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 University Miercurea Ciuc

August 2018

Abstract

This paper proposes specification tests for instrumental variable models that are robust to the presence of heteroskedasticity. The tests can be seen as generalizations of the Anderson-Rubin test. Our approach is based on the jackknife principle. We show that under the null the proposed statistics have Gaussian limiting distributions. Moreover, a simulation study shows that the proposed statistics have competitive finite sample properties in terms of size and power. Finally, we provide an empirical application using college proximity instruments to estimate the returns to education.

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

*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
‡Faculty of Economic and Human Sciences, Piata Libertătii 1, 530104 Miercurea Ciuc, Romania, zsosan@gmail.com

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 play a key role (see, e.g., Kleibergen, 2002, and references therein). The Anderson-Rubin test (Anderson & 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 & Gospodinov, 2011; D. W. K. Andrews & Stock, 2007; Kleibergen, 2002). This problem may exacerbate when heteroskedasticity is present.

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 & Crudu (2015). The analysis is closely related to the papers by Anatolyev & Gospodinov (2011) and Hausman et al. (2012).

First of all, we show that the many instrument results in Anatolyev & 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 this sense, our test statistics may be seen as generalizations of the AR test.

Over the years a number of improvements on the basic formulation of the AR test have been introduced. Kleibergen (2002) proposes a modification of the AR statistic that is robust to the presence of many instruments (see also Moreira, 2009). Moreira (2003), on the other hand, suggests replacing standard asymptotic critical values with conditional quantiles. The resulting conditional likelihood ratio (CLR) test enjoys excellent power properties (see D. W. K. Andrews et al., 2006). Other tests have been proposed by Staiger & Stock (1997), Wang & Zivot (1998) and Zivot et al. (1998). However, those tests tend to be conservative and are generally outperformed by the CLR test in terms of power (Stock et al., 2002). We are not aware of any study that generalize the AR test to the case of many instrumental variables and heteroskedasticity. Probably, the closest paper 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 plan of the paper is as follows: Section 2 introduces the model, Section 3 describes 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 and auxiliary results are relegated to the 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 $\mathrm{E}[\boldsymbol{X}] = \boldsymbol{Z}\boldsymbol{\Pi}$. Such assumptions are made for convenience and may be generalized, see e.g. Chao et al. (2014), Hausman et al. (2012), Bekker (1994). The rows of the disturbance couple $(\boldsymbol{\varepsilon}, \boldsymbol{U})$, 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

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

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

$$\widehat{\boldsymbol{\beta}}_{SJIVE} = \arg\min_{\boldsymbol{\beta}} Q_{SJIVE}(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} \frac{(\boldsymbol{y} - \boldsymbol{X\beta})'\boldsymbol{C}(\boldsymbol{y} - \boldsymbol{X\beta})}{(\boldsymbol{y} - \boldsymbol{X\beta})'\boldsymbol{B}(\boldsymbol{y} - \boldsymbol{X\beta})} \tag{5}$$

and, given the projection matrix $\boldsymbol{P} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$ and the diagonal matrix $\boldsymbol{D}$ containing the diagonal elements of $\boldsymbol{P}$,

$$\begin{aligned}
\boldsymbol{C} &= \boldsymbol{A} - \boldsymbol{B} \\
\boldsymbol{A} &= \boldsymbol{P} + \boldsymbol{\Delta} \\
\boldsymbol{B} &= (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{D}(\boldsymbol{I}_n - \boldsymbol{D})^{-1}(\boldsymbol{I}_n - \boldsymbol{P}) \\
\boldsymbol{\Delta} &= \boldsymbol{PD}(\boldsymbol{I}_n - \boldsymbol{D})^{-1}\boldsymbol{P} - \frac{1}{2}\boldsymbol{PD}(\boldsymbol{I}_n - \boldsymbol{D})^{-1} - \frac{1}{2}\boldsymbol{D}(\boldsymbol{I}_n - \boldsymbol{D})^{-1}\boldsymbol{P}.
\end{aligned}$$

To compute the SJIVE consider the partition $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$, where $\boldsymbol{X}_2 = \boldsymbol{Z}_2$ and $\boldsymbol{Z} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2)$. Define $\boldsymbol{C}^* = \boldsymbol{C} - \boldsymbol{AX}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{A}$, then

$$\widehat{\boldsymbol{\beta}}_{SJIVE} = \left(\boldsymbol{X}'\boldsymbol{CX} - \lambda\boldsymbol{X}'\boldsymbol{BX}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Cy} - \lambda\boldsymbol{X}'\boldsymbol{By}\right) \tag{6}$$

where

$$\lambda = \lambda_{\min}\Big\{\Big((\boldsymbol{y}, \boldsymbol{X}_1)'\boldsymbol{B}(\boldsymbol{y}, \boldsymbol{X}_1)\Big)^{-1}\Big((\boldsymbol{y}, \boldsymbol{X}_1)'\boldsymbol{C}^*(\boldsymbol{y}, \boldsymbol{X}_1)\Big)\Big\}.$$

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 & Gospodinov (2011) to the heteroskedastic case and we introduce our main results.

3.1 Assumptions

The assumptions we use are similar to those in Bekker & Crudu (2015). Additional assumptions are included to generalize some results due to Anatolyev & Gospodinov (2011).

Assumption 1. *The generic diagonal element of the projection matrix* $\boldsymbol{P}$, P_{ii}, *satisfies* $\max_i P_{ii} \leq 1 - 1/c_u$.

Assumption 2. *The covariance matrices of the disturbances are bounded,* $\boldsymbol{\Sigma}_i \leq c_u \boldsymbol{I}_{g+1}$, *and the variances satisfy* $\sigma_i^2 \geq \underline{\sigma}^2$ *for any* i.

Assumption 3. $\mathrm{E}[\varepsilon_i^4] \leq c_u$ *and* $\mathrm{E}\left[\|\boldsymbol{U}_i\|^4\right] \leq c_u$.

Let us define the signal matrix as $\boldsymbol{H} = \boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Pi}$. Let $r_{\min} = \lambda_{\min}(\boldsymbol{H})$ and $r_{\max} = \lambda_{\max}(\boldsymbol{H})$ be the smallest eigenvalue and the largest eigenvalue of the signal matrix respectively.

Assumption 4. $k/r_{\min} \to \kappa$ *and* κ *being a constant.*

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

Assumption 1 is a technical condition on the projection matrix $\boldsymbol{P}$. It requires that the main diagonal elements of $\boldsymbol{P}$ be bounded away from 1, not just to be less than or equal to 1. This assumption is rather standard in the literature (e.g., Bekker & Crudu, 2015; Hausman et al., 2012). Assumption 2 and Assumption 3 are regularity conditions; the former bounds the covariance matrix of the disturbances, while the latter bounds the fourth moments of errors. Assumptions 4 and 5 deal with either strong or weak instruments. In particular, the latter takes into consideration the possibility of different growth rates (see Bekker & Crudu, 2015; Chao & Swanson, 2005).

