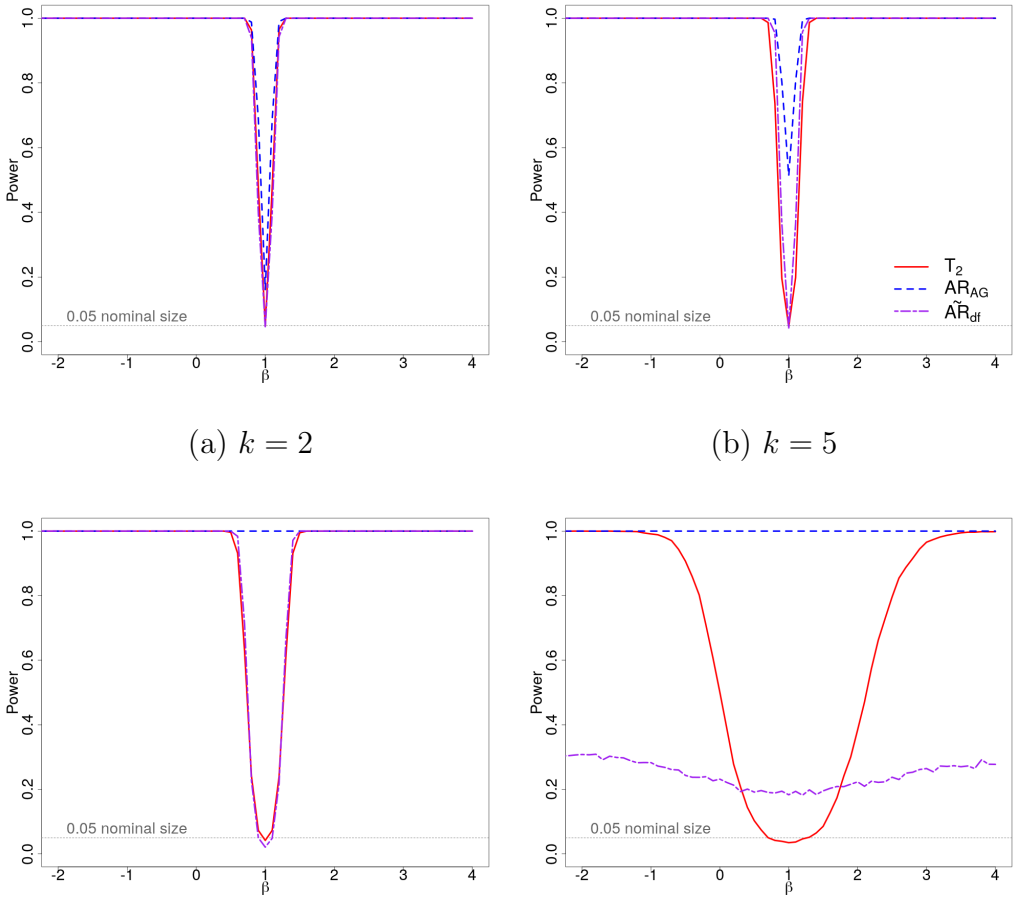
Figure 3: Power curves with heteroskedasticity and $\pi = 0.1$, $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$.



(a) $k = 2$ (b) $k = 5$

(c) $k = 100$ (d) $k = 700$

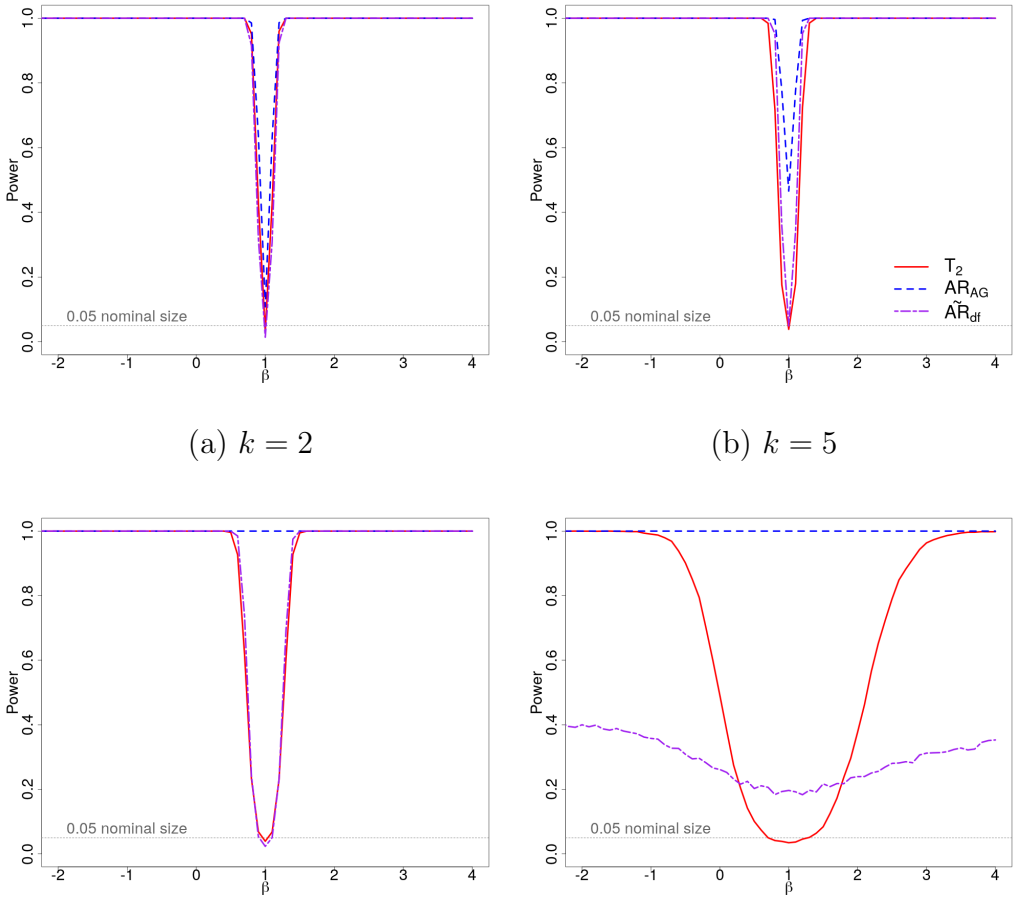
Figure 4: Power curves with heteroskedasticity and $\pi = 0.1$, $H_0 : \beta_1 = \beta_{10}$.



(a) $k = 2$ (b) $k = 5$

(c) $k = 100$ (d) $k = 700$

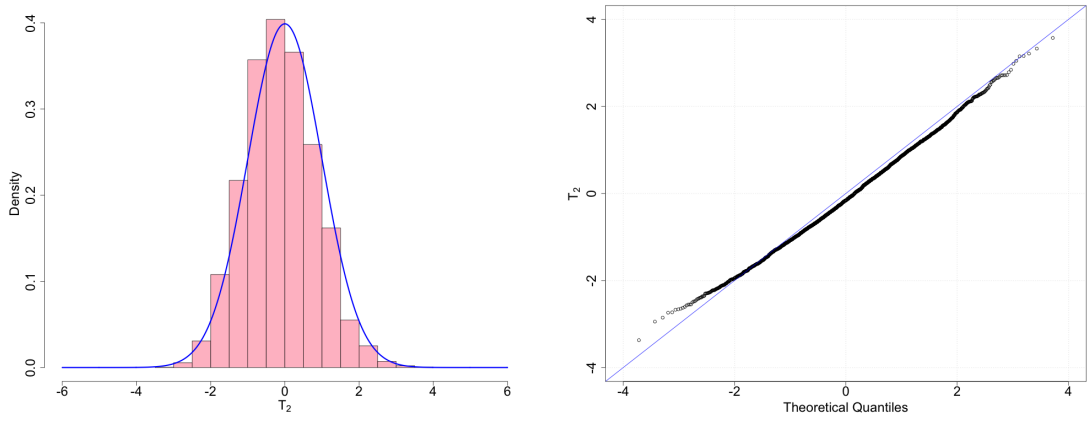
Figure 5: Power curves with heteroskedasticity and $\pi = 1$, $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$.



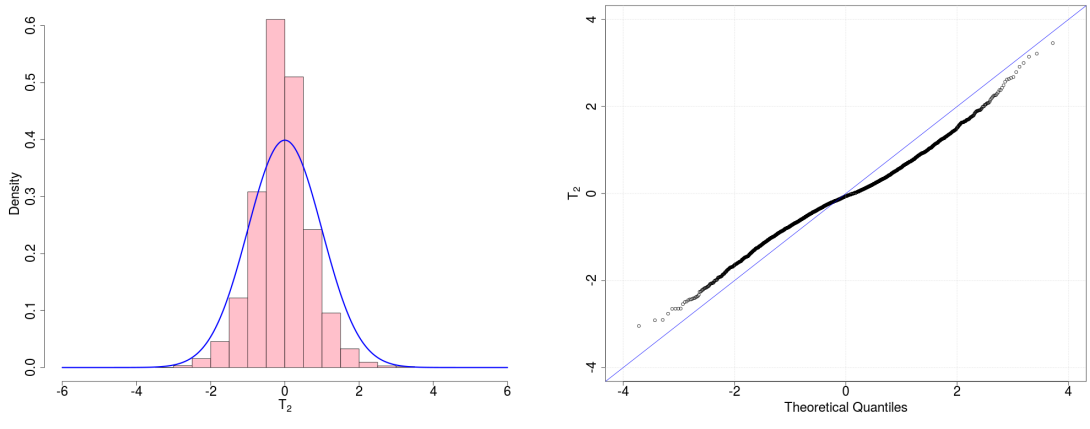
(a) $k = 2$ (b) $k = 5$

(c) $k = 100$ (d) $k = 700$

Figure 6: Power curves with heteroskedasticity and $\pi = 1$, $H_0 : \beta_1 = \beta_{10}$.

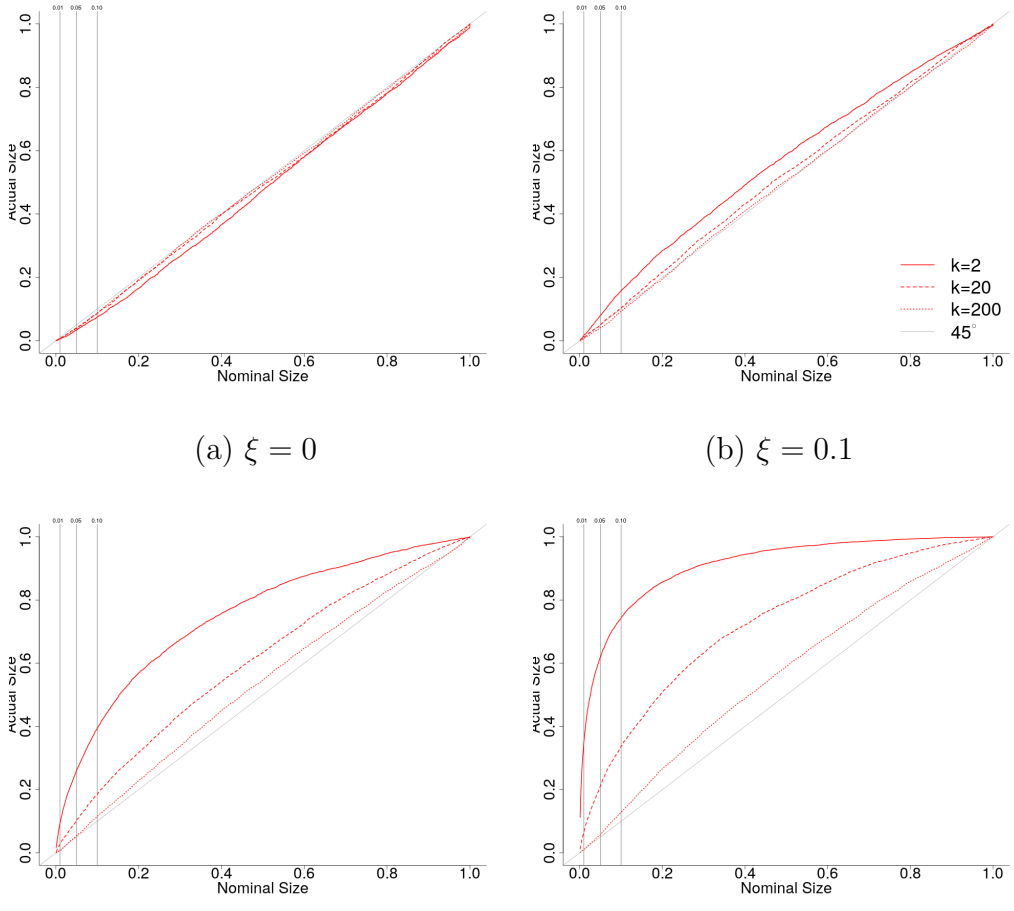


(a) $n = 400$, $\frac{\sqrt{k}}{n\boldsymbol{\pi}_w'\boldsymbol{\pi}_w} = 0.1$



(b) $n = 400$, $\frac{\sqrt{k}}{n\boldsymbol{\pi}_w'\boldsymbol{\pi}_w} = 1$

Figure 7: Histograms and QQ-plots for T_2 and JIV1 plug-in. The blue curve superimposed on the histograms is the standard normal distribution.



(a) $\xi = 0$ (b) $\xi = 0.1$

(c) $\xi = 0.2$ (d) $\xi = 0.3$

Figure 8: PP-plots for T_2 with a slow (inconsistent) plug-in.

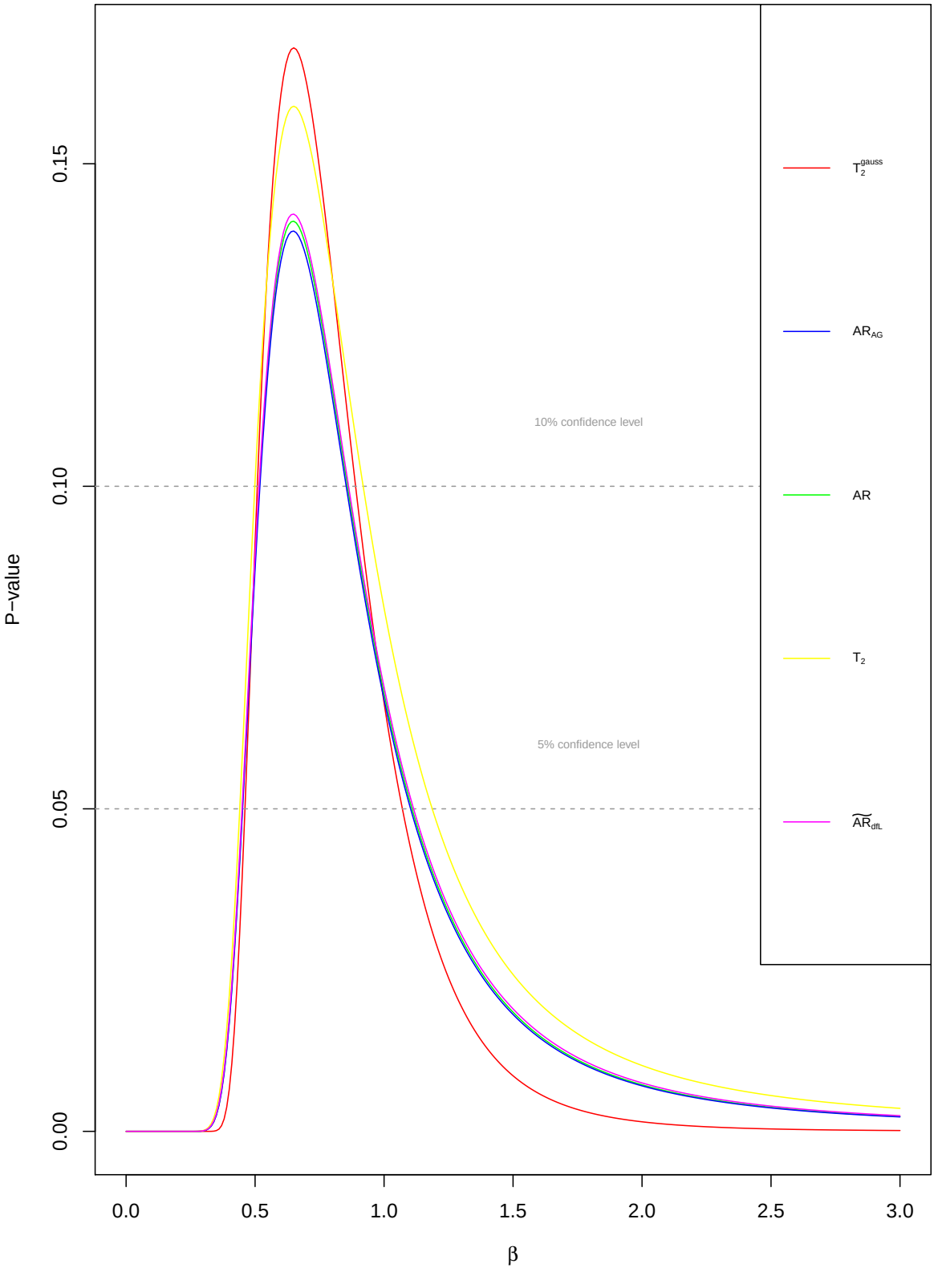


Figure 9: T_2, AR_{AG} and AR P-values for different values of β for the model with four instruments.