3.2 The AR test

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

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

and it is chi square distributed with k degrees of freedom. In the many instruments context and in the presence of homoskedasticity, the behaviour of the AR test has been studied by D. W. K. Andrews & Stock (2007) and Anatolyev & Gospodinov (2011), among others. The following result generalizes the results in Lemma 1 of Anatolyev & 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 and the limit* $\frac{1}{n}\sum_{i=1}^n (P_{ii} - \frac{k}{n})^2 \to 0$ *hold. In addition, assume that the limits* $\lim_{n\to\infty} \overline{\sigma}_n^2 = \sigma_0^2$ *and* $\lim_{n\to\infty} W_n = W_0$ *exist. Then, the statistic provided by Anatolyev & Gospodinov (2011) has the limit* [1]

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

Remark 1. *The asymptotic distribution result in Proposition 1 has two important implic-*

[1] We note that under homoskedasticity

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

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

ations. First, it means that this test statistic is not valid under heteroskedasticity. In fact the asymptotic size of this test is

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

Second, the test statistic T_1 proposed in Section 3.3 is also valid under homoskedasticity. Its asymptotic distribution requires the assumption that the main diagonal elements $P_{ii},\ i = 1, ..., n,$ of the projection matrix $\boldsymbol{P} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$ should be bounded away from 1. The test statistic proposed by Anatolyev & Gospodinov (2011) requires the stronger assumption that the main diagonal elements of $\boldsymbol{P}$ converge to $\lambda = \lim \frac{k}{n}$. Therefore, even under homoskedasticity T_1 has broader applicability than the test statistic proposed by Anatolyev and Gospodinov. 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 & Van der Ploeg, 2005). Anatolyev & Yaskov (2017, Section 5.1) show that in this case the main diagonal elements of $\boldsymbol{P}$ do not converge to $\lambda = \lim \frac{k}{n}$. In the 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.*

3.3 Inference with heteroskedasticity and many instruments

The test statistic we propose is based on the numerator of the objective function in equation (5). This is,

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

Consider testing the following null hypothesis

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \tag{8}$$

where $\boldsymbol{\beta}$ is the true parameter. The test statistic is defined as

$$T_1 = \frac{1}{\sqrt{k}} \frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0)'\boldsymbol{C}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0)}{\sqrt{\widehat{V}(\boldsymbol{\beta}_0)}}, \quad \widehat{V}(\boldsymbol{\beta}_0) = \frac{2}{k}\boldsymbol{\varepsilon}_0^{(2)'}\boldsymbol{C}^{(2)}\boldsymbol{\varepsilon}_0^{(2)} \tag{9}$$

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 us with the asymptotic distribution of the T_1 test statistic. [2]

Theorem 1. *If Assumptions 1, 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 $q_{1-\alpha}^{\mathcal{N}(0,1)}$ be the $(1-\alpha)$-th quantile of the Normal distribution. Then, the null hypothesis is rejected if $T_1 \geq q_{1-\alpha}^{\mathcal{N}(0,1)}$.

Sometimes we are 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

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

In this case, we need a consistent estimator for $\boldsymbol{\beta}_2$, say $\widetilde{\boldsymbol{\beta}}_2$. Under the null, a consistent estimator is, for example, the SJIVE. For the null hypothesis $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_{10}$ consider

[2] We can 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 this avenue since, as suggested in Bekker & Crudu (2015), $\boldsymbol{C}$ allows us to retain the whole signal matrix.

$\widetilde{\boldsymbol{\beta}} = \left(\boldsymbol{\beta}_{10}', \widetilde{\boldsymbol{\beta}}_2'\right)'$, $\widetilde{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}$ and let the modified test statistic, denoted as T_2, be

$$T_2 = \frac{1}{\sqrt{k}} \frac{\widetilde{\boldsymbol{\varepsilon}}' \boldsymbol{C} \widetilde{\boldsymbol{\varepsilon}}}{\sqrt{\widehat{V}(\widetilde{\boldsymbol{\beta}})}}, \quad \text{where } \widehat{V}(\widetilde{\boldsymbol{\beta}}) = \frac{2}{k} \widetilde{\boldsymbol{\varepsilon}}^{(2)'} \boldsymbol{C}^{(2)} \widetilde{\boldsymbol{\varepsilon}}^{(2)}. \tag{11}$$

The following theorem provides us with 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 $T_2 \geq q_{1-\alpha}^{\mathcal{N}(0,1)}$.

4 Monte Carlo simulations

We study the finite sample properties of T_1 and T_2 in terms of size and power. We make inference on the full parameter vector (see Figure 2 and Figure 4) and on the sole parameter associated to the endogenous variable (see Figure 3 and Figure 5). The proposed tests are compared to the AR test is the version proposed by Anatolyev & Gospodinov (2011), denoted as AR_{AG}. This comparison is useful for two reasons: first, we are able to see whether or not our tests work in the homoskedastic context; second, it gives us a clear idea of how much we gain by using a proper test in the heteroskedastic case.[3] We consider two data generating processes (DGPs), the first is a single parameter model that uses dummy instruments as in Bekker & Van der Ploeg (2005), while the second is a two parameter model used for instance in Hausman et al. (2012) and Bekker & Crudu (2015). The DGPs produced in this Section are characterized by the fact that the sum of the diagonal elements of $\boldsymbol{P}$ does not converge to $\lambda = \lim \frac{k}{n}$, as shown in Anatolyev & Yaskov (2017). The size properties of T_1 and T_2 are investigated by means of PP-plots as described in Davidson & MacKinnon (1998). We use the DGP in Section 4.2 to study the power properties of T_1 and T_2.

[3] The AR_{AG} test works quite well in the homoskedastic case but it is not designed to work in the presence of heteroskedasticity.

4.1 DGP I

In this section we consider a DGP similar to Bekker & 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

$$\boldsymbol{y} = \boldsymbol{x}\beta + \boldsymbol{\varepsilon}$$
$$\boldsymbol{x} = \boldsymbol{Z}\boldsymbol{\pi} + \boldsymbol{v}$$

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 $(n, k) \in \{(52, 7), (150, 40), (450, 120)\}$. 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.

4.2 DGP II

Let us consider the Monte Carlo set up of Hausman et al. (2012). The DGP is given by

$$\boldsymbol{y} = \boldsymbol{\iota}\gamma + \boldsymbol{x}\beta + \boldsymbol{\varepsilon}$$
$$\boldsymbol{x} = \boldsymbol{z}\pi + \boldsymbol{v}$$

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

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

where $\rho = 0.3$, $\psi = 0.86$ and conditional on $\boldsymbol{z}$, independent of $\boldsymbol{v}$, $\boldsymbol{w}_1 \sim \mathcal{N}(\mathbf{0},\ \mathrm{Diag}(\boldsymbol{z})^2)$ where $\mathrm{Diag}(\boldsymbol{z})$ is a diagonal matrix where the diagonal elements are the elements of $\boldsymbol{z}$ and $\boldsymbol{w}_2 \sim \mathcal{N}(\mathbf{0},\ \psi^2 \boldsymbol{I}_n)$. 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 $(1, z_i, z_i^2, z_i^3, z_i^4, z_i b_{1i}, \ldots, z_i b_{\ell i})$, $\ell = 95$, where, independent of other random variables, the elements $b_{1i}, \ldots, b_{\ell i}$ are i.i.d. Bernoulli distributed with $p = 1/2$. 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.[4]