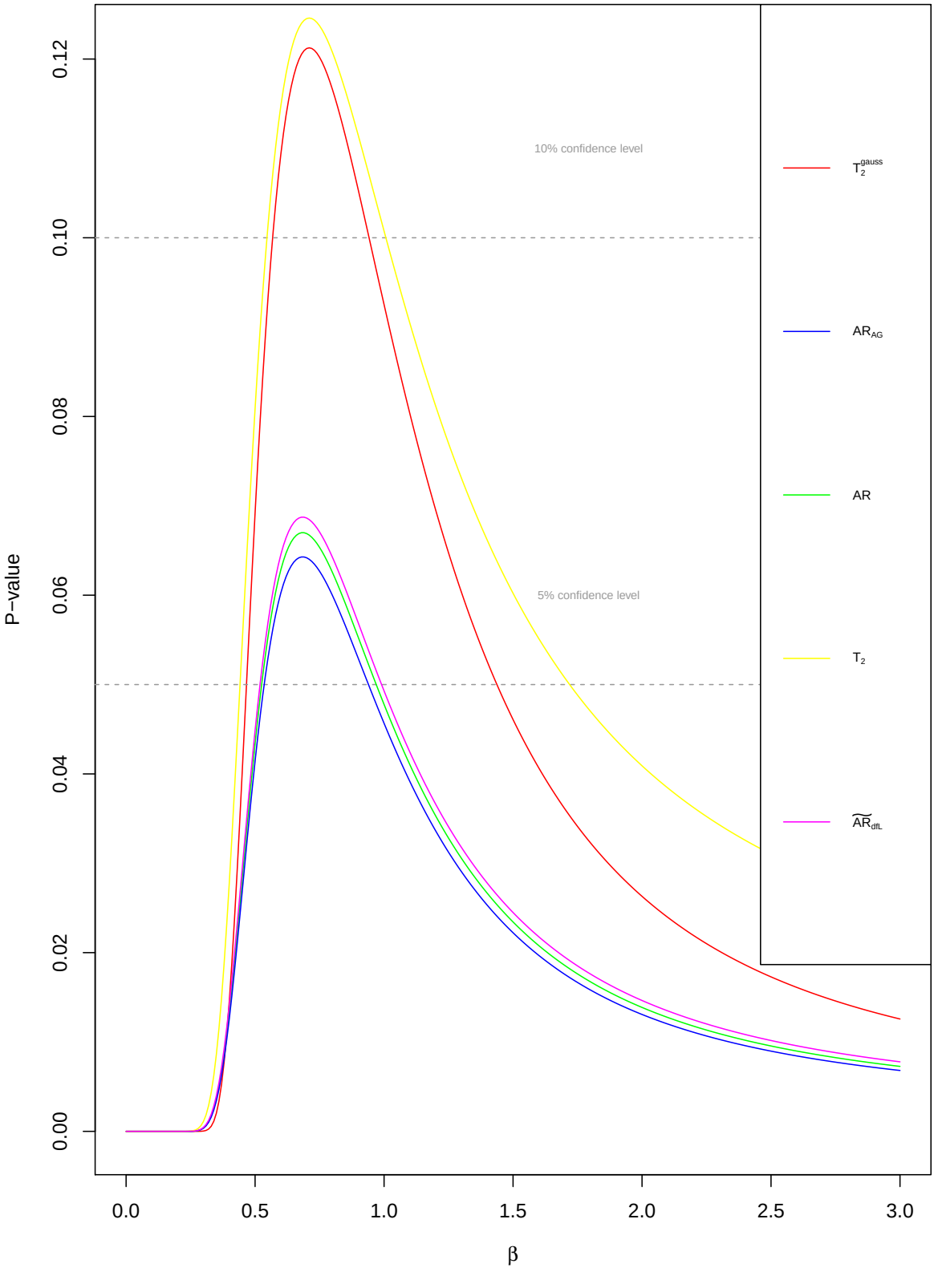


Figure 10: T_2, AR_{AG} and AR P-values for different values of β for the model with fourteen instruments.

References

Anatolyev, S. (2018) Many Instruments and/or Regressors: A Friendly Guide. *Journal of Economic Surveys* 33, 689–726.

Anatolyev, S. and Gospodinov, N. (2011) Specification Testing in Models with Many Instruments. *Econometric Theory* 27, 427–441.

Anatolyev, S. and Yaskov, P. (2017) Asymptotics of diagonal elements of projection matrices under many instruments/regressors. *Econometric Theory* 33, 717–738.

Anderson, T.W. and Rubin, H. (1949) Estimators of the Parameters of a Single Equation in a Complete Set of Stochastic Equations. *The Annals of Mathematical Statistics* 21, 570–582.

Andrews, D.W.K., Marmer, V. and Yu, Z. (2019) On optimal inference in the linear IV model. *Quantitative Economics* 10, 457–485.

Andrews, D.W.K., Moreira, M.J. and Stock, J.H. (2006) Optimal Two-Sided Invariant Similar Tests for Instrumental Variable Regression. *Econometrica* 73, 715–752.

Andrews, D.W.K. and Stock, J. (2005) Inference with Weak Instruments. In R. Blundell, W.K. Newey and T. Persson (eds.), *Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society*, volume III, Cambridge University Press, Cambridge.

Andrews, D.W.K. and Stock, J.H. (2007) Testing with Many Weak Instruments. *The Journal of Econometrics* 138, 24–46.

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

Angrist, J.D. and Pischke, J.S. (2008) *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Bekker, P.A. (1994) Alternative approximations to the distributions of instrumental variable estimators. *Econometrica* 54, 657–682.

Bekker, P.A. and Crudu, F. (2015) Jackknife Instrumental Variable Estimation with Heteroskedasticity. *The Journal of Econometrics* 185, 332–342.

Bekker, P.A. and Van der Ploeg, J. (2005) Instrumental variable estimation based on grouped data. *Statistica Neerlandica* 59, 239–267.

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

Bun, M., Farbmacher, H. and Poldermans, R. (2018) Finite sample properties of the Anderson and Rubin (1949) test. *working paper* .

Card, D. (1995) Using Geographic Variation in College Proximity to Estimate the Return to Schooling. In L. Christofides, E. Grant and R. Swidinsky (eds.), *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp*, University of Toronto Press, Toronto, 201–222.

Chao, J.C., Hausman, J.A., Newey, W.K., Swanson, N.R. and Woutersen, T. (2014) Testing Overidentifying Restrictions with Many Instruments and Heteroskedasticity. *The Journal of Econometrics* 178, 15–21.

Chao, J.C. and Swanson, N.R. (2005) Consistent estimation with a large number of weak instruments. *Econometrica* 73, 1673–1692.

Chao, J.C., Swanson, N.R., Hausman, J.A., Newey, W.K. and Woutersen, T. (2012) Asymptotic distribution of JIVE in a heteroskedastic IV regression with many instruments. *Econometric Theory* 28, 42–86.

Davidson, R. and MacKinnon, J.G. (1998) Graphical Methods for Investigating the Size and Power of Hypothesis Tests. *The Manchester School* 66, 1–26.

Donald, S.G., Imbens, G.W. and Newey, W.K. (2003) Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics* 117, 55–93.

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

Guggenberger, P., Kleibergen, F., Mavroeidis, S. and Chen, L. (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 Smith, R.J. (2005) Generalized Empirical Likelihood Estimators and Tests Under Partial, Weak, and Strong Identification. *Econometric Theory* 21, 667–709.

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

Imbens, G.W. (2014) Instrumental Variables: An Econometrician's Perspective. *Statistical Science* 29, 323–358.

Imbens, G.W. and Rubin, D. (1997) Estimating outcome distributions for compliers in instrumental variables models. *Review of Economic Studies* 64, 555–574.

Jiang, Y., Kang, H. and Small, D. (2016) *ivmodel: Statistical Inference and Sensitivity Analysis for Instrumental Variables Model*. URL `https://CRAN.R-project.org/package=ivmodel`, r package version 1.2.

Kang, H., Zhang, A., Cai, T.T. and Small, D.S. (2016) Instrumental Variables Estimation With Some Invalid Instruments and its Application to Mendelian Randomization. *Journal of the American Statistical Association* 111, 132–144.

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

Kleibergen, F. (2004) Testing Subsets of Structural Parameters in the Instrumental Variables. *The Review of Economics and Statistics* 86, 418–423.

Kleibergen, F. (2005) Testing Parameters in GMM Without Assuming They Are Identified. *Econometrica* 73, 1103–1123.

Lee, Y. and Okui, R. (2012) Hahn–Hausman test as a specification test. *Journal of Econometrics* 167, 133–139.

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

Moreira, M.J. (2009) Tests with Correct Size When Instruments Can Be Arbitrarily Weak. *The Journal of Econometrics* 152, 131–140.

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

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

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

Van Hasselt, M. (2010) Many instruments asymptotic approximations under nonnormal error distributions. *Econometric Theory* 26, 633–645.

Von Hinke, S., Davey Smith, G., Lawlor, D.A., Propper, C. and Windmeijer, F. (2016) Genetic markers as instrumental variables. *Journal of Health Economics* 45, 131–148.

Wang, J. and Zivot, E. (1998) Inference on a Structural Parameter in Instrumental Variables Regression with Weak Instruments. *Econometrica* 66, 1389–1404.

Windmeijer, F., Farbmacher, H., Davies, N. and Davey Smith, G. (2017) On the Use of the Lasso for Instrumental Variables Estimation with Some Invalid Instruments. Bristol economics discussion papers, Department of Economics, University of Bristol, UK.

Zivot, E., Startz, R. and Nelson, C.R. (1998) Valid Confidence Intervals and Inference in the Presence of Weak Instruments. *International Economic Review* 39, 1119–1144.

Supplement to "Inference in instrumental variables models with heteroskedasticity and many instruments"

Federico Crudu*
Università di Siena and CRENoS

Giovanni Mellace†
University of Southern Denmark

Zsolt Sándor‡
Sapientia Hungarian University of Transylvania

November 2019

Abstract

This supplement contains the proofs of the auxiliary lemmas, some additional theoretical results and further Monte Carlo experiments that complement the results in the main text.

*Department of Economics and Statistics, Piazza San Francesco 7/8, 53100 Siena, Italy, federico.crudu@unisi.it

†Department of Business and Economics, Campusvej 55, 5230 Odense M, Denmark, giome@sam.sdu.dk

‡Department of Business Sciences, Piata Libertăii 1, 530104 Miercurea Ciuc, Romania, sandorzsolt@cs.sapientia.ro

A Proofs of Lemmas

This Section contains some auxiliary lemmas that are useful to prove the main results of the paper.

Lemma A.2. *Let* $\widehat{V}(\boldsymbol{\beta}_0) = \frac{2}{k}\boldsymbol{\varepsilon}_0^{(2)\prime}\boldsymbol{C}^{(2)}\boldsymbol{\varepsilon}_0^{(2)}$. *If Assumptions 1, 3 hold,* $\widehat{V}(\boldsymbol{\beta}_0) - V_n = O_p\left(\frac{1}{\sqrt{k}}\right)$; *consequently* $\widehat{V}(\boldsymbol{\beta}_0) - V_n \xrightarrow{p} 0$, *where*

$$V_n = \frac{2}{k}\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}^2\sigma_i^2\sigma_j^2.$$

Proof. Let $\eta_i = \varepsilon_i^2 - \sigma_i^2$; then

$$\widehat{V}(\boldsymbol{\beta}_0) - V_n = \frac{2}{k}\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}^2\left(\varepsilon_i^2\varepsilon_j^2 - \sigma_i^2\sigma_j^2\right) = \frac{2}{k}\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}^2\left(\eta_i\eta_j + \sigma_i^2\eta_j + \sigma_j^2\eta_i\right).$$

So

$$\begin{aligned}\left|V_n - \widehat{V}(\boldsymbol{\beta}_0)\right| &\leq \frac{2}{k}\left|\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}^2\eta_i\eta_j\right| + \frac{2}{k}\left|\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}^2\sigma_i^2\eta_j\right| + \frac{2}{k}\left|\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}^2\sigma_j^2\eta_i\right| \\ &\equiv A_1 + A_2 + A_3.\end{aligned}$$

Since

$$\mathrm{E}\left[\eta_i^2\right] = \mathrm{E}\left[\varepsilon_i^4\right] - \sigma_i^4,$$

from Assumption 3 we have $\mathrm{E}\left[\eta_i^2\right] \leq c_u$. So

$$\mathrm{E}\left[A_1^2\right] = \frac{8}{k^2}\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}^4\,\mathrm{E}\left[\eta_i^2\right]\mathrm{E}\left[\eta_j^2\right] \leq \frac{c_u}{k^2}\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}^4.$$

Note that for $i \neq j$ we have

$$C_{ij} = \frac{P_{ij}}{2}\left(\frac{1}{1-P_{ii}} + \frac{1}{1-P_{jj}}\right),$$

which from Assumption 1 implies

$$|C_{ij}| = \frac{|P_{ij}|}{2}\left(\frac{1}{1-P_{ii}} + \frac{1}{1-P_{jj}}\right) \leq c_u\,|P_{ij}| \quad \text{for any } i,j, \tag{A.1}$$

so

$$\mathrm{E}\left[A_1^2\right] \leq \frac{c_u}{k^2}\sum_{i=1}^{n}\sum_{j=1}^{n} P_{ij}^4.$$

From Assumption 1 and the fact that $\boldsymbol{P} = \boldsymbol{P}^2$, we have

$$P_{hh} \geq P_{hh}^2 = \left(\sum_{i=1}^{n} P_{hi}^2\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} P_{hi}^2 P_{hj}^2 \geq \sum_{i=1}^{n} P_{hi}^4,$$

so

$$\sum_{i=1}^{n}\sum_{j=1}^{n} P_{ij}^4 \leq tr\left(\boldsymbol{P}\right) = k \quad \text{and} \quad \sum_{i,j,h} P_{hi}^2 P_{hj}^2 \leq k. \tag{A.2}$$

Therefore,

$$\mathrm{E}\left[A_1^2\right] \leq \frac{c_u}{k}.$$

Now, by Cauchy-Schwarz $\left(\mathrm{E}\left[\varepsilon_i^2\right]\right)^2 \leq \mathrm{E}\left[\varepsilon_i^4\right]$, thus $\sigma_i^2 \leq c_u$, so from Assumption 3, (A.1) and (A.2)

$$\mathrm{E}\left[A_2^2\right] = \frac{4}{k^2}\sum_{i,j,h} C_{hi}^2 C_{ij}^2 \sigma_h^2 \sigma_j^2\,\mathrm{E}\left[\eta_i^2\right] \leq \frac{4c_u^2}{k^2}\sum_{i,j,h} C_{hi}^2 C_{ij}^2 \leq \frac{c_u}{k^2}\sum_{i,j,h} P_{hi}^2 P_{ij}^2 \leq \frac{c_u}{k}.$$

We can obtain a similar inequality for A_3, so by the Markov and triangle inequalities we obtain that $\widehat{V}\left(\boldsymbol{\beta}_0\right) - V_n = O_p\left(\frac{1}{\sqrt{k}}\right)$, therefore, $\widehat{V}\left(\boldsymbol{\beta}_0\right) - V_n \xrightarrow{p} 0$. □

Lemma A.3. *Let* $\widehat{V}\left(\widetilde{\boldsymbol{\beta}}\right) = \frac{2}{k}\widetilde{\boldsymbol{\varepsilon}}^{(2)\prime}\boldsymbol{C}^{(2)}\widetilde{\boldsymbol{\varepsilon}}^{(2)}$. *If* $\widetilde{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ *and Assumptions 1, 3 hold, then* $\widehat{V}\left(\widetilde{\boldsymbol{\beta}}\right) - V_n \xrightarrow{p} 0$.

Proof. Let

$$\widetilde{V}_n = \frac{2}{k}\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}^2 \varepsilon_i^2 \varepsilon_j^2.$$

Then

$$\widehat{V}\left(\widetilde{\boldsymbol{\beta}}\right) - \widetilde{V}_n = \frac{2}{k}\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}^2 \left(\widetilde{\varepsilon}_i^2 \widetilde{\varepsilon}_j^2 - \varepsilon_i^2 \varepsilon_j^2\right).$$

Note that

$$\left|\widetilde{\varepsilon}_i^2 \widetilde{\varepsilon}_j^2 - \varepsilon_i^2 \varepsilon_j^2\right| \leq \widetilde{\varepsilon}_j^2 \left|\widetilde{\varepsilon}_i^2 - \varepsilon_i^2\right| + \varepsilon_i^2 \left|\widetilde{\varepsilon}_j^2 - \varepsilon_j^2\right|,$$

$$\left|\widetilde{\varepsilon}_i + \varepsilon_i\right| \leq \left|\widetilde{\varepsilon}_i - \varepsilon_i\right| + 2\left|\varepsilon_i\right|,$$

$$\widetilde{\varepsilon}_j^2 = \left|\widetilde{\varepsilon}_j^2 - \varepsilon_j^2 + \varepsilon_j^2\right| \leq \left|\widetilde{\varepsilon}_j^2 - \varepsilon_j^2\right| + \varepsilon_j^2$$

and

$$\begin{aligned}\left|\widetilde{\varepsilon}_i^2 - \varepsilon_i^2\right| &= \left|\widetilde{\varepsilon}_i - \varepsilon_i\right| \cdot \left|\widetilde{\varepsilon}_i + \varepsilon_i\right| = \left|\boldsymbol{X}_i'\left(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\right)\right| \cdot \left|\widetilde{\varepsilon}_i + \varepsilon_i\right| \leq \left|\boldsymbol{X}_i'\left(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\right)\right| \cdot \left(\left|\boldsymbol{X}_i'\left(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\right)\right| + 2\left|\varepsilon_i\right|\right) \\ &\leq \left\|\boldsymbol{X}_i\right\| \left\|\left(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\right)\right\| \cdot \left(\left\|\boldsymbol{X}_i\right\| \left\|\left(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\right)\right\| + 2\left|\varepsilon_i\right|\right) \equiv d_i \left\|\left(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\right)\right\|,\end{aligned}$$

where

$$d_i = \left\|\boldsymbol{X}_i\right\| \left(\left\|\boldsymbol{X}_i\right\| + 2\left|\varepsilon_i\right|\right)$$

because $\left\|\left(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\right)\right\| \leq 1$ with probability approaching 1 as $n \to \infty$. So

$$\left|\widetilde{\varepsilon}_i^2 \widetilde{\varepsilon}_j^2 - \varepsilon_i^2 \varepsilon_j^2\right| \leq d_i d_j \left\|\left(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\right)\right\|^2 + \left(d_i \varepsilon_j^2 + d_j \varepsilon_i^2\right) \left\|\left(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\right)\right\|,$$

therefore,

$$\left|\widehat{V}\left(\widetilde{\boldsymbol{\beta}}\right)-\widetilde{V}_n\right| \leq \frac{2\left\|\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)\right\|^2}{k}\sum_{i=1}^{n}\sum_{j=1}^{n}C_{ij}^2 d_i d_j + \frac{2\left\|\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)\right\|}{k}\sum_{i=1}^{n}\sum_{j=1}^{n}C_{ij}^2 d_i \varepsilon_j^2. \quad \text{(A.3)}$$

Note that by the Cauchy-Schwarz inequality

$$\mathrm{E}\left[d_i d_j\right] \leq \sqrt{\mathrm{E}\left[d_i^2\right]\mathrm{E}\left[d_j^2\right]},$$

where

$$\begin{aligned}\mathrm{E}\left[d_i^2\right] &= \mathrm{E}\left[\left\|\boldsymbol{X}_i\right\|^4\right] + 4\,\mathrm{E}\left[\left\|\boldsymbol{X}_i\right\|^3\left|\varepsilon_i\right|\right] + 4\,\mathrm{E}\left[\left\|\boldsymbol{X}_i\right\|^2\varepsilon_i^2\right] \\ &\leq \mathrm{E}\left[\left\|\boldsymbol{X}_i\right\|^4\right] + 4\sqrt{\mathrm{E}\left[\left\|\boldsymbol{X}_i\right\|^4\right]\mathrm{E}\left[\left\|\boldsymbol{X}_i\right\|^2\varepsilon_i^2\right]} + 4\sqrt{\mathrm{E}\left[\left\|\boldsymbol{X}_i\right\|^4\right]\mathrm{E}\left[\varepsilon_i^4\right]}.\end{aligned}$$

Assumption 3 and Minkowski's inequality imply $\mathrm{E}\left[\left\|\boldsymbol{X}_i\right\|^4\right] \leq c_u$. Hence

$$\mathrm{E}\left[d_i d_j\right] \leq c_u, \quad \mathrm{E}\left[d_i^2\right] \leq c_u,$$

so by Assumption 1 and (A.1)

$$\mathrm{E}\left[\frac{1}{k}\sum_{i=1}^{n}\sum_{j=1}^{n}C_{ij}^2 d_i d_j\right] \leq \frac{1}{k}\sum_{i=1}^{n}\sum_{j=1}^{n}C_{ij}^2\,\mathrm{E}\left[d_i d_j\right] \leq c_u\left(\frac{1}{k}\sum_{i=1}^{n}\sum_{j=1}^{n}C_{ij}^2\right) \leq c_u$$

and

$$\mathrm{E}\left[\frac{1}{k}\sum_{i=1}^{n}\sum_{j=1}^{n}C_{ij}^2 d_i^2\right] \leq c_u.$$

Then by Markov's and the triangle inequalities $\widehat{V}\left(\widetilde{\boldsymbol{\beta}}\right) - V_n \xrightarrow{p} 0$. □

Lemma A.4. *Under Assumptions 1, 3*

1. $\mathrm{E}\left[\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2\right] = \boldsymbol{H}_{22},\ \mathrm{Var}\left[\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2\right] \leq c_u\boldsymbol{H}_{22} + c_u k\boldsymbol{I}_{g_2} + c_u r_{\max}\boldsymbol{I}_{g_2}$,

2. $\mathrm{E}\left[\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{\varepsilon}\right] = \boldsymbol{0},\ \mathrm{Var}\left[\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{\varepsilon}\right] \leq c_u\boldsymbol{H}_{22} + c_u k\boldsymbol{I}_{g_2}$.

Proof. The model $\boldsymbol{X}_2 = \boldsymbol{Z\Pi}_2 + \boldsymbol{U}_2$ implies that

$$\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2 = (\boldsymbol{Z\Pi}_2)'\boldsymbol{C}\boldsymbol{Z\Pi}_2 + (\boldsymbol{Z\Pi}_2)'\boldsymbol{C}\boldsymbol{U}_2 + \boldsymbol{U}_2'\boldsymbol{C}\boldsymbol{Z\Pi}_2 + \boldsymbol{U}_2'\boldsymbol{C}\boldsymbol{U}_2. \tag{A.4}$$

Therefore,

$$\mathrm{E}\left[\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2\right] = \boldsymbol{\Pi}_2'\boldsymbol{Z}'\boldsymbol{C}\boldsymbol{Z\Pi}_2 + \boldsymbol{E}\left[\boldsymbol{U}_2'\boldsymbol{C}\boldsymbol{U}_2\right].$$

Since $\boldsymbol{Z}'\boldsymbol{C}\boldsymbol{Z} = \boldsymbol{Z}'\boldsymbol{Z}$ we have that

$$\boldsymbol{\Pi}_2'\boldsymbol{Z}'\boldsymbol{C}\boldsymbol{Z\Pi}_2 = \boldsymbol{H}_{22}. \tag{A.5}$$

Also,

$$\mathrm{E}\left[\boldsymbol{U}_2'\boldsymbol{C}\boldsymbol{U}_2\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} E\left[U_2'e_ie_i'Ce_je_j'U_2\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}E\left[\boldsymbol{U}_{2i}\boldsymbol{U}_{2j}'\right] = \boldsymbol{O}$$

because the main diagonal elements of $\boldsymbol{C}$ are 0, so $\mathrm{E}\left[\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2\right] = \boldsymbol{H}_{22}$. Further,

$$\begin{aligned}\mathrm{Var}\left[\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2\right] &= \mathrm{E}\left[\left\{(\boldsymbol{Z\Pi}_2)'\boldsymbol{C}\boldsymbol{U}_2 + \boldsymbol{U}_2'\boldsymbol{C}\boldsymbol{Z\Pi}_2 + \boldsymbol{U}_2'\boldsymbol{C}\boldsymbol{U}_2\right\}\left\{(\boldsymbol{Z\Pi}_2)'\boldsymbol{C}\boldsymbol{U}_2 + \boldsymbol{U}_2'\boldsymbol{C}\boldsymbol{Z\Pi}_2 + \boldsymbol{U}_2'\boldsymbol{C}\boldsymbol{U}_2\right\}'\right] \\ &\leq 3\,\mathrm{E}\left[(\boldsymbol{Z\Pi}_2)'\boldsymbol{C}\boldsymbol{U}_2\boldsymbol{U}_2'\boldsymbol{C}\boldsymbol{Z\Pi}_2\right] + 3\,\mathrm{E}\left[\boldsymbol{U}_2'\boldsymbol{C}\boldsymbol{Z\Pi}_2\,(\boldsymbol{Z\Pi}_2)'\boldsymbol{C}\boldsymbol{U}_2\right] + 3\,\mathrm{E}\left[\boldsymbol{U}_2'\boldsymbol{C}\boldsymbol{U}_2\boldsymbol{U}_2'\boldsymbol{C}\boldsymbol{U}_2\right]\end{aligned}$$

by the Cauchy-Schwarz inequality. By Assumption 3 $\mathrm{E}\left[\boldsymbol{U}_2\boldsymbol{U}_2'\right] \leq c_u\boldsymbol{I}_n$ and from the definition of $\boldsymbol{C}$ it holds that (see Bekker and Crudu (2015), p.337)

$$\boldsymbol{Z}'\boldsymbol{C}^2\boldsymbol{Z} = \boldsymbol{Z}'\left\{\boldsymbol{I}_n + \frac{1}{4}\left(\boldsymbol{I}_n - \boldsymbol{D}\right)^{-1}\left(\boldsymbol{I}_n - \boldsymbol{P}\right)\left(\boldsymbol{I}_n - \boldsymbol{D}\right)^{-1}\right\}\boldsymbol{Z}.$$

Further, by Assumption 1 $\left(\boldsymbol{I}_n - \boldsymbol{D}\right)^{-1} \leq c_u\boldsymbol{I}_n$, and therefore,

$$\boldsymbol{Z}'\boldsymbol{C}^2\boldsymbol{Z} \leq c_u\boldsymbol{Z}'\boldsymbol{Z}, \tag{A.6}$$

so the first expectation is

$$\mathrm{E}\left[(\boldsymbol{Z\Pi}_2)'\boldsymbol{C}\boldsymbol{U}_2\boldsymbol{U}_2'\boldsymbol{C}\boldsymbol{Z\Pi}_2\right] \leq c_u\left(\boldsymbol{Z\Pi}_2\right)'\boldsymbol{C}^2\boldsymbol{Z\Pi}_2 \leq c_u\boldsymbol{H}_{22}. \tag{A.7}$$

The second expectation is

$$\mathrm{E}\left[\boldsymbol{U}_2'\boldsymbol{C}\boldsymbol{Z\Pi}_2\left(\boldsymbol{Z\Pi}_2\right)'\boldsymbol{C}\boldsymbol{U}_2\right] = \sum_{i=1}^{n} a_{ii}\,\mathrm{E}\left[\boldsymbol{U}_{2i}\boldsymbol{U}_{2i}'\right],$$

where a_{ii} denotes the i-th main diagonal component of $\boldsymbol{C}\boldsymbol{Z\Pi}_2\left(\boldsymbol{Z\Pi}_2\right)'\boldsymbol{C}$ and $\boldsymbol{U}_{2i}'$ is the i-th row of $\boldsymbol{U}_2$. By Assumption 3 and (A.6) we obtain that

$$\begin{aligned}\mathrm{E}\left[\boldsymbol{U}_2'\boldsymbol{C}\boldsymbol{Z\Pi}_2\left(\boldsymbol{Z\Pi}_2\right)'\boldsymbol{C}\boldsymbol{U}_2\right] &\leq c_u\,\mathrm{tr}\left(\boldsymbol{C}\boldsymbol{Z\Pi}_2\left(\boldsymbol{Z\Pi}_2\right)'\boldsymbol{C}\right)\boldsymbol{I}_{g_2} = c_u\,\mathrm{tr}\left(\left(\boldsymbol{Z\Pi}_2\right)'\boldsymbol{C}^2\boldsymbol{Z\Pi}_2\right)\boldsymbol{I}_{g_2}\\ &\leq c_u\,\mathrm{tr}\left(\boldsymbol{H}_{22}\right)\boldsymbol{I}_{g_2} \leq c_u r_{\max}\boldsymbol{I}_{g_2}.\end{aligned} \tag{A.8}$$

$$\begin{aligned}\mathrm{E}\left[\boldsymbol{U}_2'\boldsymbol{C}\boldsymbol{U}_2\boldsymbol{U}_2'\boldsymbol{C}\boldsymbol{U}_2\right] &= \sum_{i,j,k,\ell}\mathrm{E}\left[\boldsymbol{U}_2'\boldsymbol{e}_i\boldsymbol{e}_i'\boldsymbol{C}\boldsymbol{e}_j\boldsymbol{e}_j'\boldsymbol{U}_2\boldsymbol{U}_2'\boldsymbol{e}_k\boldsymbol{e}_k'\boldsymbol{C}\boldsymbol{e}_\ell\boldsymbol{e}_\ell'\boldsymbol{U}_2\right]\\ &= \sum_{i,j,k,\ell} C_{ij}C_{k\ell}\,\mathrm{E}\left[\boldsymbol{U}_{2i}\boldsymbol{U}_{2j}'\boldsymbol{U}_{2k}\boldsymbol{U}_{2\ell}'\right]\\ &= \sum_{i\neq j} C_{ij}^2\,\mathrm{E}\left[\boldsymbol{U}_{2i}\boldsymbol{U}_{2j}'\boldsymbol{U}_{2i}\boldsymbol{U}_{2j}'\right] + \sum_{i\neq j} C_{ij}^2\,\mathrm{E}\left[\boldsymbol{U}_{2i}\boldsymbol{U}_{2j}'\boldsymbol{U}_{2j}\boldsymbol{U}_{2i}'\right].\end{aligned}$$

By Assumption 3, the Cauchy-Schwarz inequality and (21) we obtain that

$$\mathrm{E}\left[\boldsymbol{U}_2'\boldsymbol{C}\boldsymbol{U}_2\boldsymbol{U}_2'\boldsymbol{C}\boldsymbol{U}_2\right] \leq c_u\left(\sum_{i\neq j} C_{ij}^2\right)\boldsymbol{I}_{g_2} \leq c_u k\boldsymbol{I}_{g_2}. \tag{A.9}$$

By collecting the results from (A.7), (A.8), (A.9) we obtain the first result.

2. The model $\boldsymbol{X}_2 = \boldsymbol{Z\Pi}_2 + \boldsymbol{U}_2$ implies $\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{\varepsilon} = \boldsymbol{\Pi}_2'\boldsymbol{Z}'\boldsymbol{C}\boldsymbol{\varepsilon} + \boldsymbol{U}_2'\boldsymbol{C}\boldsymbol{\varepsilon}$. Similar to part 1., since the main diagonal elements of $\boldsymbol{C}$ are 0, we have $\mathrm{E}\left[\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{\varepsilon}\right] = \boldsymbol{0}$. Regarding the

variance we have

$$\begin{aligned}\operatorname{Var}\left[\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{\varepsilon}\right] &= (\boldsymbol{Z}\boldsymbol{\Pi}_2)'\boldsymbol{C}\,\mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right]\boldsymbol{C}\boldsymbol{Z}\boldsymbol{\Pi}_2 + \mathrm{E}\left[\boldsymbol{U}_2'\boldsymbol{C}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right]\boldsymbol{C}\boldsymbol{Z}\boldsymbol{\Pi}_2 + (\boldsymbol{Z}\boldsymbol{\Pi}_2)'\boldsymbol{C}\,\mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{C}\boldsymbol{U}_2\right] \\ &\quad + \mathrm{E}\left[\boldsymbol{U}_2'\boldsymbol{C}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{C}\boldsymbol{U}_2\right]. \end{aligned} \tag{A.10}$$

By Assumption 3 and (A.6) the first term is

$$\boldsymbol{H}_{22}^{-1/2}(\boldsymbol{Z}\boldsymbol{\Pi}_2)'\boldsymbol{C}\,\mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right]\boldsymbol{C}\boldsymbol{Z}\boldsymbol{\Pi}_2\boldsymbol{H}_{22}^{-1/2} \leq c_u\boldsymbol{H}_{22}^{-1/2}(\boldsymbol{Z}\boldsymbol{\Pi}_2)'\boldsymbol{Z}\boldsymbol{\Pi}_2\boldsymbol{H}_{22}^{-1/2} = c_u\boldsymbol{I}_{g_2}. \tag{A.11}$$

The second and third terms from (A.10) are 0. This is because

$$E\left[\boldsymbol{U}_2'\boldsymbol{C}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right]\boldsymbol{C} = \sum_{i,j,k}\mathrm{E}\left[\boldsymbol{U}_2'\boldsymbol{e}_i\boldsymbol{e}_i'\boldsymbol{C}\boldsymbol{e}_j\boldsymbol{e}_j'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{e}_k\right]\boldsymbol{e}_k'\boldsymbol{C} = \sum_{i,j,k}\mathrm{E}\left[\boldsymbol{U}_{2i}C_{ij}\varepsilon_j\varepsilon_k\right]\boldsymbol{e}_k'\boldsymbol{C}.$$

Since the main diagonal elements of $\boldsymbol{C}$ are 0, the expectations from the above sum are 0. Consequently, $\boldsymbol{C}\,\mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{C}\boldsymbol{U}_2\right] = \boldsymbol{O}$ as well. The fourth term from the expression in (A.10) is

$$\begin{aligned}\mathrm{E}\left[\boldsymbol{U}_2'\boldsymbol{C}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{C}\boldsymbol{U}_2\right] &= \sum_{i,j,k,\ell}\mathrm{E}\left[\boldsymbol{U}_2'\boldsymbol{e}_i\boldsymbol{e}_i'\boldsymbol{C}\boldsymbol{e}_j\boldsymbol{e}_j'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{e}_k\boldsymbol{e}_k'\boldsymbol{C}\boldsymbol{e}_\ell\boldsymbol{e}_\ell'\boldsymbol{U}_2\right] = \sum_{i,j,k,\ell}\mathrm{E}\left[\boldsymbol{U}_{2i}C_{ij}\varepsilon_j\varepsilon_kC_{k\ell}\boldsymbol{U}_{2\ell}'\right] \\ &= \sum_{i\neq j}C_{ij}^2\left(\mathrm{E}\left[\varepsilon_j^2\boldsymbol{U}_{2i}\boldsymbol{U}_{2i}'\right] + \mathrm{E}\left[\varepsilon_i\boldsymbol{U}_{2i}\varepsilon_j\boldsymbol{U}_{2j}'\right]\right) = \sum_{i\neq j}C_{ij}^2\left(\sigma_j^2\boldsymbol{\Sigma}_{22i} + \boldsymbol{\sigma}_{12i}\boldsymbol{\sigma}_{12j}'\right).\end{aligned}$$

By the Cauchy-Schwarz inequality, Assumption 3, and Equation (21) in the main text we obtain that

$$\mathrm{E}\left[\boldsymbol{U}_2'\boldsymbol{C}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{C}\boldsymbol{U}_2\right] \leq c_u\sum_{i\neq j}C_{ij}^2\boldsymbol{I}_{g_2} \leq c_uk\boldsymbol{I}_{g_2}. \tag{A.12}$$

Collecting the results from (A.11) and (A.12), we obtain the result. □

B Auxiliary Results

This Section includes a set of theorems, examples, remarks and propositions associated to the main results of the paper. First we discuss the case when the OLS estimator $\widetilde{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{y}_0$ can be used as a plug-in estimator. This occurs in the practically relevant situation where the null hypothesis contains all parameters corresponding to endogenous variables. The case when there is a single endogenous regressor in the model and the null hypothesis contains exactly its coefficient is a common example.

Theorem B.1. *Suppose that $\boldsymbol{X}_2$ is exogenous, $\boldsymbol{H}_{22}/n = O(1)$ and $(\boldsymbol{X}_2'\boldsymbol{X}_2/n)^{-1} = O_p(1)$. Then under Assumptions 1, 2, 3 we have that $T_2 \to_d \mathcal{N}(0,1)$.*

Proof. We need to show that $\frac{\Delta}{\sqrt{k}} = o_p(1)$ where Δ is defined in Equation (24). First note that $\mathrm{E}\left[\frac{\boldsymbol{X}_2'\boldsymbol{\varepsilon}}{\sqrt{n}}\right] = \boldsymbol{0}$ because $\boldsymbol{X}_2$ is exogenous, and

$$\mathrm{Var}\left[\frac{\boldsymbol{X}_2'\boldsymbol{\varepsilon}}{\sqrt{n}}\right] \leq c_u\frac{\boldsymbol{H}_{22}}{n} + \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2\boldsymbol{\Sigma}_{i22} = O(1)$$

due to Assumption 3 and $\boldsymbol{H}_{22}/n = O(1)$. Consequently,

$$\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\right) = (\boldsymbol{X}_2'\boldsymbol{X}_2/n)^{-1}\left(\boldsymbol{X}_2'\boldsymbol{\varepsilon}/\sqrt{n}\right)^{-1} = O_p(1). \tag{B.1}$$

Using $\boldsymbol{H}_{22}/n = O(1)$ and by Lemma A.4 we have

$$\mathrm{E}\left[\frac{\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2}{n\sqrt{k}}\right] = o(1), \quad \mathrm{Var}\left[\frac{\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2}{n\sqrt{k}}\right] = o(1)$$

and

$$\mathrm{E}\left[\frac{\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{\varepsilon}}{\sqrt{kn}}\right] = \boldsymbol{0}, \quad \mathrm{Var}\left[\frac{\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{\varepsilon}}{\sqrt{kn}}\right] = o(1).$$

Therefore,

$$\frac{1}{\sqrt{k}}\Delta = \frac{1}{\sqrt{k}}\left(\widetilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\right)' \boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2\left(\widetilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\right) - \frac{2}{\sqrt{k}}\left(\widetilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\right)' \boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{\varepsilon} = o_p(1),$$

so B_2 in (23) is $o_p(1)$. B_1 from (23) is $o_p(1)$ due to the consistency of $\widetilde{\boldsymbol{\beta}}_2$ from (B.1) and the result follows. □

The following theorem provides sufficient conditions for the asymptotic distribution of the T_2 statistic when the JIV1 estimator is used as plug-in. Recall that $r_{\min} = \lambda_{\min}(\boldsymbol{H}_{22})$ and $r_{\max} = \lambda_{\max}(\boldsymbol{H}_{22})$.