4.3 Simulation results

The simulations produced via DGP I and displayed in Figure 1 show that the T_1 statistic hits the conventional nominal rejection levels for any combination of k and n, apart from the small sample size case where $(n, k) = (52, 7)$. On the other hand, in comparison with the T_1 statistic, the AR_{AG} clearly tends to overreject. This result hold whether we have homoskedasticity or heterosskedasticity. It is interesting to notice that this result is in line with the predictions of Example 1 in Section 3.2. With respect to DGP II, we observe that in the presence of homoskedasticity the T_1 test and the AR_{AG} test are nearly identical (left panel in Figure 2). This result repeats if we compare T_2 and AR_{AG} (left panel in Figure 3). On the other hand, we notice that in the heteroskedastic case (right panel in Figure 2 and Figure 3) the AR_{AG} statistic dramatically overrejects at any conventional level and instrument combination, while the T_1 and T_2 statistics are close to the corresponding

[4]We computed results also for $\ell = 5, 15, 35, 55, 75$ and we noticed that the various p-value curves would converge from the p-value curve associated to the smallest number of instruments ($k = 5$) to the p-value curve to the largest number of instruments ($k = 100$). This result replicates in all cases, including the power curves. For clarity of exposition we only display the curves associated to $k = 5$ and $k = 100$.

nominal levels for $k = 5$ and nearly equal for $k = 100$. As shown in Figure 4 and Figure 5, the power properties of the considered statistics are similar only in the homoskedastic case. In general, we notice that there is a trade off between size and power with respect to k. This is, as k grows the empirical size approaches the nominal size, but the power curves tend to get wider. This is not necessarily a justification for using a small k as it would imply that the test rejects too often.

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.[5] 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 [6]. 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. While 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

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

[6] Another reason not to control for experience when estimating returns of education, at least in 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 & Pischke (2008).

four variables, while in the second it includes fourteen variables. We use our T_2 statistic, the AR_{AG} and the standard AR 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 6. With only 4 excluded instruments all the tests give very similar results, in particular T_2, AR_{AG} and AR are not able to reject values in the intervals $[.47, 1.07]$, $[.48, 1.00]$, and $[.46, 1.10]$ at the 5% significance level, and at the $[.51, .88]$, $[.53, .83]$, and $[.52, .85]$ at the 10% significance level, respectively. The implied effect 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 & Rubin (1997).

The results with fourteen instruments are reported in figure 7. Probably due to the presence of heteroskedasticity, adding instruments deteriorates the performances of both the AR_{AG}, and the AR tests that rejects every single value of β at the 10% significance level and only values in the intervals $[.61, .79]$, and $[.53, .96]$ at the 5% significance level, respectively. On the other hand, increasing the number of instruments does not seem to have a big impact on our T_2 test that do not reject values of β in the interval [.47,1.43] at the 5% level and [.57,.94] at the 10%. These results are in line with what we find in our simulation study where, in the presence of heteroskedasticity, the AR_{AG} test tends to reject too often in comparison with the T_1 or the T_2 test.

6 Conclusion

This paper introduces two specification tests 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 & Crudu (2015). We prove that, after appropriate rescaling, the limiting distribution of the test statistics is standard normal. Moreover, simulation evidence shows that, in finite samples, the proposed tests outperform their competitors, such as the AR test proposed in Anatolyev & Gospodinov (2011).

In our empirical application, the standard Anderson-Rubin test and its modification by Anatolyev & 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 test potentially outperforms 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.

A Appendix

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 $\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{\ell=1}^{n}$.

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

$$\sqrt{k}\left(\frac{AR}{k} - 1\right) = \frac{\frac{1}{\sqrt{k}}\left(\frac{n-k}{k}\boldsymbol{\varepsilon}'\boldsymbol{P}\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}'\left(\boldsymbol{I} - \boldsymbol{P}\right)\boldsymbol{\varepsilon}\right)}{\frac{1}{k}\boldsymbol{\varepsilon}'\left(\boldsymbol{I} - \boldsymbol{P}\right)\boldsymbol{\varepsilon}} = \frac{n}{k}\frac{\frac{1}{\sqrt{k}}\left(\boldsymbol{\varepsilon}'\boldsymbol{P}\boldsymbol{\varepsilon} - \frac{k}{n}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}\right)}{\frac{1}{k}\boldsymbol{\varepsilon}\left(\boldsymbol{I} - \boldsymbol{P}\right)\boldsymbol{\varepsilon}}. \tag{12}$$

Note that

$$\frac{1}{\sqrt{k}}\left(\boldsymbol{\varepsilon}'\boldsymbol{P}\boldsymbol{\varepsilon}-\frac{k}{n}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}\right)=\frac{1}{\sqrt{k}}\sum_{i\neq j}P_{ij}\varepsilon_i\varepsilon_j+\frac{1}{\sqrt{k}}\sum_{i=1}^{n}\left(P_{ii}-\frac{k}{n}\right)\varepsilon_i^2\equiv E_1+E_2. \tag{13}$$

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

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

(the latter inequality comes from (21)), which is bounded away from 0. Consequently, W_n is bounded between two positive numbers. We obtain that $E_1/\sqrt{W_n}\xrightarrow{d}\mathcal{N}\left(0,1\right)$.

Regarding E_2 we note that by Assumption 3

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

Using the assumption $\frac{1}{k}\sum_{i=1}^{n}\left(P_{ii}-\frac{k}{n}\right)^2\to 0$, we obtain that $\mathrm{Var}[E_2]=o\left(1\right)$. Consequently, $E_2 = o_p(1)$; therefore,

$$\frac{E_1 + E_2}{\sqrt{W_n}} \xrightarrow{d} \mathcal{N}(0,1). \tag{14}$$

Regarding the denominator involved in (12) 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 (13) 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

$$\frac{1}{n}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} - \overline{\sigma}_n^2 = O_p\left(\frac{1}{\sqrt{k}}\right). \tag{15}$$

Consequently,

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

Now, from equation (12) 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

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

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 (14) implies that

$$\left(1-\frac{k}{n}\right)\frac{\overline{\sigma}_n^2}{\sqrt{W_n}}\sqrt{k}\left(\frac{AR}{k}-1\right)\xrightarrow{d}\mathcal{N}\left(0,1\right). \tag{16}$$

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. □

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

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

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

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

The expression

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

from (13) 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 $\mathrm{E}\left[\varepsilon_i^2\right] = \sigma^2$ 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

$$\begin{aligned} E_2 &= \frac{\ell_3}{2\ell_2 + 3\ell_3} \frac{\ell_2}{\sqrt{\ell}} \frac{\sum_{g:n_g=2} \sum_{i\in G_g} \varepsilon_i^2}{2\ell_2} - \frac{\ell_2}{2\ell_2 + 3\ell_3} \frac{\ell_3}{\sqrt{\ell}} \frac{\sum_{g:n_g=3} \sum_{i\in G_g} \varepsilon_i^2}{3\ell_3} \\ &= \frac{\ell_2 \ell_3}{\ell n} \sqrt{\ell} \left(\frac{\sum_{g:n_g=2} \sum_{i\in G_g} \varepsilon_i^2}{2\ell_2} - \sigma^2 - \left[\frac{\sum_{g:n_g=3} \sum_{i\in G_g} \varepsilon_i^2}{3\ell_3} - \sigma^2 \right] \right). \end{aligned} \tag{17}$$

By the CLT for IID observations

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

where $v = \mathrm{Var}\left[\varepsilon_i^2\right]$. 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

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

Therefore, from (17) 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.