Theorem B.2. *If Assumptions 1, 2, 3 and $\sqrt{k}/r_{\min} \to 0$, $r_{\max}/k = O(1)$ are satisfied, then the JIV1 estimator $\widetilde{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{y}_0$ is consistent and $T_2 \to_d \mathcal{N}(0,1)$.*

Proof. First we show consistency, that is, $\widetilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2 = (\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{\varepsilon} = o_p(1)$. From Lemma A.4 it follows that

$$\mathrm{E}\left[\boldsymbol{H}_{22}^{-1}\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2\right] = \boldsymbol{I}_{g_2},\ \mathrm{Var}\left[\boldsymbol{H}_{22}^{-1}\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2\right] \le c_u\boldsymbol{H}_{22}^{-1} + c_u k\boldsymbol{H}_{22}^{-2} + c_u r_{\max}\boldsymbol{H}_{22}^{-2}.$$

Since $\boldsymbol{H}_{22}^{-1} \le \frac{1}{r_{\min}}\boldsymbol{I}_{g_2}$ and by assumptions $\dfrac{k}{r_{\min}^2} \to 0$, $r_{\max}/k = O(1)$ we have that $\mathrm{Var}\left[\boldsymbol{H}_{22}^{-1}\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2\right] \to \boldsymbol{O}$, so $\boldsymbol{H}_{22}^{-1}\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2 \to_p \boldsymbol{I}_{g_2}$, and therefore, $\left(\boldsymbol{H}_{22}^{-1}\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2\right)^{-1} = O_p(1)$. From Lemma A.4 it also follows that

$$\mathrm{E}\left[\boldsymbol{H}_{22}^{-1}\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{\varepsilon}\right] = \boldsymbol{0},\ \mathrm{Var}\left[\boldsymbol{H}_{22}^{-1}\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{\varepsilon}\right] \le c_u\boldsymbol{H}_{22}^{-1} + c_u k\boldsymbol{H}_{22}^{-2}.$$

This variance goes to $\boldsymbol{O}$ for the same reason as above, so $\boldsymbol{H}_{22}^{-1}\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{\varepsilon} = o_p(1)$. Therefore, $\widetilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2 = o_p(1)$.

Let now

$$\widehat{V}(\boldsymbol{\beta}) = \frac{2}{k}\boldsymbol{\varepsilon}^{(2)\prime}\boldsymbol{C}^{(2)}\boldsymbol{\varepsilon}^{(2)}, \quad \boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}.$$

Note that Δ in (23) now is

$$\Delta = -\boldsymbol{\varepsilon}'\boldsymbol{C}\boldsymbol{X}_2 \left(\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2\right)^{-1} \boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{\varepsilon}.$$

This can be written as

$$\Delta = -\boldsymbol{\varepsilon}'\boldsymbol{C}\boldsymbol{X}_2\boldsymbol{H}_{22}^{-1/2} \left(\boldsymbol{H}_{22}^{-1/2}\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2\boldsymbol{H}_{22}^{-1/2}\right)^{-1} \boldsymbol{H}_{22}^{-1/2}\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{\varepsilon}. \tag{B.2}$$

From Lemma A.4 we know that

$$\mathrm{E}\left[\boldsymbol{H}_{22}^{-1/2}\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2\boldsymbol{H}_{22}^{-1/2}\right] = \boldsymbol{I}_{g_2},$$

$$\mathrm{Var}\left[\boldsymbol{H}_{22}^{-1/2}\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2\boldsymbol{H}_{22}^{-1/2}\right] \leq c_u\boldsymbol{H}_{22}^{-1} + c_u k\boldsymbol{H}_{22}^{-2} + c_u r_{\max}\boldsymbol{H}_{22}^{-2}.$$

Since $\boldsymbol{H}_{22}^{-1} \leq \frac{1}{r_{\min}} I_{g_2}$ and by assumptions $\dfrac{k}{r_{\min}^2} \to 0$, $r_{\max}/k = O(1)$ we have that

$$\mathrm{Var}\left[\boldsymbol{H}_{22}^{-1/2}\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2\boldsymbol{H}_{22}^{-1/2}\right] \to \boldsymbol{O},$$

so $\boldsymbol{H}_{22}^{-1/2}\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2\boldsymbol{H}_{22}^{-1/2} \to_p \boldsymbol{I}_{g_2}$. Consequently, $\left(\boldsymbol{H}_{22}^{-1/2}\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2\boldsymbol{H}_{22}^{-1/2}\right)^{-1} = O_p(1)$. By Lemma A.4 we know that

$$\mathrm{E}\left[\boldsymbol{H}_{22}^{-1/2}\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{\varepsilon}\right] = \boldsymbol{0}. \tag{B.3}$$

Next we show that under $\dfrac{\sqrt{k}}{r_{\min}} \to 0$ it holds that

$$\frac{1}{\sqrt{k}}\mathrm{Var}\left(\boldsymbol{H}_{22}^{-1/2}\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{\varepsilon}\right) = o(1). \tag{B.4}$$

From Lemma A.4 we know that

$$\frac{1}{\sqrt{k}}\mathrm{Var}\left[\boldsymbol{H}_{22}^{-1/2}\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{\varepsilon}\right] \leq \frac{c_u}{\sqrt{k}}\boldsymbol{I}_{g_2} + c_u\sqrt{k}\boldsymbol{H}_{22}^{-1} \leq \left(\frac{c_u}{\sqrt{k}} + c_u\frac{\sqrt{k}}{r_{\min}}\right)\boldsymbol{I}_{g_2},$$

where the last inequality follows from $\boldsymbol{H}_{22}^{-1} \leq \frac{1}{r_{\min}} \boldsymbol{I}_{g_2}$. So (B.4) holds under $k \to \infty$, $\frac{\sqrt{k}}{r_{\min}} \to 0$, and therefore, taking also (B.3) into account we obtain

$$\frac{1}{k^{1/4}} \boldsymbol{H}_{22}^{-1/2} \boldsymbol{X}_2' \boldsymbol{C} \boldsymbol{\varepsilon} = o_p(1).$$

Consequently,

$$B_2 = \frac{1}{\sqrt{k}} \frac{\Delta}{\sqrt{\widehat{V}\left(\widetilde{\boldsymbol{\beta}}\right)}} = \left(\frac{1}{k^{1/4}} \boldsymbol{H}_{22}^{-1/2} \boldsymbol{X}_2' \boldsymbol{C} \boldsymbol{\varepsilon} \right)' \frac{1}{k^{1/4}} \boldsymbol{H}_{22}^{-1/2} \boldsymbol{X}_2' \boldsymbol{C} \boldsymbol{\varepsilon} \cdot O_p(1) = o_p(1).$$

□

This result is not very different from Theorem 2 in the main text, but it is useful because, on the one hand, the convenient expression of the JIV1 estimator allows us to explain why underrejection of the null hypothesis occurs in most cases.[1] On the other hand, this result allows us to better link the weak instrument assumption $\sqrt{k}/r_{\min} \to 0$ to the asymptotic distribution of T_2. Specifically, the proof of this result suggests that the assumption $\sqrt{k}/r_{\min} \to 0$ appears to be necessary for the asymptotic standard normality of the statistic T_2.[2]

Derivation of Example 1. Suppose that there are ℓ groups with group g having n_g observations and

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{\iota}_{n_1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \boldsymbol{\iota}_{n_\ell} \end{pmatrix},$$

[1] We discuss this in more detail in Remark B.1 below.
[2] See also Remark B.2.

where $\boldsymbol{\iota}_m$ is an $m \times 1$ vector of ones. In this case

$$\boldsymbol{P} = \begin{pmatrix} \frac{1}{n_1}\boldsymbol{\iota}_{n_1}\boldsymbol{\iota}'_{n_1} & & \boldsymbol{O} \\ & \ddots & \\ \boldsymbol{O} & & \frac{1}{n_\ell}\boldsymbol{\iota}_{n_\ell}\boldsymbol{\iota}'_{n_\ell} \end{pmatrix}.$$

The expression

$$E_2 = \frac{1}{\sqrt{k}} \sum_{i=1}^{n} \left(P_{ii} - \frac{k}{n} \right) \varepsilon_i^2$$

from (17) can be written as

$$E_2 = \frac{1}{\sqrt{\ell}} \sum_{g=1}^{\ell} \sum_{i \in G_g} \left(\frac{1}{n_g} - \frac{\ell}{n} \right) \varepsilon_i^2,$$

where G_g is the set of observations belonging to group g.

Suppose now that the groups have either 2 or 3 observations. In this case

$$\begin{aligned} E_2 &= \frac{1}{\sqrt{\ell}} \sum_{g:n_g=2} \sum_{i \in G_g} \left(\frac{1}{2} - \frac{\ell}{n} \right) \varepsilon_i^2 + \frac{1}{\sqrt{\ell}} \sum_{g:n_g=3} \sum_{i \in G_g} \left(\frac{1}{3} - \frac{\ell}{n} \right) \varepsilon_i^2 \\ &= \left(\frac{1}{2} - \frac{\ell}{n} \right) \frac{1}{\sqrt{\ell}} \sum_{g:n_g=2} \sum_{i \in G_g} \varepsilon_i^2 + \left(\frac{1}{3} - \frac{\ell}{n} \right) \frac{1}{\sqrt{\ell}} \sum_{g:n_g=3} \sum_{i \in G_g} \varepsilon_i^2. \end{aligned}$$

Suppose homoskedasticity with $\mathrm{E}[\varepsilon_i^2] = \sigma^2$ and let ℓ_2 and ℓ_3 denote the number of 2-observation and 3-observation groups, respectively. In this case

$$E_2 = \left(\frac{1}{2} - \frac{\ell}{n} \right) \frac{2\ell_2}{\sqrt{\ell}} \frac{\sum_{g:n_g=2} \sum_{i \in G_g} \varepsilon_i^2}{2\ell_2} + \left(\frac{1}{3} - \frac{\ell}{n} \right) \frac{3\ell_3}{\sqrt{\ell}} \frac{\sum_{g:n_g=3} \sum_{i \in G_g} \varepsilon_i^2}{3\ell_3}$$

Note that $\ell = \ell_2 + \ell_3$ and $n = 2\ell_2 + 3\ell_3$, so

$$\begin{aligned} E_2 &= \frac{\ell_3}{2\ell_2 + 3\ell_3} \frac{\ell_2}{\sqrt{\ell}} \frac{\sum_{g:n_g=2} \sum_{i \in G_g} \varepsilon_i^2}{2\ell_2} - \frac{\ell_2}{2\ell_2 + 3\ell_3} \frac{\ell_3}{\sqrt{\ell}} \frac{\sum_{g:n_g=3} \sum_{i \in G_g} \varepsilon_i^2}{3\ell_3} \\ &= \frac{\ell_2 \ell_3}{\ell n} \sqrt{\ell} \left(\frac{\sum_{g:n_g=2} \sum_{i \in G_g} \varepsilon_i^2}{2\ell_2} - \sigma^2 - \left[\frac{\sum_{g:n_g=3} \sum_{i \in G_g} \varepsilon_i^2}{3\ell_3} - \sigma^2 \right] \right). \end{aligned} \tag{B.5}$$

By the CLT for IID observations

$$\sqrt{2\ell_2}\left(\frac{\sum_{g:n_g=2}\sum_{i\in G_g}\varepsilon_i^2}{2\ell_2}-\sigma^2\right)\xrightarrow{d}\mathcal{N}(0,v)\quad\text{and}$$
$$\sqrt{3\ell_3}\left(\frac{\sum_{g:n_g=3}\sum_{i\in G_g}\varepsilon_i^2}{3\ell_3}-\sigma^2\right)\xrightarrow{d}\mathcal{N}(0,v),$$

where $v=\mathrm{Var}[\varepsilon_i^2]$. The limit $\frac{\ell}{n}\to\lambda\in(0,1)$ implies that $\frac{\ell}{2\ell_2}\to\frac{\lambda}{6\lambda-2}$ and $\frac{\ell}{3\ell_3}\to\frac{\lambda}{3-6\lambda}$, so we obtain

$$\sqrt{\ell}\left(\frac{\sum_{g:n_g=2}\sum_{i\in G_g}\varepsilon_i^2}{2\ell_2}-\sigma^2\right)\xrightarrow{d}\mathcal{N}\left(0,\frac{\lambda}{6\lambda-2}v\right)\quad\text{and}$$
$$\sqrt{\ell}\left(\frac{\sum_{g:n_g=3}\sum_{i\in G_g}\varepsilon_i^2}{3\ell_3}-\sigma^2\right)\xrightarrow{d}\mathcal{N}\left(0,\frac{\lambda}{3-6\lambda}v\right).$$

Therefore, from (B.5) we obtain

$$E_2\xrightarrow{d}\mathcal{N}\left(0,\frac{(3\lambda-1)(1-2\lambda)}{6\lambda}v\right).$$

Since its variance does not vanish in the limit, E_2 will not converge to 0 in probability. □

Example B.1. *In this example we consider the indicator instruments discussed in Example 1 under heteroskedasticity when there are only groups of size 2 and 3, and we study whether E_2 defined in B.5 has convergent or divergent mean. That is, we study the limit of*

$$\mathrm{E}[E_2]=\frac{1}{\sqrt{\ell}}\frac{\ell_2\ell_3}{n}\left(\frac{\sum_{g:n_g=2}\sum_{i\in G_g}\sigma_i^2}{2\ell_2}-\frac{\sum_{g:n_g=3}\sum_{i\in G_g}\sigma_i^2}{3\ell_3}\right)$$

with respect to the growth rate of ℓ_2 and ℓ_3. First note that the assumption $\frac{1}{k}\sum_i\left(P_{ii}-\frac{k}{n}\right)^2\to 0$ from Proposition 1 is equivalent to

$$\frac{1}{\ell}\left(\sum_{g:n_g=2}\sum_{i\in G_g}\left(\frac{1}{2}-\frac{\ell}{n}\right)^2+\sum_{g:n_g=3}\sum_{i\in G_g}\left(\frac{1}{3}-\frac{\ell}{n}\right)^2\right)\to 0.$$

Further, since

$$\sum_{g:n_g=2}\sum_{i\in G_g}\left(\frac{1}{2}-\frac{\ell}{n}\right)^2+\sum_{g:n_g=3}\sum_{i\in G_g}\left(\frac{1}{3}-\frac{\ell}{n}\right)^2=\frac{\ell_2\ell_3}{6n},$$

this is equivalent to

$$\frac{\ell_2\ell_3}{\ell n}\to 0.$$

Recalling that $\ell=\ell_2+\ell_3$ *and* $n=2\ell_2+3\ell_3$*, we conclude that this can only happen if either* $\ell_2/\ell_3\to 0$ *or* $\ell_3/\ell_2\to 0$*. Suppose* $\ell_3/\ell_2\to 0$*, which implies* $\ell_2\to\infty$*.*

Suppose that the variance averages $\frac{\sum_{g:n_g=2}\sum_{i\in G_g}\sigma_i^2}{2\ell_2}$ *and* $\frac{\sum_{g:n_g=3}\sum_{i\in G_g}\sigma_i^2}{3\ell_3}$ *converge as* $n\to\infty$*; let*

$$\overline{\sigma}_2^2=\lim_{n\to\infty}\frac{\sum_{g:n_g=2}\sum_{i\in G_g}\sigma_i^2}{2\ell_2},\quad \overline{\sigma}_3^2=\lim_{n\to\infty}\frac{\sum_{g:n_g=3}\sum_{i\in G_g}\sigma_i^2}{3\ell_3}.$$

Note that

$$\frac{1}{\sqrt{\ell}}\frac{\ell_2\ell_3}{2\ell_2+3\ell_3}=\frac{1}{\sqrt{\ell_2+\ell_3}}\frac{\ell_2\ell_3}{2\ell_2+3\ell_3}=\frac{1}{\sqrt{1+\ell_3/\ell_2}}\frac{\ell_3/\sqrt{\ell_2}}{2+3\ell_3/\ell_2}.$$

Therefore, if $\ell_3/\sqrt{\ell_2}\to 0$ *then* $\frac{1}{\sqrt{\ell}}\frac{\ell_2\ell_3}{2\ell_2+3\ell_3}\to 0$*. In this case*

$$\mathrm{E}\left[E_2\right]\to 0.$$

If $\ell_3/\sqrt{\ell_2}\to\infty$ *then* $\frac{1}{\sqrt{\ell}}\frac{\ell_2\ell_3}{2\ell_2+3\ell_3}\to\infty$*. In this case* $\mathrm{E}\left[E_2\right]$ *can be unbounded; specifically*

$$\mathrm{E}\left[E_2\right]\to\begin{cases}\infty & \text{if } \overline{\sigma}_2^2>\overline{\sigma}_3^2,\\ -\infty & \text{if } \overline{\sigma}_2^2<\overline{\sigma}_3^2,\end{cases}$$

and therefore, E_2 *is not bounded in probability. Consequently, the statistic* AR_{AG} *is not bounded in probability. We summarize our findings in the following.*

Proposition B.1. *Suppose that $\ell_3/\ell_2 \to 0$ and that the variance averages $\frac{\sum_{g:n_g=2}\sum_{i\in G_g}\sigma_i^2}{2\ell_2}$ and $\frac{\sum_{g:n_g=3}\sum_{i\in G_g}\sigma_i^2}{3\ell_3}$ converge to $\overline{\sigma}_2^2$ and $\overline{\sigma}_3^2$, respectively. Then, if $\ell_3/\sqrt{\ell_2} \to 0$, $E_2 = o_p(1)$; if $\ell_3/\sqrt{\ell_2} \to \infty$ and $\overline{\sigma}_2^2 \neq \overline{\sigma}_3^2$, E_2 is not bounded in probability.*

Remark B.1. *When the plug-in is the JIV1 estimator $\widetilde{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{y}_0$ we obtain that*

$$T_2 = \frac{1}{\sqrt{k}} \frac{\boldsymbol{\varepsilon}'\boldsymbol{C}\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}'\boldsymbol{C}\boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{\varepsilon}}{\sqrt{\widehat{V}\left(\widetilde{\boldsymbol{\beta}}\right)}}. \tag{B.6}$$

The formula in (B.6) suggests that T_2 is more likely to be negative than positive in finite samples, which may explain the underrejection results in our Monte Carlo simulations. See, e.g., Figures C.7 and C.8 in Section B. Indeed, we know that $\mathrm{E}[\boldsymbol{\varepsilon}'\boldsymbol{C}\boldsymbol{\varepsilon}] = 0$ and since $\boldsymbol{H}_{22}^{-1/2}\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2\boldsymbol{H}_{22}^{-1/2} \to_p \boldsymbol{I}_{g_2}$, $\left(\boldsymbol{H}_{22}^{-1/2}\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2\boldsymbol{H}_{22}^{-1/2}\right)^{-1}$ is likely to be positive definite in sufficiently large finite samples. Therefore, $\boldsymbol{\varepsilon}'\boldsymbol{C}\boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{C}\boldsymbol{\varepsilon} \geq 0$, so the numerator of T_2 is more likely to take negative values, unless the sign of $\boldsymbol{\varepsilon}'\boldsymbol{C}\boldsymbol{\varepsilon}$ interacts with the magnitude of $\widehat{V}\left(\widetilde{\boldsymbol{\beta}}\right)$ in a special way. This suggests that the density of T_2 is shifted to the left, which leads to underrejection.