$\square$

In the proof of Theorem 1 we use the following CLT, which, as argued by Bekker & 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)* $\frac{1}{n}\mathrm{Var}[Q] \geq c_u > 0$*. 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*

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

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

$$|C_{ij}| = \frac{|P_{ij}|}{2} \left(\frac{1}{1 - P_{ii}} + \frac{1}{1 - P_{jj}} \right) \leq c_u \, |P_{ij}| \quad \text{for any } i, j, \tag{18}$$

so

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

Since $\boldsymbol{P}$ is idempotent $\boldsymbol{P}^4 = \boldsymbol{P}$, so based on the main diagonal elements we have

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

so

$$\sum_{i=1}^{n} \sum_{j=1}^{n} P_{ij}^4 \leq tr\left(\boldsymbol{P}\right) = k \quad \text{and} \quad \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{h=1}^{n} P_{hi}^2 P_{hj}^2 \leq k. \tag{19}$$

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, (18) and (19)

$$\begin{aligned}\mathrm{E}\left[A_2^2\right] &= \frac{4}{k^2}\sum_{i=1}^n\sum_{j=1}^n\sum_{h=1}^n C_{hi}^2 C_{ij}^2 \sigma_h^2 \sigma_j^2\, \mathrm{E}\left[\eta_i^2\right] \leq \frac{4c_u^2}{k^2}\sum_{i=1}^n\sum_{j=1}^n\sum_{h=1}^n C_{hi}^2 C_{ij}^2 \\ &\leq \frac{c_u}{k^2}\sum_{i=1}^n\sum_{j=1}^n\sum_{h=1}^n P_{hi}^2 P_{ij}^2 \leq \frac{c_u}{k}.\end{aligned}$$

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, $\widehat{V}\left(\boldsymbol{\beta}_0\right) - V_n \xrightarrow{p} 0$. □

Proof of Theorem 1. Under the null hypothesis we have

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

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

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

where

$$\begin{aligned}\sum_{i=1}^n\sum_{j=1}^n C_{ij}^2 = &\quad \sum_{i\neq j}\frac{P_{ij}^2}{4}\left(\frac{1}{1-P_{ii}} + \frac{1}{1-P_{jj}}\right)^2 \geq \sum_{i\neq j}\frac{P_{ij}^2}{4}\left(1+1\right)^2 = \sum_{i\neq j}P_{ij}^2 \\ = &\quad \sum_{i=1}^n\sum_{j=1}^n P_{ij}^2 - \sum_{i=1}^n P_{ii}^2 = tr\left(\boldsymbol{P}\right) - \sum_{i=1}^n P_{ii}^2 = k - \sum_{i=1}^n P_{ii}^2. \qquad (20)\end{aligned}$$

By Assumption 1

$$\sum_{i=1}^{n} P_{ii}^2 \leq \max P_{ii} \sum_{i=1}^{n} P_{ii} \leq (1 - 1/c_u)\, tr\,(\boldsymbol{P}) = (1 - 1/c_u)\, k. \tag{21}$$

So

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

therefore,

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

which is bounded away from 0 if $\lim \frac{k}{n} > 0$. In this case we can apply the CLT and complete the proof. □

For the proof of Theorem 2 we need the following result.

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

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

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

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

and

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

where

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

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

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

therefore,

$$\left|\widehat{V}\left(\widetilde{\boldsymbol{\beta}}\right) - \widetilde{V}_n\right| \le \frac{2\left\|\left(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\right)\right\|^2}{k} \sum_{i=1}^n \sum_{j=1}^n C_{ij}^2 d_i d_j + \frac{2\left\|\left(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\right)\right\|}{k} \sum_{i=1}^n \sum_{j=1}^n C_{ij}^2 d_i \varepsilon_j^2. \quad (22)$$

Note that by the Cauchy-Schwarz inequality

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

where

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

Assumption 3 and Minkowski's inequality imply $\mathrm{E}\left[\|\boldsymbol{X}_i\|^4\right] \leq c_u$. 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 (18)

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

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 $\widehat{V}\left(\widetilde{\boldsymbol{\beta}}\right) - V_n \xrightarrow{p} 0$. □

Proof of Theorem 2. Let $\Delta = \widetilde{\boldsymbol{\varepsilon}}'\boldsymbol{C}\widetilde{\boldsymbol{\varepsilon}} - \boldsymbol{\varepsilon}'\boldsymbol{C}\boldsymbol{\varepsilon}$ and

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

Then

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

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 it follows that $\sqrt{\widehat{V}\left(\boldsymbol{\beta}\right)} - \sqrt{\widehat{V}\left(\widetilde{\boldsymbol{\beta}}\right)} = o_p(1)$, while since V_n is bounded away from 0 by Assumption 3, it follows that $\frac{1}{\sqrt{\widehat{V}(\widetilde{\boldsymbol{\beta}})}} = O_p(1)$. Theorem 1 implies that $\frac{1}{\sqrt{k}} \frac{\boldsymbol{\varepsilon}'\boldsymbol{C}\boldsymbol{\varepsilon}}{\sqrt{\widehat{V}(\boldsymbol{\beta})}} = O_p(1)$, so $B_1 = o_p(1)$.

Regarding B_2, note that

$$\Delta = \left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \boldsymbol{X}'\boldsymbol{C}\boldsymbol{X} \left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) - 2\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \boldsymbol{X}'\boldsymbol{C}\boldsymbol{\varepsilon}. \tag{23}$$

The first term from Δ is

$$\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \boldsymbol{X}'\boldsymbol{C}\boldsymbol{X} \left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \boldsymbol{H}^{1/2}\boldsymbol{H}^{-1/2}\boldsymbol{X}'\boldsymbol{C}\boldsymbol{X}\boldsymbol{H}^{-1/2}\boldsymbol{H}^{1/2}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right). \tag{24}$$

First we show that $\boldsymbol{H}^{-1/2}\boldsymbol{X}'\boldsymbol{C}\boldsymbol{X}\boldsymbol{H}^{-1/2} \xrightarrow{p} \boldsymbol{I}_g$. The model $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Pi} + \boldsymbol{U}$ implies that

$$\boldsymbol{X}'\boldsymbol{C}\boldsymbol{X} = (\boldsymbol{Z}\boldsymbol{\Pi})'\boldsymbol{C}\boldsymbol{Z}\boldsymbol{\Pi} + (\boldsymbol{Z}\boldsymbol{\Pi})'\boldsymbol{C}\boldsymbol{U} + \boldsymbol{U}'\boldsymbol{C}\boldsymbol{Z}\boldsymbol{\Pi} + \boldsymbol{U}'\boldsymbol{C}\boldsymbol{U}. \tag{25}$$

Note that $\boldsymbol{Z}'\boldsymbol{C}\boldsymbol{Z} = \boldsymbol{Z}'\boldsymbol{Z}$,hence $(\boldsymbol{Z}\boldsymbol{\Pi})'\boldsymbol{C}\boldsymbol{Z}\boldsymbol{\Pi} = \boldsymbol{H}$, and since the main diagonal elements of $\boldsymbol{C}$ are 0 we have $\mathrm{E}[\boldsymbol{U}'\boldsymbol{C}\boldsymbol{U}] = \boldsymbol{O}$. Therefore,

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

Based on (25),

$$\begin{aligned}
&\mathrm{Var}\left[\boldsymbol{H}^{-1/2}\boldsymbol{X}'\boldsymbol{C}\boldsymbol{X}\boldsymbol{H}^{-1/2}\right] = \\
&= \mathrm{E}\left[\boldsymbol{H}^{-1/2}\left\{(\boldsymbol{Z}\boldsymbol{\Pi})'\boldsymbol{C}\boldsymbol{U} + \boldsymbol{U}'\boldsymbol{C}\boldsymbol{Z}\boldsymbol{\Pi} + \boldsymbol{U}'\boldsymbol{C}\boldsymbol{U}\right\}\boldsymbol{H}^{-1}\left\{(\boldsymbol{Z}\boldsymbol{\Pi})'\boldsymbol{C}\boldsymbol{U} + \boldsymbol{U}'\boldsymbol{C}\boldsymbol{Z}\boldsymbol{\Pi} + \boldsymbol{U}'\boldsymbol{C}\boldsymbol{U}\right\}'\boldsymbol{H}^{-1/2}\right] \\
&\leq \frac{1}{r_{\min}}\boldsymbol{H}^{-1/2}\,\mathrm{E}\left[\left\{(\boldsymbol{Z}\boldsymbol{\Pi})'\boldsymbol{C}\boldsymbol{U} + \boldsymbol{U}'\boldsymbol{C}\boldsymbol{Z}\boldsymbol{\Pi} + \boldsymbol{U}'\boldsymbol{C}\boldsymbol{U}\right\}\left\{(\boldsymbol{Z}\boldsymbol{\Pi})'\boldsymbol{C}\boldsymbol{U} + \boldsymbol{U}'\boldsymbol{C}\boldsymbol{Z}\boldsymbol{\Pi} + \boldsymbol{U}'\boldsymbol{C}\boldsymbol{U}\right\}'\right]\boldsymbol{H}^{-1/2},
\end{aligned}$$

where the inequality holds due to $\boldsymbol{H}^{-1} \leq \frac{1}{r_{\min}} I_g$. By the Cauchy-Schwarz inequality

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

By Assumption 2 $\mathrm{E}[\boldsymbol{U}\boldsymbol{U}'] \leq c_u \boldsymbol{I}_n$ and due to the definition of $\boldsymbol{C}$ it holds that

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

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

$$\boldsymbol{Z}'\boldsymbol{C}^2\boldsymbol{Z} \leq c_u \boldsymbol{Z}'\boldsymbol{Z}, \tag{27}$$

so the first expectation is

$$\mathrm{E}\left[(\boldsymbol{Z\Pi})'\boldsymbol{C}\boldsymbol{U}\boldsymbol{U}'\boldsymbol{C}\boldsymbol{Z\Pi}\right] \leq c_u (\boldsymbol{Z\Pi})'\boldsymbol{C}^2\boldsymbol{Z\Pi} \leq c_u \boldsymbol{H}. \tag{28}$$

The second expectation $\mathrm{E}\left[\boldsymbol{U}'\boldsymbol{C}\boldsymbol{Z\Pi}(\boldsymbol{Z\Pi})'\boldsymbol{C}\boldsymbol{U}\right]$ is just the transpose of the first expectation, so also

$$\mathrm{E}\left[\boldsymbol{U}'\boldsymbol{C}\boldsymbol{Z\Pi}(\boldsymbol{Z\Pi})'\boldsymbol{C}\boldsymbol{U}\right] \leq c_u \boldsymbol{H}. \tag{29}$$

The third expectation

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

By Assumption 3, the Cauchy-Schwarz inequality, (20) and (21) we obtain that

$$\mathrm{E}[\boldsymbol{U}'\boldsymbol{C}\boldsymbol{U}\boldsymbol{U}'\boldsymbol{C}\boldsymbol{U}] \leq c_u\left(\sum_{i\neq j} C_{ij}^2\right)\boldsymbol{I}_g \leq c_u k \boldsymbol{I}_g. \tag{30}$$

Collecting the results from (28), (29), (30) we obtain that

$$\mathrm{Var}\left[\boldsymbol{H}^{-1/2}\boldsymbol{X}'\boldsymbol{C}\boldsymbol{X}\boldsymbol{H}^{-1/2}\right] \leq \frac{1}{r_{\min}}\boldsymbol{H}^{-1/2}(c_u\boldsymbol{H} + c_u k\boldsymbol{I}_g)\boldsymbol{H}^{-1/2} = \frac{1}{r_{\min}}\left(c_u + c_u\frac{k}{r_{\min}}\right)\boldsymbol{I}_g.$$

Therefore, Assumption 4 (many strong instruments case) implies that Var $\left[\boldsymbol{H}^{-1/2}\boldsymbol{X}'\boldsymbol{C}\boldsymbol{X}\boldsymbol{H}^{-1/2}\right] = O\left(\frac{1}{k}\right)$ while Assumption 5 (many weak instruments case) implies that Var $\left[\boldsymbol{H}^{-1/2}\boldsymbol{X}'\boldsymbol{C}\boldsymbol{X}\boldsymbol{H}^{-1/2}\right] = o(1)$. In either case we obtain that Var $\left[\boldsymbol{H}^{-1/2}\boldsymbol{X}'\boldsymbol{C}\boldsymbol{X}\boldsymbol{H}^{-1/2}\right] \to 0$, which together with (26) implies that $\boldsymbol{H}^{-1/2}\boldsymbol{X}'\boldsymbol{C}\boldsymbol{X}\boldsymbol{H}^{-1/2} \xrightarrow{p} \boldsymbol{I}_g$.

Note that under Assumption 4 $\boldsymbol{H}^{1/2}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = O_p(1)$ while under Assumption 5 $\frac{1}{\sqrt{k}}\boldsymbol{H}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = O_p(1)$ (see Section 4 in Bekker & Crudu, 2015). Therefore, under either Assumption 4 or Assumption 5, from (24) we conclude that

$$\frac{1}{\sqrt{k}}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\boldsymbol{X}'\boldsymbol{C}\boldsymbol{X}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = o_p(1). \tag{31}$$

The second term from Δ involves

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

Note that since the main diagonal elements of $\boldsymbol{C}$ are 0, we have

$$\mathrm{E}\left[\boldsymbol{X}'\boldsymbol{C}\boldsymbol{\varepsilon}\right] = \boldsymbol{0}. \tag{32}$$

Indeed, the model $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Pi} + \boldsymbol{U}$ implies

$$\mathrm{E}\left[\boldsymbol{X}'\boldsymbol{C}\boldsymbol{\varepsilon}\right] = \mathrm{E}\left[\boldsymbol{U}'\boldsymbol{C}\boldsymbol{\varepsilon}\right] = \sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}\left[\boldsymbol{U}'\boldsymbol{e}_i\boldsymbol{e}_i'\boldsymbol{C}\boldsymbol{e}_j\boldsymbol{e}_j'\boldsymbol{\varepsilon}\right] = \sum_{i=1}^{n}\sum_{j=1}^{n}C_{ij}\,\mathrm{E}\left[\boldsymbol{U}_i\varepsilon_j\right] = \boldsymbol{0}.$$