Remark B.2. *The assumption $\frac{\sqrt{k}}{r_{\min}} \to 0$ in Theorem B.2 appears to be necessary. Suppose that this assumption is violated while Assumptions 1, 2, 3 hold; for simplicity consider the case when $g_2 = 1$ and denote the single endogenous variable as $\boldsymbol{x}_2$. Moreover, $\boldsymbol{x}_2 = \boldsymbol{Z}\boldsymbol{\pi}_2 + \boldsymbol{u}_2$. In this case $r = r_{\min} = H_{22}$ and suppose that $\sqrt{k}/r = \tau_n$ with $\tau_n \geq c_\tau > 0$ for any n. One important special case is when τ_n converges to a positive number; another special case is when τ_n goes to ∞.*

1. *We note first that in this case the JIV1-type estimator $\widetilde{\beta}_2 = \beta_2 + (\boldsymbol{x}_2'\boldsymbol{C}\boldsymbol{x}_2)^{-1}\boldsymbol{x}_2'\boldsymbol{C}\boldsymbol{\varepsilon}$ is not consistent. Indeed, by (A.4)*

$$\frac{1}{\sqrt{k}}\boldsymbol{x}_2'\boldsymbol{C}\boldsymbol{x}_2 = \frac{1}{\sqrt{k}}(\boldsymbol{Z}\boldsymbol{\pi}_2)'\boldsymbol{C}\boldsymbol{Z}\boldsymbol{\pi}_2 + \frac{1}{\sqrt{k}}(\boldsymbol{Z}\boldsymbol{\pi}_2)'\boldsymbol{C}\boldsymbol{u}_2 + \frac{1}{\sqrt{k}}\boldsymbol{u}_2'\boldsymbol{C}\boldsymbol{Z}\boldsymbol{\pi}_2 + \frac{1}{\sqrt{k}}\boldsymbol{u}_2'\boldsymbol{C}\boldsymbol{u}_2.$$

Note that

$$\mathrm{E}\left[\frac{1}{\sqrt{k}}\left(\boldsymbol{Z}\boldsymbol{\pi}_2\right)'\boldsymbol{C}\boldsymbol{u}_2\right]=0,$$
$$\mathrm{Var}\left[\frac{1}{\sqrt{k}}\left(\boldsymbol{Z}\boldsymbol{\pi}_2\right)'\boldsymbol{C}\boldsymbol{u}_2\right]=\frac{1}{k}\mathrm{E}\left[\left(\boldsymbol{Z}\boldsymbol{\pi}_2\right)'\boldsymbol{C}\boldsymbol{u}_2\boldsymbol{u}_2'\boldsymbol{C}\boldsymbol{Z}\boldsymbol{\pi}_2\right]\leq\frac{c_u}{k}r\leq\frac{c_u}{c_\tau\sqrt{k}},$$

where the first inequality follows from (A.7), so

$$\frac{1}{\sqrt{k}}\boldsymbol{x}_2'\boldsymbol{C}\boldsymbol{x}_2=\frac{1}{\tau_n}+\frac{1}{\sqrt{k}}\boldsymbol{u}_2'\boldsymbol{C}\boldsymbol{u}_2+o_p\left(1\right).$$

Since by Lemma A.1 $\frac{1}{\sqrt{k}}\boldsymbol{u}_2'\boldsymbol{C}\boldsymbol{u}_2\rightarrow_d N_1$ *with* N_1 *normal with mean* 0*, we obtain that*

$$\frac{1}{\sqrt{k}}\boldsymbol{x}_2'\boldsymbol{C}\boldsymbol{x}_2=\frac{1}{\tau_n}+N_1+o_p\left(1\right). \tag{B.7}$$

Further,

$$\frac{1}{\sqrt{k}}\boldsymbol{x}_2'\boldsymbol{C}\boldsymbol{\varepsilon}=\frac{1}{\sqrt{k}}\left(\boldsymbol{Z}\boldsymbol{\pi}_2\right)'\boldsymbol{C}\boldsymbol{\varepsilon}+\frac{1}{\sqrt{k}}\boldsymbol{u}_2'\boldsymbol{C}\boldsymbol{\varepsilon},$$

where the first term is $o_p\left(1\right)$ *for similar reasons as above and the second term is asymptotically normal with mean* 0 *(from Lemma A2 in Chao* et al.*, 2012). Therefore, we can write*

$$\frac{1}{\sqrt{k}}\boldsymbol{x}_2'\boldsymbol{C}\boldsymbol{\varepsilon}=N_2+o_p\left(1\right), \tag{B.8}$$

with N_2 *normal with mean* 0*. So, from (B.7) and (B.8)*

$$\widetilde{\beta}_2-\beta_2=\frac{N_2+o_p\left(1\right)}{1/\lambda_n+N_1+o_p\left(1\right)},$$

which in general is not $o_p\left(1\right)$*, so* $\widetilde{\beta}_2$ *is not consistent. Therefore, we cannot prove that* $\widehat{V}\left(\widetilde{\boldsymbol{\beta}}\right)-\widehat{V}\left(\boldsymbol{\beta}\right)=o_p\left(1\right)$ *in the way we do above (Lemmas A.2 and A.3).*

2. *In order to derive the limit of* T_2 *in Theorem B.2 we proved that* $\frac{1}{\sqrt{k}}\boldsymbol{\varepsilon}'\boldsymbol{C}\boldsymbol{x}_2(\boldsymbol{x}_2'\boldsymbol{C}\boldsymbol{x}_2)^{-1}\boldsymbol{x}_2'\boldsymbol{C}\boldsymbol{\varepsilon} = o_p(1)$. *In this case (B.7) and (B.8) imply*

$$\frac{1}{\sqrt{k}}\boldsymbol{\varepsilon}'\boldsymbol{C}\boldsymbol{x}_2(\boldsymbol{x}_2'\boldsymbol{C}\boldsymbol{x}_2)^{-1}\boldsymbol{x}_2'\boldsymbol{C}\boldsymbol{\varepsilon} = \left(\frac{1}{\sqrt{k}}\boldsymbol{\varepsilon}'\boldsymbol{C}\boldsymbol{x}_2\right)^2\left(\frac{1}{\sqrt{k}}\boldsymbol{x}_2'\boldsymbol{C}\boldsymbol{x}_2\right)^{-1} = \frac{(N_2 + o_p(1))^2}{1/\lambda_n + N_1 + o_p(1)},$$

which is not $o_p(1)$ *in general.*

Example B.2. *This example is motivated by the fact that, in practice, applied researchers may erroneously choose an inconsistent plug-in estimator. It is reasonable to think that such a choice may affect the behaviour of* T_2. *In order to simplify the analysis we assume that the plug-in is consistent but it converges at an arbitrary slow rate to the true value. The assumption of consistency allows us to use Lemma A.3. Let us consider a simple two-regressor model*

$$\boldsymbol{y} = \boldsymbol{x}_1\beta_1 + \boldsymbol{x}_2\beta_2 + \boldsymbol{\varepsilon}$$

where $\boldsymbol{x}_1$ *and* $\boldsymbol{x}_2$ *may both be endogenous and suppose that we want to test the following null* $H_0 : \beta_1 = \beta_{10}$. *Let us assume that there exists an estimator for* β_2, *say* $\widetilde{\beta}_2$, *such that* $\sqrt{a_n}(\widetilde{\beta}_2 - \beta_2) = O_p(1)$ *where* $a_n \to \infty$ *as* $n \to \infty$. *This situation defines a consistent but potentially slowly converging estimator. Let us also define*

$$\boldsymbol{x}_2 = \boldsymbol{Z}\boldsymbol{\pi}_2 + \boldsymbol{u}_2$$

with $H_{22} = \boldsymbol{\pi}_2'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\pi}_2$, *in this case* $r = r_{\min} = H_{22}$ *and* $r \to \infty$ *as* $n \to \infty$. *Let us suppose that* a_n *and* r *diverge to infinity possibly at different rates. If we assume that* $\widetilde{\boldsymbol{\beta}} = (\beta_{10}, \widetilde{\beta}_2)'$ *is consistent we can use Lemmas A.2 and A.3. From Equation (23) in the main text we notice that the fact that* T_2 *converges to a standard normal would now depend only on the behaviour of* Δ *(see Equation (24) in the main text), which in this case is*

$$\Delta = \left(\widetilde{\beta}_2 - \beta_2\right)\boldsymbol{x}_2'\boldsymbol{C}\boldsymbol{x}_2\left(\widetilde{\beta}_2 - \beta_2\right) - 2\left(\widetilde{\beta}_2 - \beta_2\right)\boldsymbol{x}_2'\boldsymbol{C}\boldsymbol{\varepsilon}.$$

Since $\frac{\boldsymbol{x}_2'\boldsymbol{C}\boldsymbol{x}_2}{r} \to_p 1$, *we get*

$$\frac{1}{\sqrt{k}}\left(\widetilde{\beta}_2-\beta_2\right)\boldsymbol{x}_2'\boldsymbol{C}\boldsymbol{x}_2\left(\widetilde{\beta}_2-\beta_2\right)=\frac{1}{\sqrt{k}}\frac{r}{a_n}\sqrt{a_n}\left(\widetilde{\beta}_2-\beta_2\right)\frac{\boldsymbol{x}_2'\boldsymbol{C}\boldsymbol{x}_2}{r}\sqrt{a_n}\left(\widetilde{\beta}_2-\beta_2\right)=O_p\left(\frac{r}{a_n\sqrt{k}}\right). \tag{B.9}$$

Moreover,

$$\frac{1}{\sqrt{k}}\left(\widetilde{\beta}_2-\beta_2\right)\boldsymbol{x}_2'\boldsymbol{C}\boldsymbol{\varepsilon}=\frac{\sqrt{r}}{\sqrt{ka_n}}\sqrt{a_n}\left(\widetilde{\beta}_2-\beta_2\right)\frac{\boldsymbol{x}_2'\boldsymbol{C}\boldsymbol{\varepsilon}}{\sqrt{r}}=o_p\left(\frac{\sqrt{r}}{\sqrt{a_n}k^{1/4}}\right). \tag{B.10}$$

This means that if $\frac{r}{a_n\sqrt{k}}$ *diverges,* Δ *does not go to zero and* T_2 *would not converge to a standard normal. Notice that* Δ *does not go to zero when* a_n *grows slower or at the same rate of the boundary condition* $r/\sqrt{k}$*. In this case the distribution of* T_2 *will be shifted to the right causing the test to overreject.*

C Monte Carlo Experiments

This Section collects some complementary Monte Carlo results on the finite sample properties of T_1 and T_2. The simulations consider two DGPs and both the homoskedastic and heteroskedastic case. Apart from the T_1 and T_2 statistics we include the AR_{AG} test of Anatolyev and Gospodinov (2011) and three test statistics due to Bun *et al.* (2018)

$$\begin{aligned}\widehat{AR} &= n\widehat{g}(\boldsymbol{\beta})'\widehat{\boldsymbol{\Omega}}(\boldsymbol{\beta})^{-1}\widehat{g}(\boldsymbol{\beta})\\ \widetilde{AR} &= n\widehat{g}(\boldsymbol{\beta})'\widetilde{\boldsymbol{\Omega}}(\boldsymbol{\beta})^{-1}\widehat{g}(\boldsymbol{\beta})\\ \widetilde{AR}_{df} &= n\widehat{g}(\boldsymbol{\beta})'\widetilde{\boldsymbol{\Omega}}_{df}(\boldsymbol{\beta})^{-1}\widehat{g}(\boldsymbol{\beta})\end{aligned}$$

where $\widehat{\boldsymbol{\Omega}}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{g}(\boldsymbol{\beta})\boldsymbol{g}(\boldsymbol{\beta})'$, $\widetilde{\boldsymbol{\Omega}}(\boldsymbol{\beta}) = \widehat{\boldsymbol{\Omega}}(\boldsymbol{\beta}) - \widehat{g}(\boldsymbol{\beta})\widehat{g}(\boldsymbol{\beta})'$ and $\widetilde{\boldsymbol{\Omega}}_{df}(\boldsymbol{\beta}) = \frac{n}{n-k}\widetilde{\boldsymbol{\Omega}}(\boldsymbol{\beta})$. In our case the moment condition model is defined as $\boldsymbol{g}_i(\boldsymbol{\beta}) = \boldsymbol{Z}_i(y_i - \boldsymbol{X}_i'\boldsymbol{\beta})$ and $\widehat{g}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{g}_i(\boldsymbol{\beta})$. The evaluation of the performance is made in terms of size and power. Furthermore, the second DGP is also used to assess the quality of the asymptotic approx-

imations as presented in Corollary 1 in the main text.

The first DGP (DGP I) is similar to Bekker and Van der Ploeg (2005) where the instruments are dummies. In this experiment the observations are stratified in k groups where each group contains n_j observations and $n = \sum_{j=1}^{k} n_j$ and each group contains a different number of observations. Let us define the model

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{x}\beta + \boldsymbol{\varepsilon} \\ \boldsymbol{x} &= \boldsymbol{Z}\boldsymbol{\pi} + \boldsymbol{v} \end{aligned} \tag{C.1}$$

where the true value of β is zero and $\boldsymbol{Z}$ is a $n \times k$ matrix of dummy variables, such that each of its rows is a versor. Moreover, for each group, the disturbances are jointly normally distributed with zero mean and variance covariance matrix equal to

$$\boldsymbol{\Sigma}_j = \begin{pmatrix} \sigma_j^2 & \rho\sigma_j\sigma_{vj} \\ \rho\sigma_j\sigma_{vj} & \sigma_{vj} \end{pmatrix}, \quad j = 1, \ldots, k.$$

We choose $\rho = 0.5$ and $(k, n) \in \{(7, 146), (40, 140), (60, 168)\}$. The parameters σ_j and σ_{vj} are sampled independently from a uniform distribution $\mathcal{U}(0.5, 1)$. We consider both the homoskedastic case where $\boldsymbol{\Sigma}_j$ is the same for any j and the corresponding heteroskedastic case. Furthermore, the elements of $\boldsymbol{\pi}$ are sampled from $\mathcal{U}(0.05, 0.1)$. The experiment is replicated 5000 times.

The second DGP (DGP II) (Hausman *et al.*, 2012) is given by

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{\iota}\gamma + \boldsymbol{x}\beta + \boldsymbol{\varepsilon} \\ \boldsymbol{x} &= \boldsymbol{z}\pi + \boldsymbol{v} \end{aligned} \tag{C.2}$$

where $\gamma = \beta = 1$, while $\pi = 0.1$ in the analysis of size and $\pi \in \{0.1, 1\}$ in the analysis of power. The sample size is $n = 800$, $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0},\ \boldsymbol{I}_n)$ and independently $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}, 0.1^2 \times \boldsymbol{I}_n)$.

The disturbances vector $\boldsymbol{\varepsilon}$ is generated as

$$\boldsymbol{\varepsilon} = \rho \boldsymbol{v} + \sqrt{\frac{1-\rho^2}{\phi^2+\psi^4}}(\phi \boldsymbol{w}_1 + \psi \boldsymbol{w}_2), \tag{C.3}$$

where $\rho = 0.3$, $\psi = 0.86$ and conditional on $\boldsymbol{z}$, independent of $\boldsymbol{v}$, $\boldsymbol{w}_1 \sim \mathcal{N}(\boldsymbol{0},\ \mathrm{Diag}(\boldsymbol{z})^2)$ where $\mathrm{Diag}(\boldsymbol{z})$ is a diagonal matrix where the diagonal elements are the elements of $\boldsymbol{z}$ and $\boldsymbol{w}_2 \sim \mathcal{N}(\boldsymbol{0},\ \psi^2 \boldsymbol{I}_n)$. Moreover, $\phi \in \{0, 1.38072\}$, where $\phi = 0$ is the homoskedastic case. The instrument matrix $\boldsymbol{Z}$ is given by matrices with rows $(1, z_i, z_i^2, z_i^3, z_i^4)$ and $(1, z_i, z_i^2, z_i^3, z_i^4, z_i b_{1i}, \ldots, z_i b_{\ell i})$, $\ell = 95, 695$, where, independent of other random variables, the elements $b_{1i}, \ldots, b_{\ell i}$ are i.i.d. Bernoulli distributed with $p = 1/2$.[3] We consider also two rather extreme situations: $k = 2$ and $k = 700$. We replicate our experiments 5000 times. When using the T_1 test and the T_2 test we consider $H_0 : (\gamma, \beta)' = (1, 1)'$ and $H_0 : \beta = 1$ respectively.

C.1 Simulation results

We first discuss the quality of the potential approximations for T_1 and T_2 when $k = 2$, then we provide some interpretation of the simulations by separately analysing the results on size and power. We also discuss the behaviour of T_2 when an inconsistent plug-in is used.

Approximations. In Figure C.1, we explore the behaviour of T_1 for $k = 2$ and $n = 50, 100, 200, 400, 800$. It seems clear that, in this case, the chi square approximation for T_1 is more accurate than its Gaussian counterpart. This result is less evident in the case of T_2, since, as shown in Corollary 1, three alternative chi square approximations are available. Nonetheless, Figure C.2 panel (c) suggests that result (iii) in Corollary 1 may cause the test to reject too often. On the other hand, the approximations in (ii) and (iv) of Corollary 1 deliver more reliable results (Figure C.2 panels (b) and (d)).