Next we show that Var $\left[\boldsymbol{H}^{-1}\boldsymbol{X}'\boldsymbol{C}\boldsymbol{\varepsilon}\right] = o(1)$. Note that the model $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Pi} + \boldsymbol{U}$ implies

$$\begin{aligned}\mathrm{Var}\left[\boldsymbol{H}^{-1}\boldsymbol{X}'\boldsymbol{C}\boldsymbol{\varepsilon}\right] &= \boldsymbol{H}^{-1}(\boldsymbol{Z}\boldsymbol{\Pi})'\boldsymbol{C}\,\mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right]\boldsymbol{C}\boldsymbol{Z}\boldsymbol{\Pi}\boldsymbol{H}^{-1} \\ &\quad + \boldsymbol{H}^{-1}\left\{\mathrm{E}\left[\boldsymbol{U}'\boldsymbol{C}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right]\boldsymbol{C}\boldsymbol{Z}\boldsymbol{\Pi} + (\boldsymbol{Z}\boldsymbol{\Pi})'\boldsymbol{C}\,\mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{C}\boldsymbol{U}\right] + \mathrm{E}\left[\boldsymbol{U}'\boldsymbol{C}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{C}\boldsymbol{U}\right]\right\}\boldsymbol{H}^{-1}\end{aligned} \tag{33}$$

By Assumption 2 and (27) the first term of 33 is

$$\boldsymbol{H}^{-1}(\boldsymbol{Z}\boldsymbol{\Pi})'\boldsymbol{C}\,\mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right]\boldsymbol{C}\boldsymbol{Z}\boldsymbol{\Pi}\boldsymbol{H}^{-1} \le c_u\boldsymbol{H}^{-1}(\boldsymbol{Z}\boldsymbol{\Pi})'\boldsymbol{Z}\boldsymbol{\Pi}\boldsymbol{H}^{-1} \le c_u\frac{1}{r_{\min}}\boldsymbol{I}_g, \tag{34}$$

where the latter inequality is due to $\boldsymbol{H}^{-1} \leq \frac{1}{r_{\min}} \boldsymbol{I}_g$. The first and second terms from the expression in $\{\cdot\}$ in (33) are 0 because

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

Since the main diagonal elements of $\boldsymbol{C}$ are 0, we have that $\mathrm{E}\left[\boldsymbol{U}'\boldsymbol{C}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right]\boldsymbol{C} = \boldsymbol{O}$. Consequently, $\boldsymbol{C}\,\mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{C}\boldsymbol{U}\right] = \boldsymbol{O}$ as well. The third term from the expression in $\{\cdot\}$ in (33) is

$$\begin{aligned}\mathrm{E}\left[\boldsymbol{U}'\boldsymbol{C}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{C}\boldsymbol{U}\right] &= \sum_{i,j,k,\ell} \mathrm{E}\left[\boldsymbol{U}'\boldsymbol{e}_i\boldsymbol{e}_i'\boldsymbol{C}\boldsymbol{e}_j\boldsymbol{e}_j'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{e}_k\boldsymbol{e}_k'\boldsymbol{C}\boldsymbol{e}_\ell\boldsymbol{e}_\ell'\boldsymbol{U}\right] = \sum_{i,j,k,\ell} \mathrm{E}\left[\boldsymbol{U}_i C_{ij}\varepsilon_j\varepsilon_k C_{k\ell}\boldsymbol{U}_\ell'\right] \\ &= \sum_{i\neq j} C_{ij}^2\left(\mathrm{E}\left[\varepsilon_j^2\boldsymbol{U}_i\boldsymbol{U}_i'\right] + \mathrm{E}\left[\varepsilon_i\boldsymbol{U}_i\varepsilon_j\boldsymbol{U}_j'\right]\right) = \sum_{i\neq j} C_{ij}^2\left(\sigma_i^2\boldsymbol{\Sigma}_{22i} + \boldsymbol{\sigma}_{21i}\boldsymbol{\sigma}_{12i}\right).\end{aligned}$$

By the Cauchy-Schwarz inequality, Assumption 2, (20) and (21) we obtain that

$$\mathrm{E}\left[\boldsymbol{U}'\boldsymbol{C}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{C}\boldsymbol{U}\right] \leq c_u \sum_{i\neq j} C_{ij}^2 \boldsymbol{I}_g \leq c_u k \boldsymbol{I}_g.$$

Consequently,

$$\boldsymbol{H}^{-1}\,\mathrm{E}\left[\boldsymbol{U}'\boldsymbol{C}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{C}\boldsymbol{U}\right]\boldsymbol{H}^{-1} \leq c_u k \boldsymbol{H}^{-2} \leq c_u \frac{k}{r_{\min}^2}\boldsymbol{I}_g, \tag{35}$$

where the latter inequality is due to $\boldsymbol{H}^{-1} \leq \frac{1}{r_{\min}} \boldsymbol{I}_g$. Collecting the results from (34) and (35), we obtain that

$$\mathrm{Var}\left(\boldsymbol{H}^{-1}\boldsymbol{X}'\boldsymbol{C}\boldsymbol{\varepsilon}\right) \leq c_u\left(\frac{1}{r_{\min}} + \frac{k}{r_{\min}^2}\right)\boldsymbol{I}_g. \tag{36}$$

We also obtain that

$$\mathrm{Var}\left[\frac{1}{\sqrt{k}}\boldsymbol{H}^{-1/2}\boldsymbol{X}'\boldsymbol{C}\boldsymbol{\varepsilon}\right] \leq c_u\left(\frac{1}{k} + \frac{1}{r_{\min}}\right)\boldsymbol{I}_g. \tag{37}$$

Under Assumption 4 (many strong instruments case) we get $\mathrm{Var}\left[\frac{1}{\sqrt{k}}\boldsymbol{H}^{-1/2}\boldsymbol{X}'\boldsymbol{C}\boldsymbol{\varepsilon}\right] =$

$O\left(\frac{1}{k}\right)$, which together with (32) implies that $\frac{1}{\sqrt{k}}\boldsymbol{H}^{-1/2}\boldsymbol{X}'\boldsymbol{C}\boldsymbol{\varepsilon} = o_p(1)$. Since $\boldsymbol{H}^{1/2}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = O_p(1)$ holds, we obtain

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

Under Assumption 5 (many weak instruments case) (36) implies $\mathrm{Var}\left[\boldsymbol{H}^{-1}\boldsymbol{X}'\boldsymbol{C}\boldsymbol{\varepsilon}\right] = o(1)$, which together with (32) implies that $\boldsymbol{H}^{-1}\boldsymbol{X}'\boldsymbol{C}\boldsymbol{\varepsilon} = o_p(1)$. Since $\frac{1}{\sqrt{k}}\boldsymbol{H}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = O_p(1)$ holds, we obtain

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

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

References

Anatolyev, S. & Gospodinov, N. (2011). Specification testing in models with many instruments. *Econometric Theory*, *27*, 427–441.

Anatolyev, S. & Yaskov, P. (2017). Asymptotics of diagonal elements of projection matrices under many instruments/regressors. *Econometric Theory*, *33*(3), 717–738.

Anderson, T. W. & 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., Marmer, V. & Yu, Z. (in press). On optimal inference in the linear iv model. *Quantitative Economics*.