Size. In the case of DGP I (Figure C.3 and Figure C.4), heteroskedasticity is rather mild and, as expected, the various statistics perform quite similarly in the homoskedastic

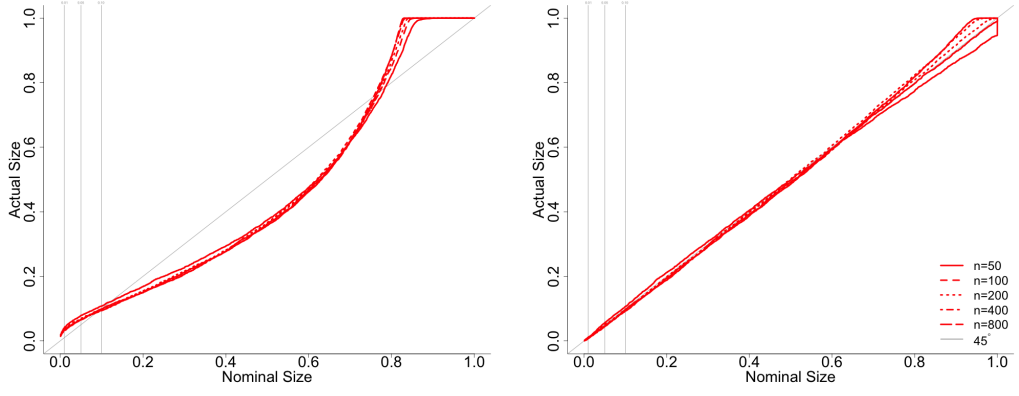
[3] The same set of instruments is used throughout the various repetitions.

and heteroskedastic case. In addition, we observe that T_1, AR_{AG} and $\widetilde{AR}_{df}$ work well for the three combinations of k and n considered. On the other hand, $\widehat{AR}$ tends to underreject as $\frac{k}{n}$ gets larger, while $\widetilde{AR}$ tends to underreject. The case of DGP II (Figure C.5 to Figure C.8) is more complex, as the type of heteroskedasticity introduced in the model may have a non trivial impact on the performance of the tests. In general, we observe that T_1 and T_2 work well in all the considered cases and AR_{AG} performs well in general under homoskedasticity and, as it is expected, it shows some tendency to overreject when $k = 700$.[4] As we introduce heteroskedasticity, the performance of the AR_{AG} test dramatically deteriorates. The tests introduced in Bun *et al.* (2018) work well for most of the cases but tend to either underreject ($\widehat{AR}$, $\widetilde{AR}$) or overreject ($\widetilde{AR}_{df}$) when k is large.

Power. Under homoskedasticity and k small the tests are indistinguishable (Figure C.9 to Figure C.12 panels (a) and (b)). The picture gets more complicated as k increases. In particular, with $\pi = 1$ all the test apart from $\widetilde{AR}$ can control size and have excellent power properties (Figure C.11 and Figure C.12 panel (c)). However, when $\pi = 0.1$, the power properties of all the tests, in particular $\widehat{AR}$ and $\widetilde{AR}_{df}$, deteriorate (Figure C.9 and Figure C.10 panel (c)). In the heteroskedastic case and when $k = 2, 5$, the T_1 and the T_2 tests along with the tests of Bun *et al.* (2018) are able to discriminate among alternatives (Figure C.13 to Figure C.16 panels (a) and (b)). To some extent the same could be said about the case where $k = 100$ (Figure C.13 to Figure C.16 panel (c)). When $k = 700$, $\pi = 0.1$ no test statistic among those considered seems to work well in this case. Only the AR_{AG} test has some power in the homoskedastic case (Figure C.9 and C.10 panel (d)). However, when $\pi = 1$, the T_1 and T_2 tests tend to outperform their competitors (Figure C.9 to Figure C.16 panel (d)).

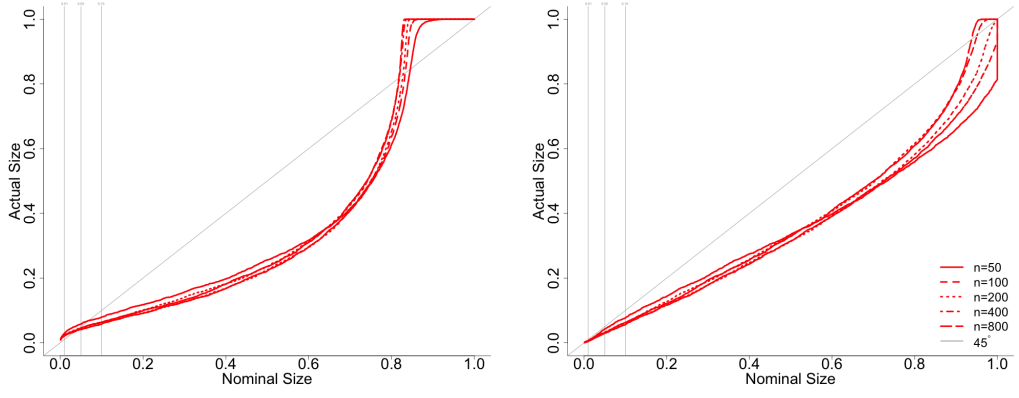
[4]It is worth noticing that, in general, for the hypothesis $H_0 : \beta_1 = \beta_{10}$ all the tests tend to underreject for small values of k.

C.2 Figures

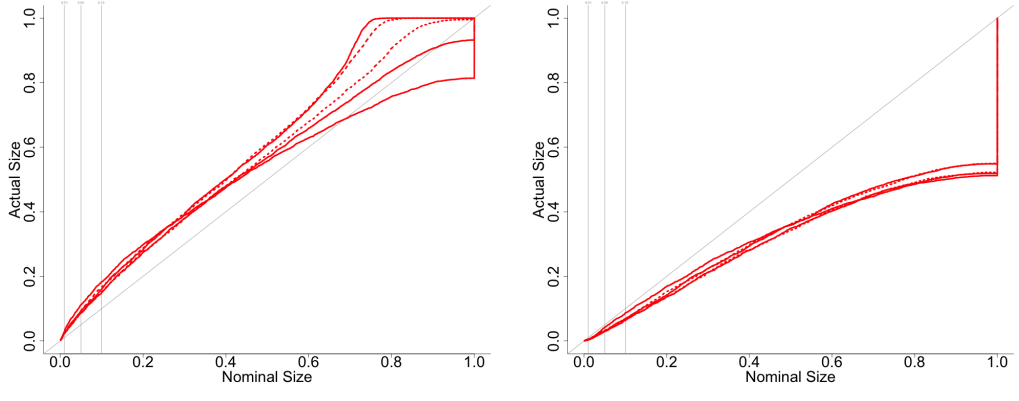


(a) T_1, $\mathcal{N}(0,1)$ (b) $\sqrt{k}T_1 + k$, χ^2_k

Figure C.1: PP-plots for T_1 under DGP II with heteroskedasticity, $k = 2$ and $n = 50, 100, 200, 400, 800$.

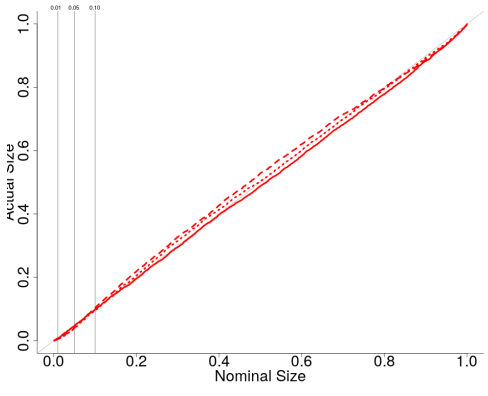


(a) T_2, $\mathcal{N}(0,1)$ (b) $\sqrt{k}T_2 + k$, χ^2_k

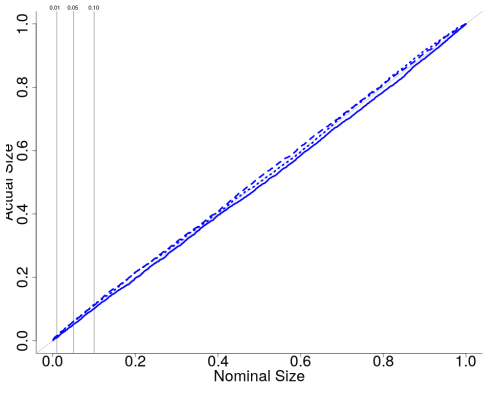


(c) $\sqrt{k}T_2 + k$, $\chi^2_{k-g_2}$ (d) $\sqrt{k-g_2}T_2 + (k-g_2)$, $\chi^2_{k-g_2}$

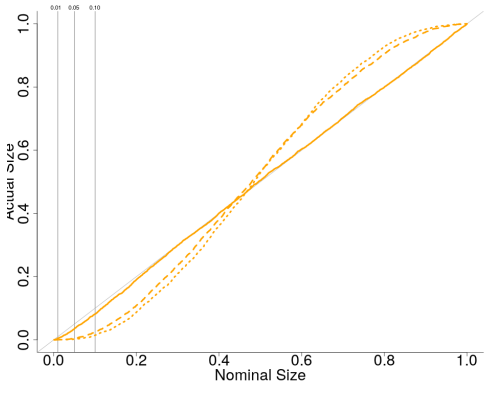
Figure C.2: PP-plots for T_2 under DGP II with heteroskedasticity, $k = 2$ and $n = 50, 100, 200, 400, 800$.



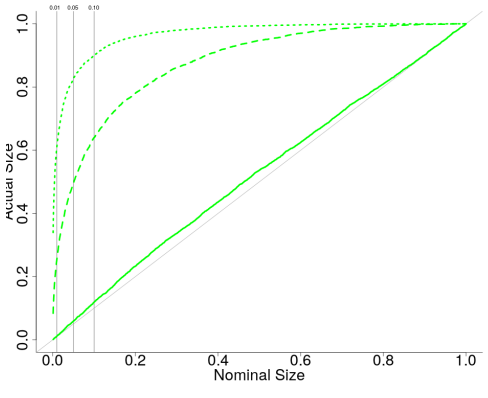
(a) T_1



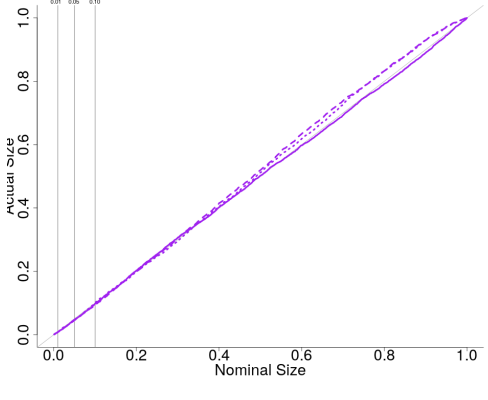
(b) AR_{AG}



(c) $\widehat{AR}$



(d) $\widetilde{AR}$



(e) $\widetilde{AR}_{df}$

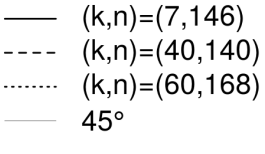
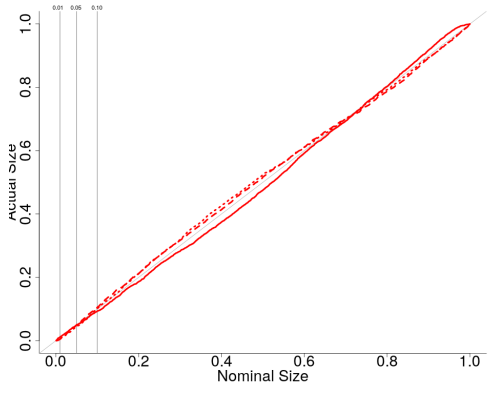
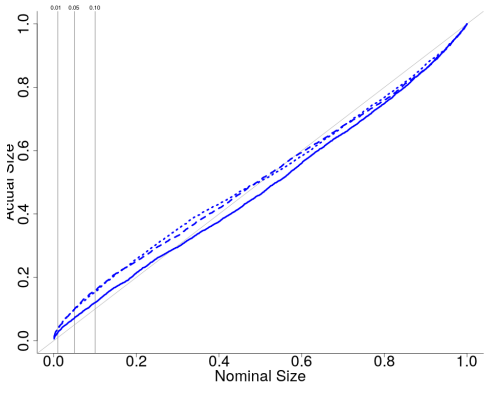


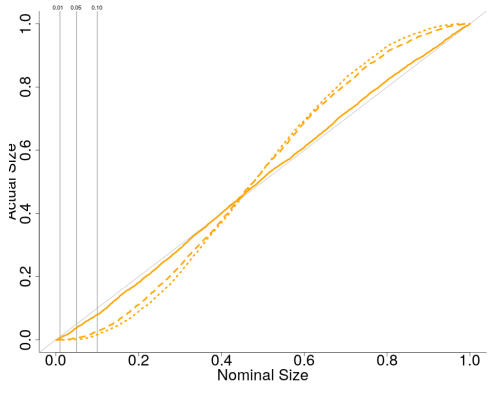
Figure C.3: PP-plots with homoskedasticity under DGP I, $H_0 : \beta = \beta_0$.



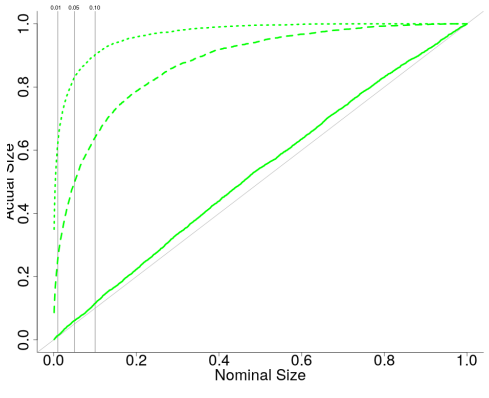
(a) T_1



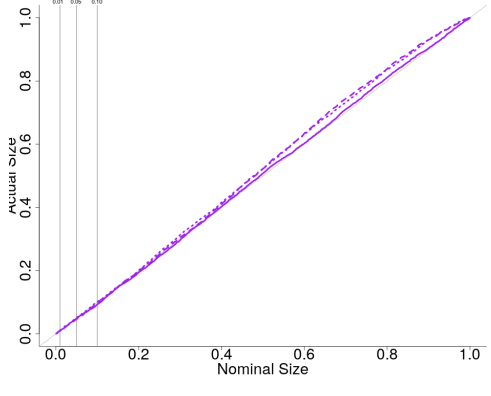
(b) AR_{AG}



(c) $\widehat{AR}$



(d) $\widetilde{AR}$



(e) $\widetilde{AR}_{df}$

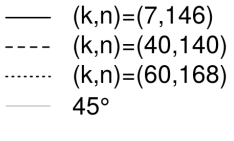
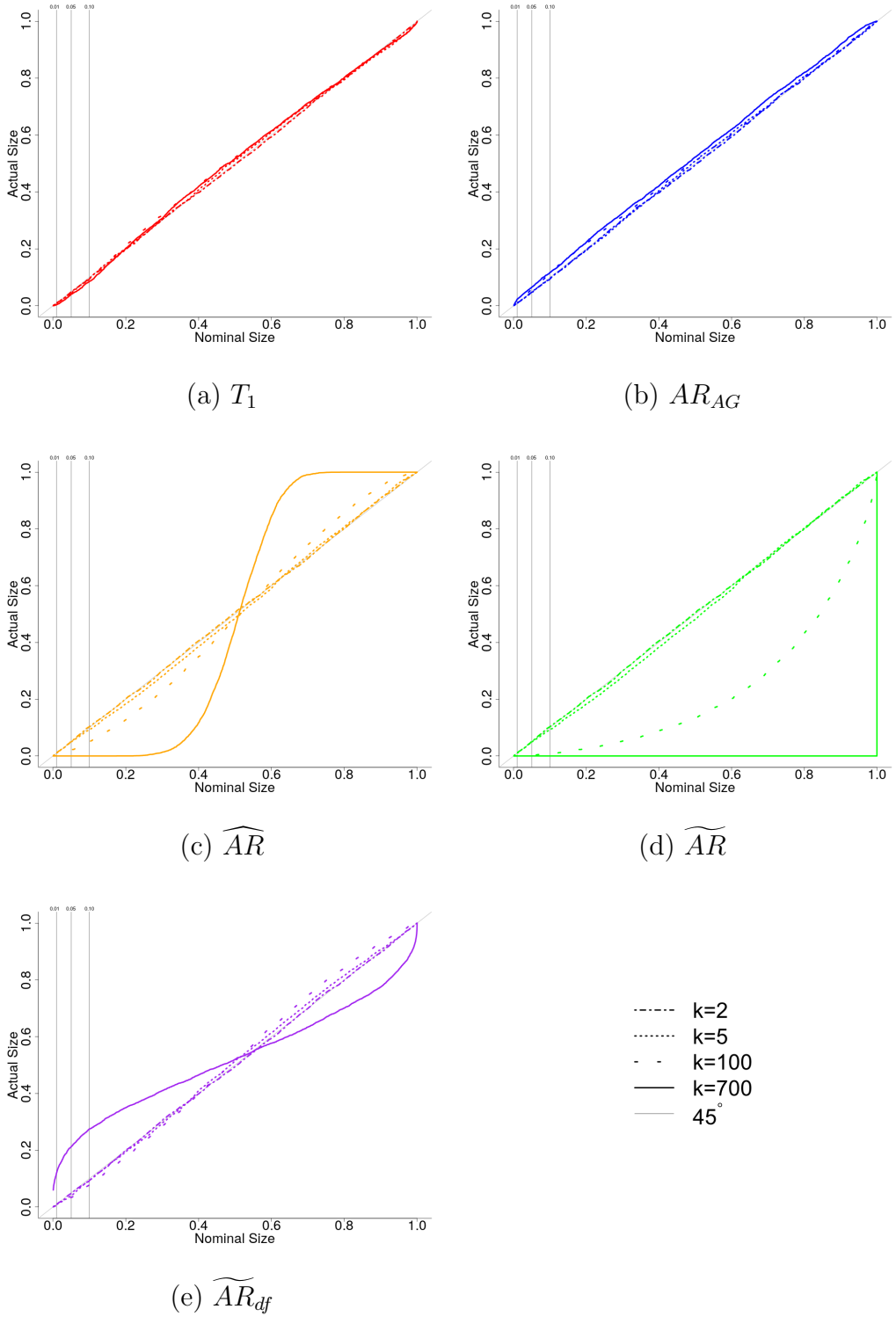


Figure C.4: PP-plots with heteroskedasticity under DGP I, $H_0 : \beta = \beta_0$.



(a) T_1

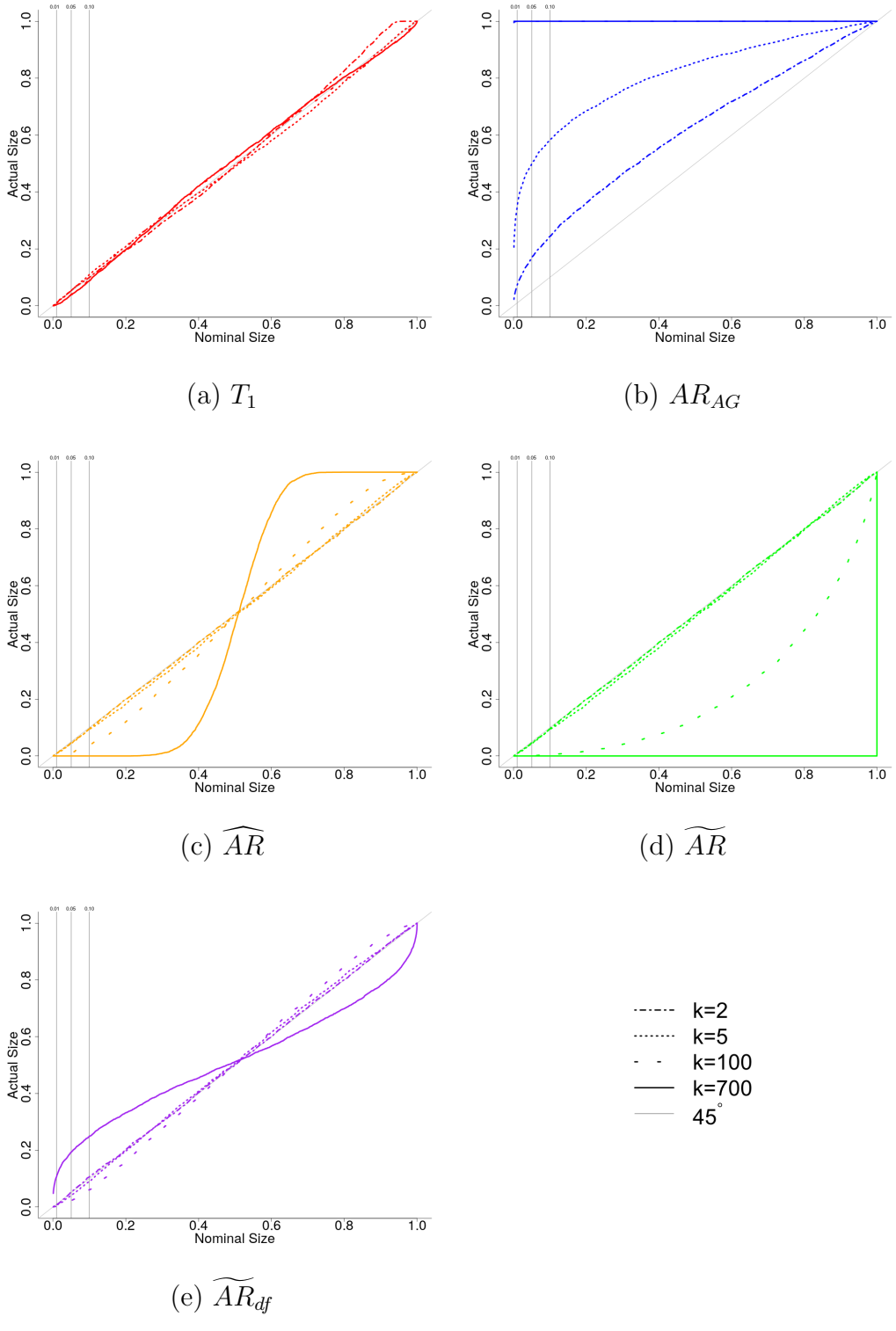
(b) AR_{AG}

(c) $\widehat{AR}$

(d) $\widetilde{AR}$

(e) $\widetilde{AR}_{df}$

Figure C.5: PP-plots with homoskedasticity under DGP II, $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$.



(a) T_1

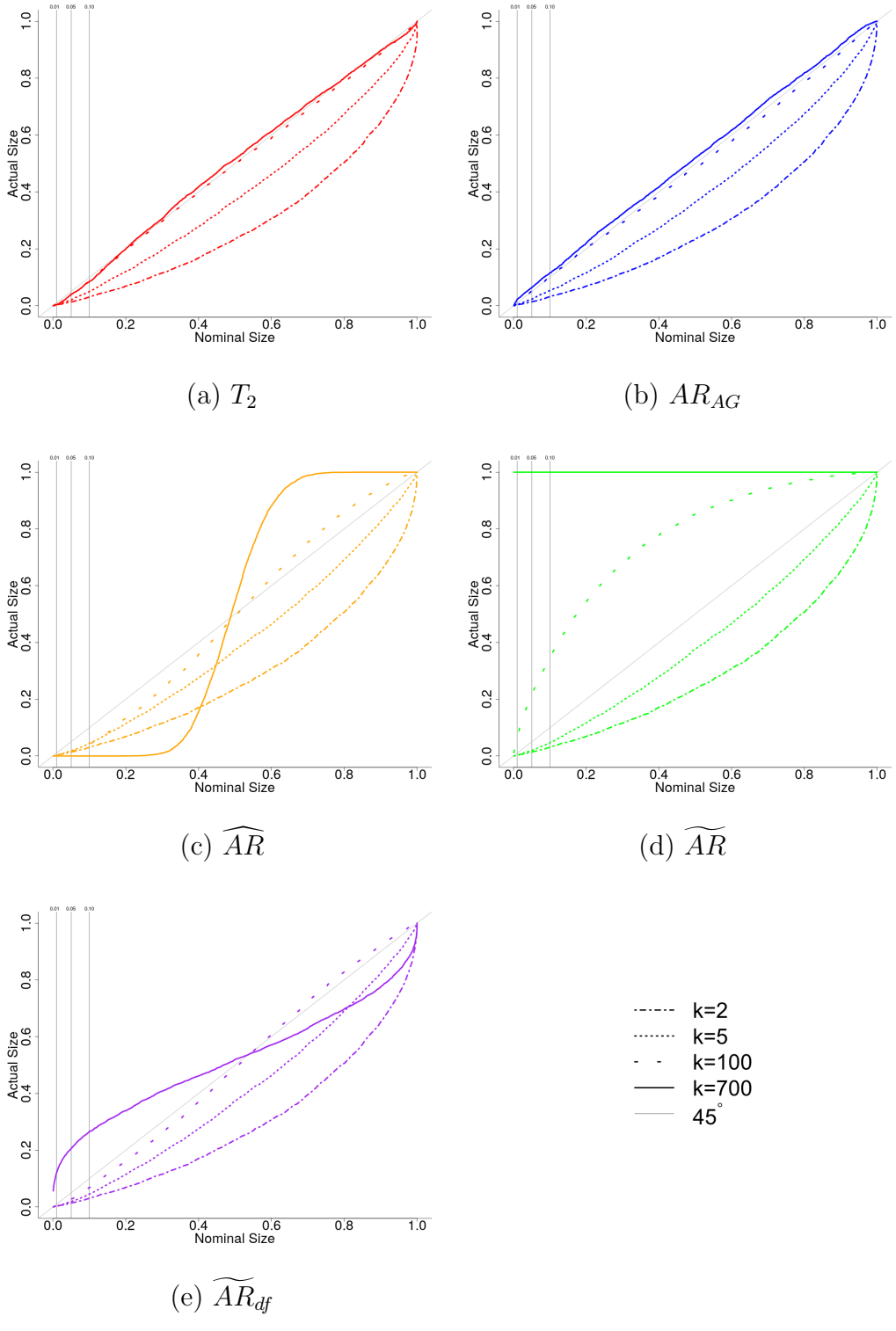
(b) AR_{AG}

(c) $\widehat{AR}$

(d) $\widetilde{AR}$

(e) $\widetilde{AR}_{df}$

Figure C.6: PP-plots with heteroskedasticity under DGP II, $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$.



(a) T_2

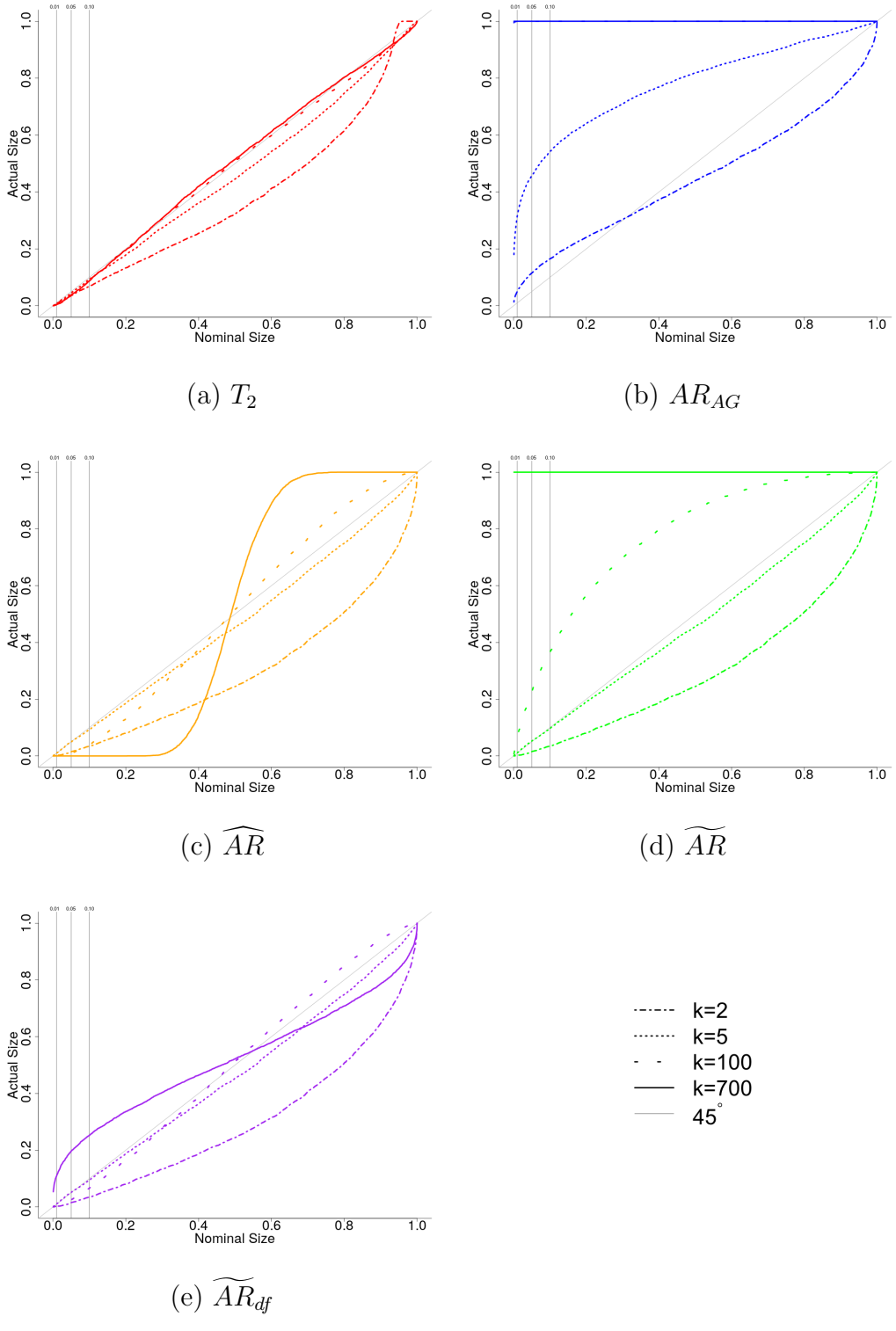
(b) AR_{AG}

(c) $\widehat{AR}$

(d) $\widetilde{AR}$

(e) $\widetilde{AR}_{df}$

Figure C.7: PP-plots with homoskedasticity under DGP II, $H_0 : \beta_1 = \beta_{10}$.



(a) T_2

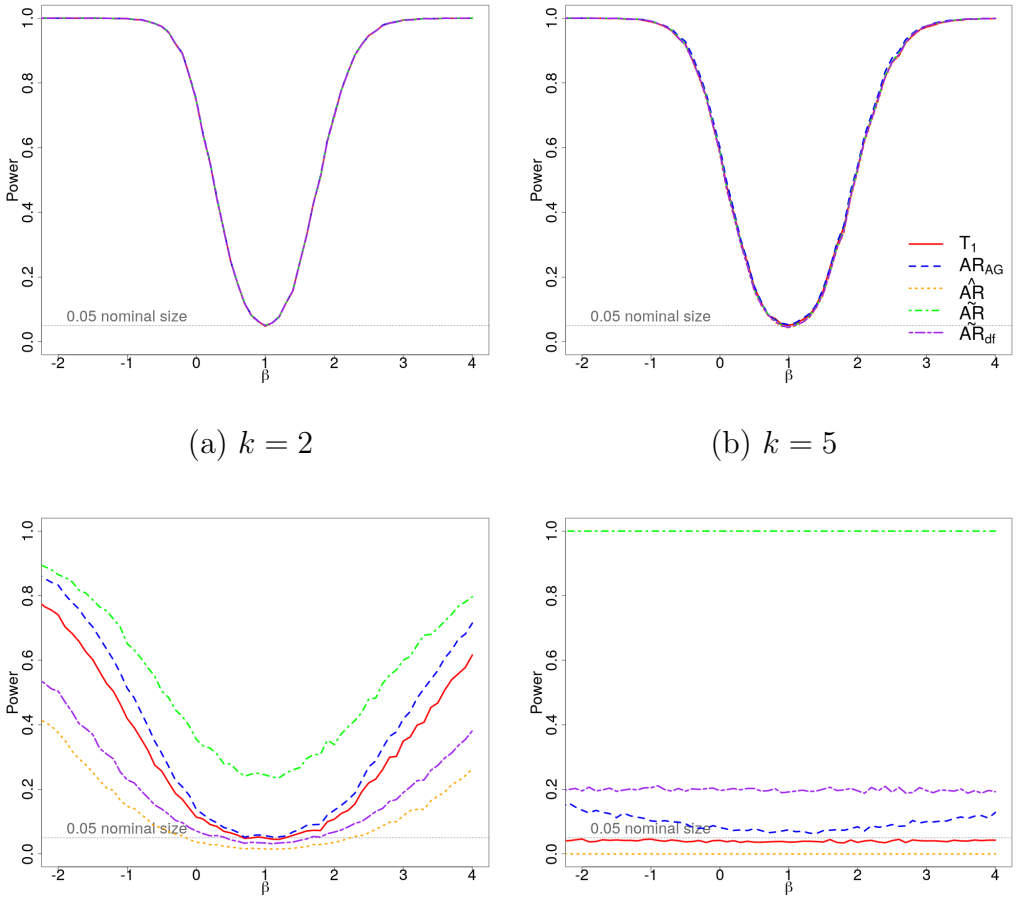
(b) AR_{AG}

(c) $\widehat{AR}$

(d) $\widetilde{AR}$

(e) $\widetilde{AR}_{df}$

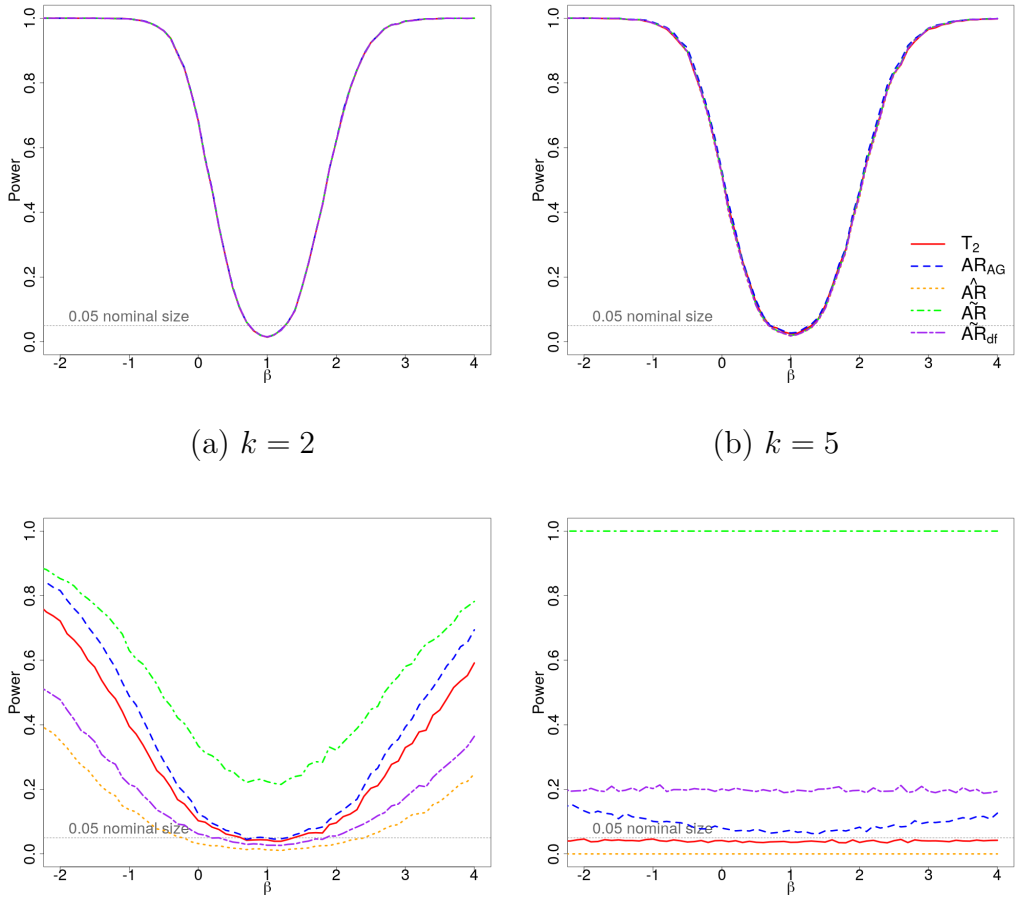
Figure C.8: PP-plots with heteroskedasticity under DGP II, $H_0 : \beta_1 = \beta_{10}$.



(a) $k = 2$ (b) $k = 5$

(c) $k = 100$ (d) $k = 700$

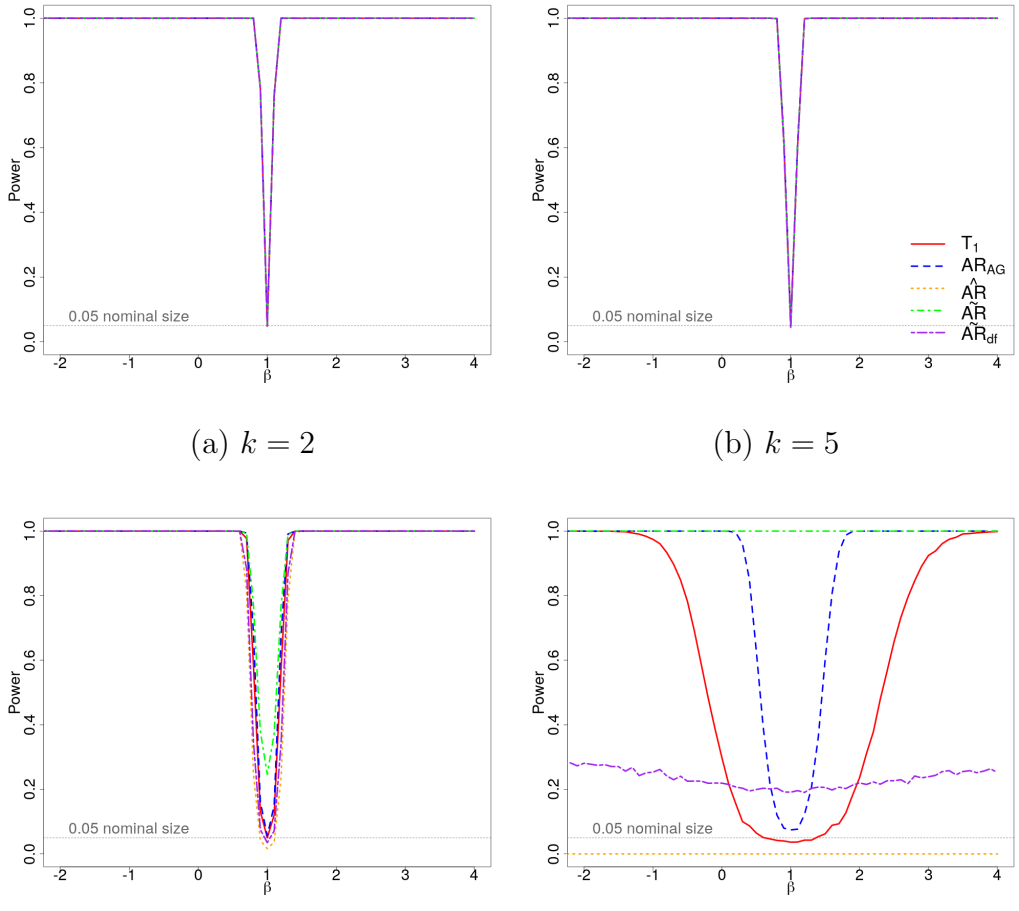
Figure C.9: Power curves with homoskedasticity and $\pi = 0.1$, $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$.