Andrews, D. W. K., Moreira, M. J. & Stock, J. H. (2006). Optimal two-sided invariant similar tests for instrumental variable regression. *Econometrica*, *73*, 715–752.

Andrews, D. W. K. & Stock, J. (2005). Inference with weak instruments. In R. Blundell, W. K. Newey & T. Persson (Eds.), *Advances in economics and econometrics, theory and applications: Ninth world congress of the Econometric Society* (Vol. III). Cambridge: Cambridge University Press.

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

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

Angrist, J. D. & 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. & Crudu, F. (2015). Jackknife instrumental variable estimation with heteroskedasticity. *The Journal of Econometrics*, *185*, 332–342.

Bekker, P. A. & Van der Ploeg, J. (2005). Instrumental variable estimation based on grouped data. *Statistica Neerlandica*, *59*(3), 239–267.

Bound, J., Jaeger, D. A. & 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. & 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 & R. Swidinsky (Eds.), *Aspects of labor market behaviour: Essays in honour of John Vanderkamp* (p. 201-222). Toronto: University of Toronto Press.

Chao, J. C., Hausman, J. A., Newey, W. K., Swanson, N. R. & Woutersen, T. (2014). Testing overidentifying restrictions with many instruments and heteroskedasticity. *The Journal of Econometrics*, *178*, 15–21.

Chao, J. C. & Swanson, N. R. (2005). Consistent estimation with a large number of weak instruments. *Econometrica*, *73*(5), 1673–1692.

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

Davidson, R. & MacKinnon, J. G. (1998). Graphical methods for investigating the size and power of hypothesis tests. *The Manchester School*, *66*, 1–26.

Guggenberger, P., Kleibergen, F., Mavroeidis, S. & Chen, L. (2012). On the asymptotic sizes of subset Anderson–Rubin and Lagrange multiplier tests in linear instrumental variables regression. *Econometrica*, *80*(6), 2649–2666.

Hausman, J. A., Newey, W. K., Woutersen, T., Chao, J. C. & 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. & Rubin, D. (1997). Estimating outcome distributions for compliers in instrumental variables models. *Review of Economic Studies*, *64*, 555-574.

Jiang, Y., Kang, H. & Small, D. (2016). ivmodel: Statistical inference and sensitivity analysis for instrumental variables model [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=ivmodel` (R package version 1.2)

Kang, H., Zhang, A., Cai, T. T. & Small, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American Statistical Association*, *111*(513), 132-144.

Keleijan, H. H. & Prucha, I. R. (2001). On the asymptotic distribution of the Moran I test statistic with applications. *The Journal of Econometrics*, *104*, 219–257.

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*(1), 418-423.

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.

Staiger, D. & Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, *65*, 557–586.

Stock, J. H., Wright, J. H. & 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. & Windmeijer, F. (2016). Genetic markers as instrumental variables. *Journal of Health Economics*, *45*(C), 131-148.

Wang, J. & 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. & 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. & Nelson, C. R. (1998). Valid confidence intervals and inference in the presence of weak instruments. *International Economic Review*, *39*, 1119–1144.

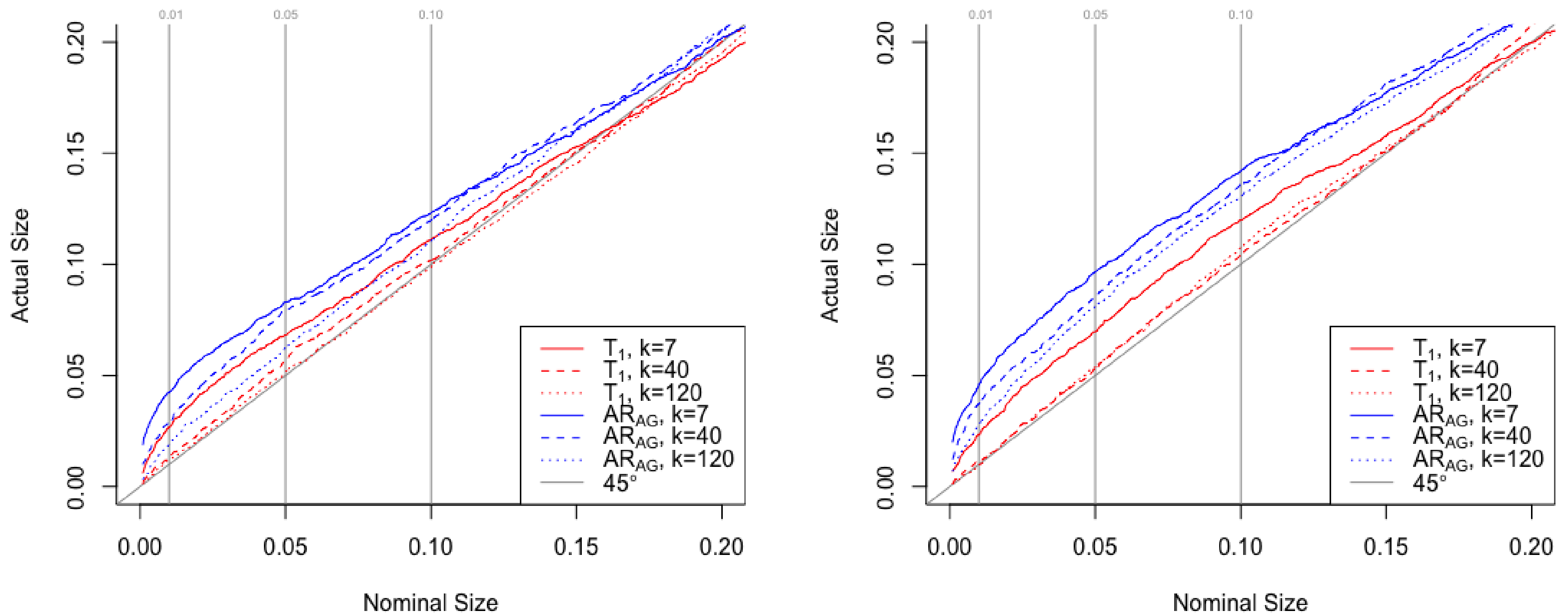

Figure 1: PP-plots of T_1 and AR_{AG} for the dummy instrument case with homoskedasticity (left) and heteroskedasticity (right) with $(n, k) \in \{(52, 7), (150, 40), (450, 120)\}$, $H_0 : \beta = \beta_0$.