(a) $k = 2$ (b) $k = 5$

(c) $k = 100$ (d) $k = 700$

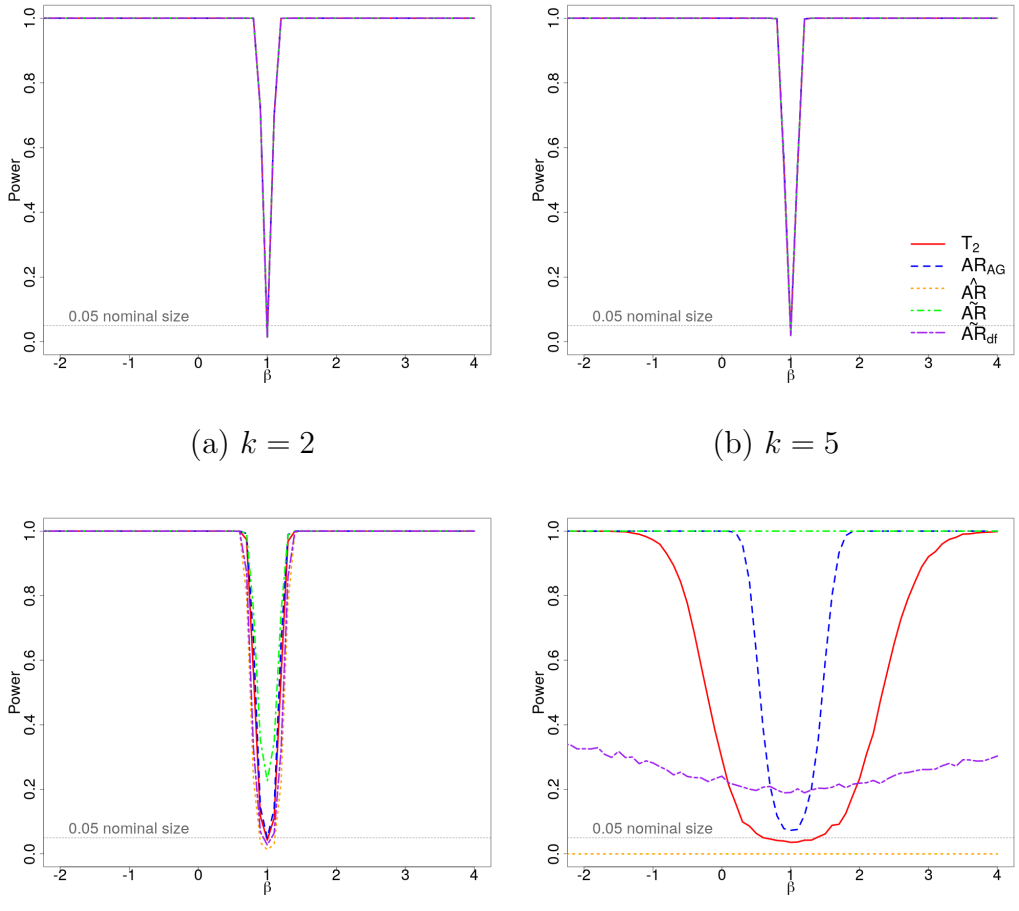
Figure C.10: Power curves with homoskedasticity and $\pi = 0.1$, $H_0 : \beta_1 = \beta_{10}$.



(a) $k = 2$ (b) $k = 5$

(c) $k = 100$ (d) $k = 700$

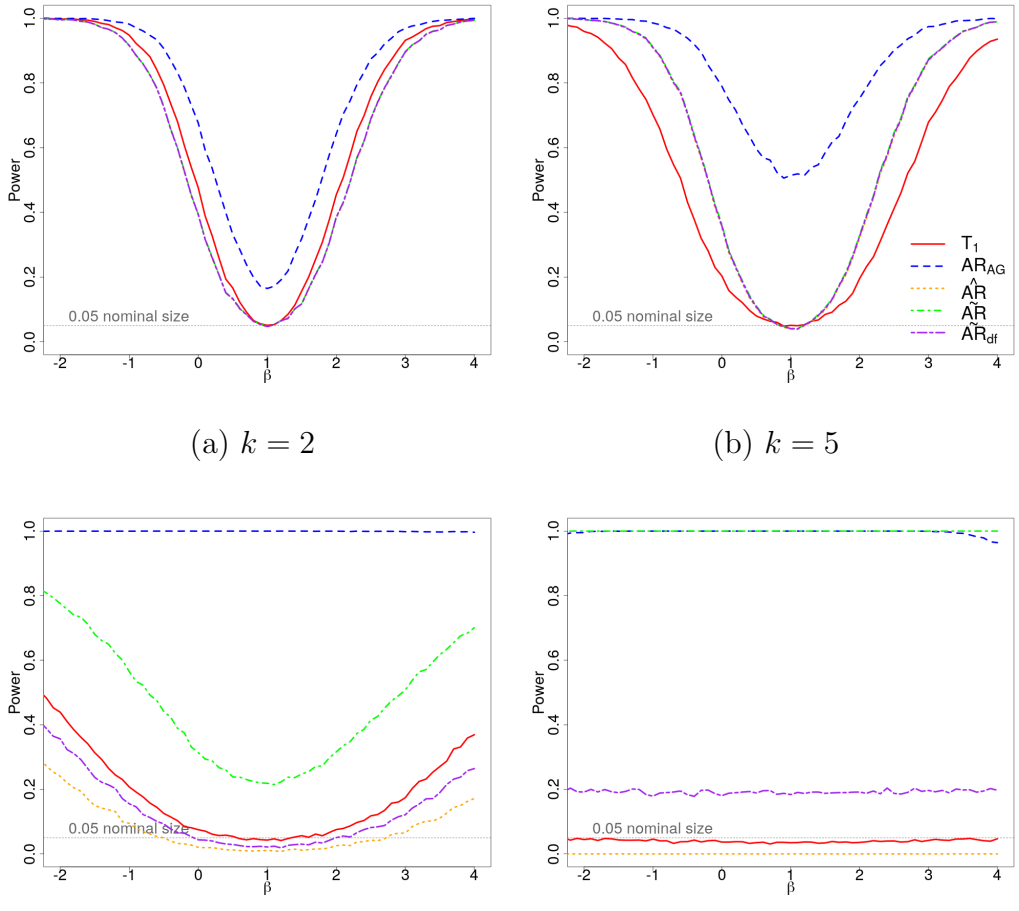
Figure C.11: Power curves with homoskedasticity and $\pi = 1$, $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$.



(a) $k = 2$ (b) $k = 700$

(c) $k = 100$ (d) $k = 700$

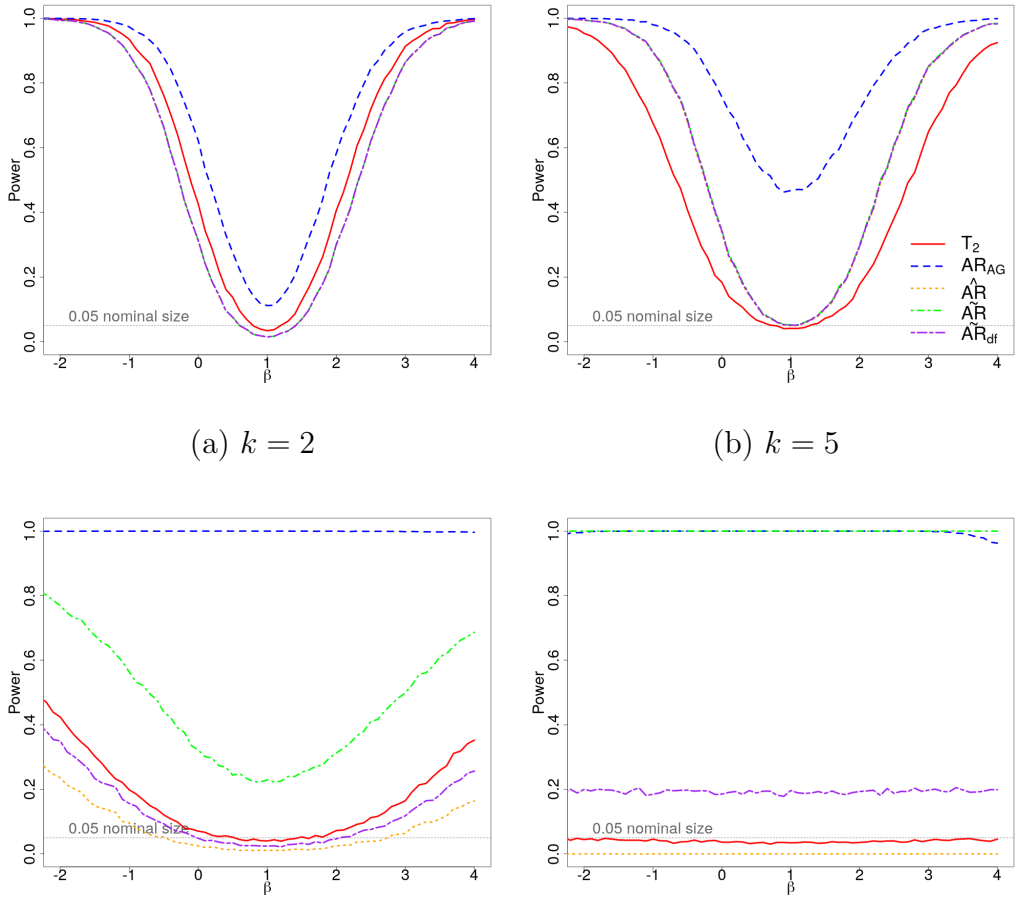
Figure C.12: Power curves with homoskedasticity and $\pi = 1$, $H_0 : \beta_1 = \beta_{10}$.



(a) $k = 2$ (b) $k = 5$

(c) $k = 100$ (d) $k = 700$

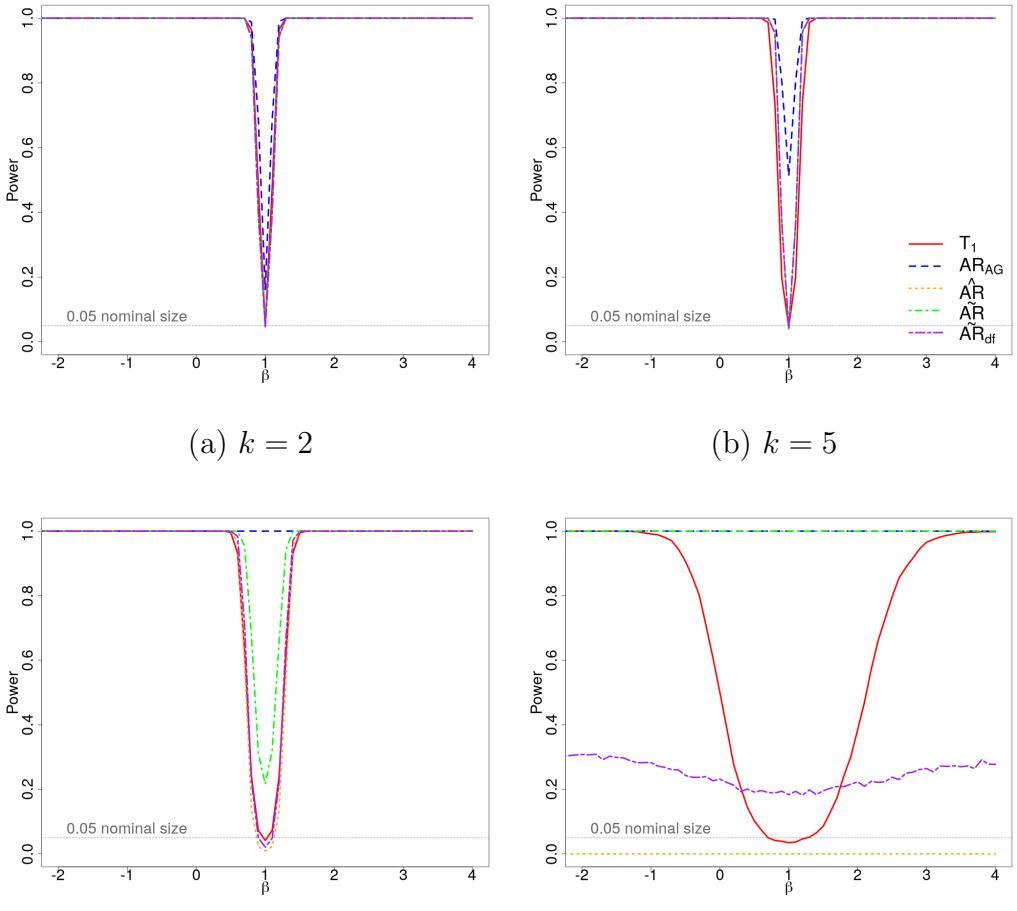
Figure C.13: Power curves with heteroskedasticity and $\pi = 0.1$, $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$.



(a) $k = 2$ (b) $k = 5$

(c) $k = 100$ (d) $k = 700$

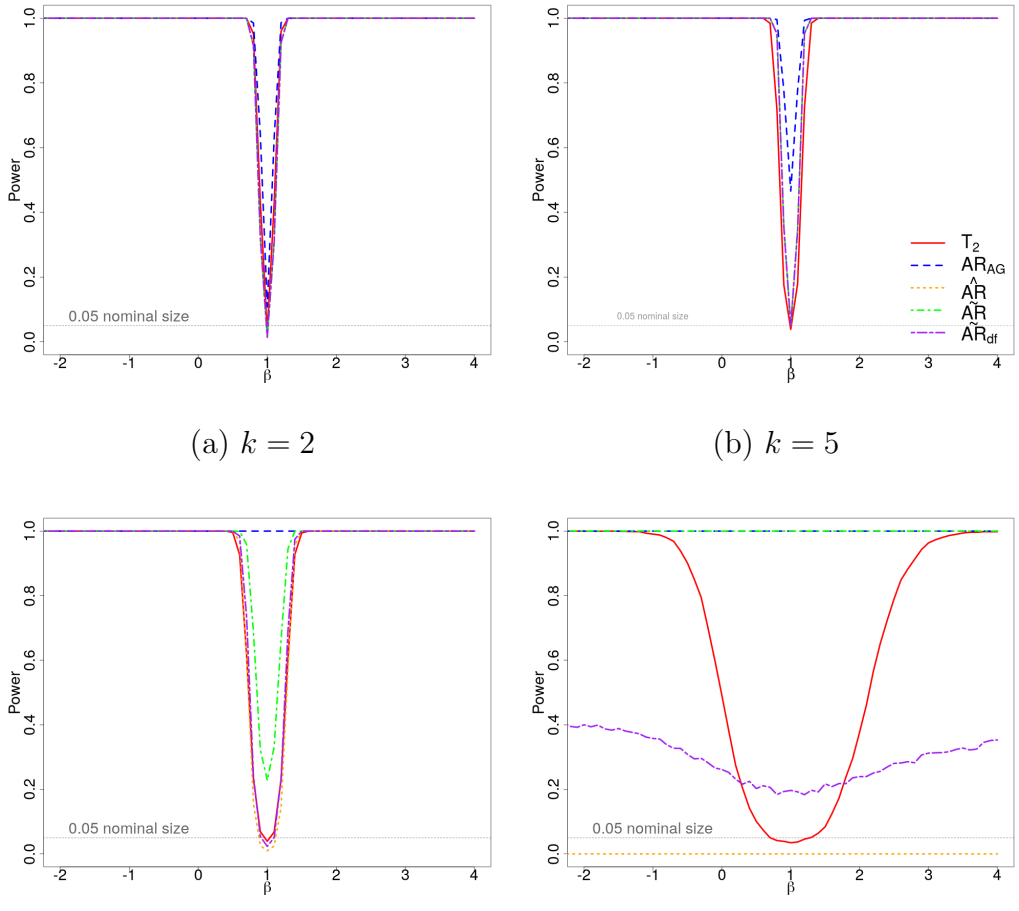
Figure C.14: Power curves with heteroskedasticity and $\pi = 0.1$, $H_0 : \beta_1 = \beta_{10}$.



(a) $k = 2$ (b) $k = 5$

(c) $k = 100$ (d) $k = 700$

Figure C.15: Power curves with heteroskedasticity and $\pi = 1$, $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$.



(a) $k = 2$ (b) $k = 5$

(c) $k = 100$ (d) $k = 700$

Figure C.16: Power curves with heteroskedasticity and $\pi = 1$, $H_0 : \beta_1 = \beta_{10}$.

References

Anatolyev, S. and Gospodinov, N. (2011) Specification Testing in Models with Many Instruments. *Econometric Theory* 27, 427–441.

Bekker, P.A. and Crudu, F. (2015) Jackknife Instrumental Variable Estimation with Heteroskedasticity. *The Journal of Econometrics* 185, 332–342.

Bekker, P.A. and Van der Ploeg, J. (2005) Instrumental variable estimation based on grouped data. *Statistica Neerlandica* 59, 239–267.

Bun, M., Farbmacher, H. and Poldermans, R. (2018) Finite sample properties of the Anderson and Rubin (1949) test. *working paper* .

Chao, J.C., Swanson, N.R., Hausman, J.A., Newey, W.K. and Woutersen, T. (2012) Asymptotic distribution of JIVE in a heteroskedastic IV regression with many instruments. *Econometric Theory* 28, 42–86.

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