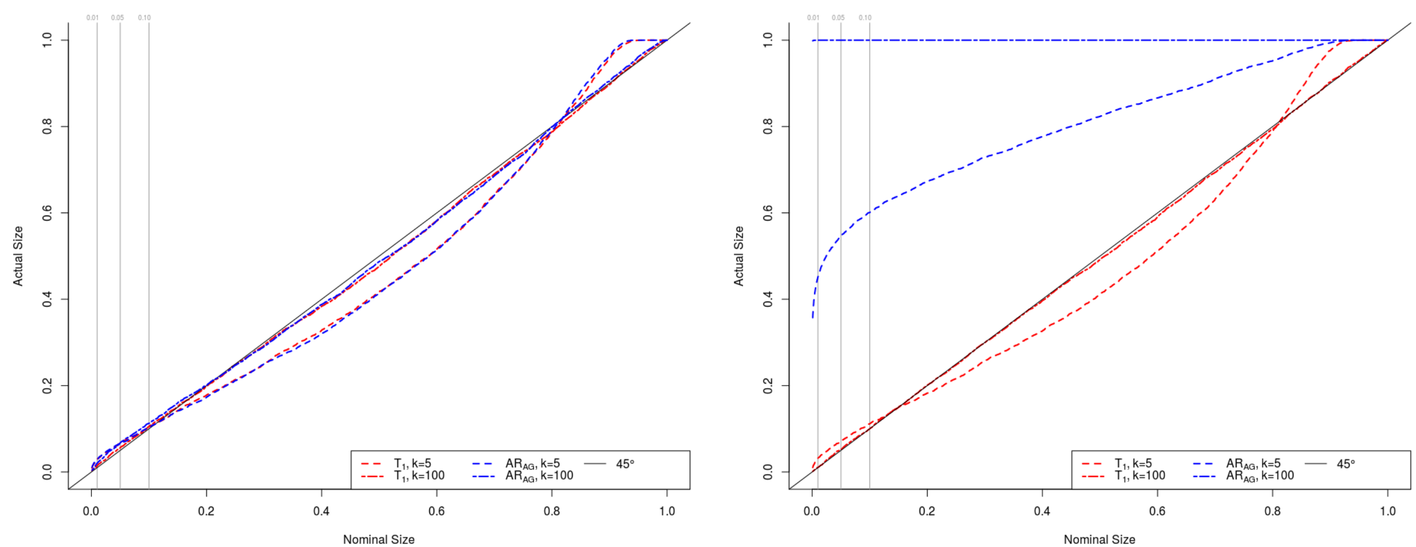

Figure 2: PP-plots of T_1 and AR_{AG} with homoskedasticity (left) and heteroskedasticity (right), $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$.

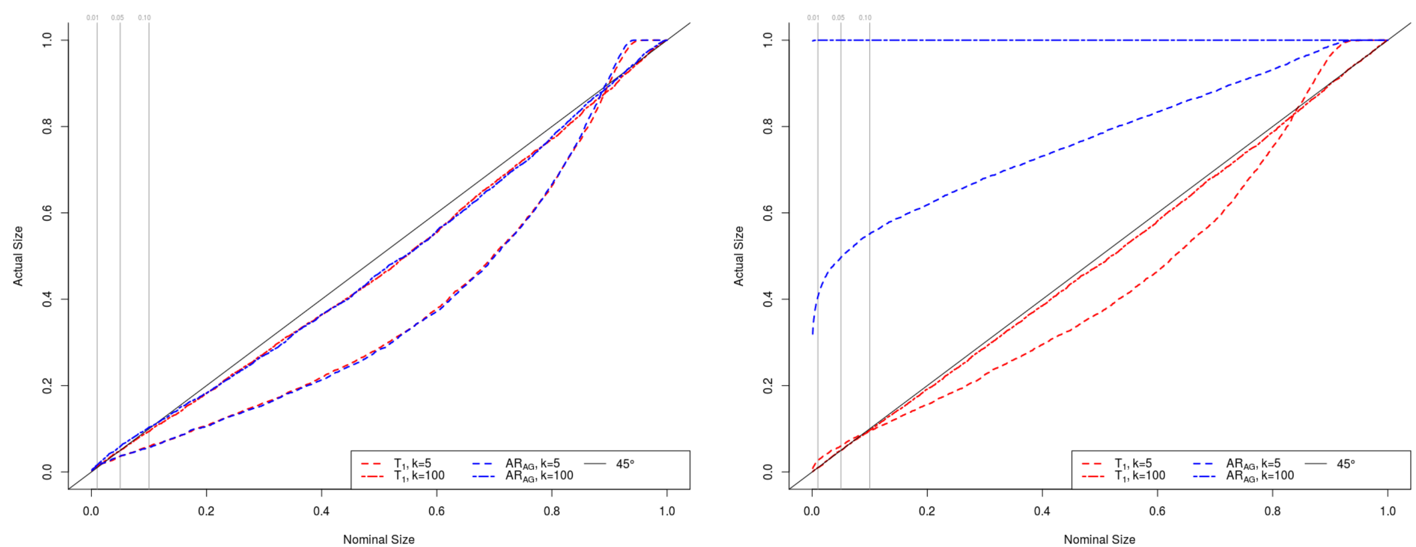

Figure 3: PP-plots of T_2 and AR_{AG} with homoskedasticity (left) and heteroskedasticity (right), $H_0 : \beta_1 = \beta_{10}$.

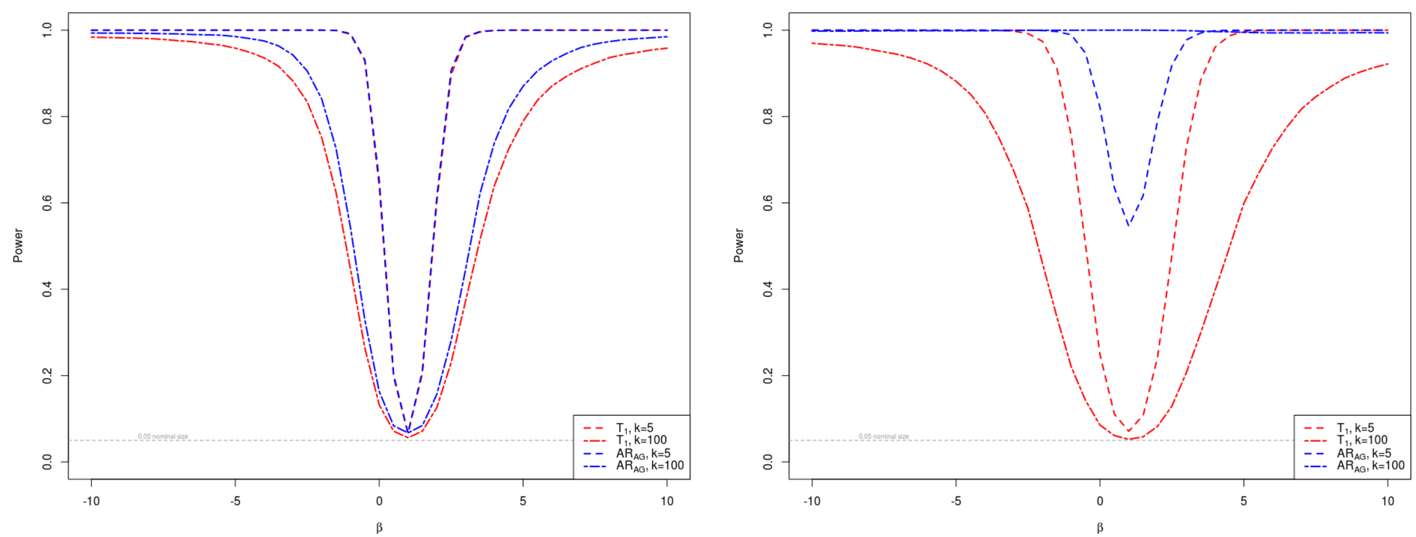

Figure 4: Power of T_1 and AR_{AG} with homoskedasticity (left) and heteroskedasticity (right), $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$.

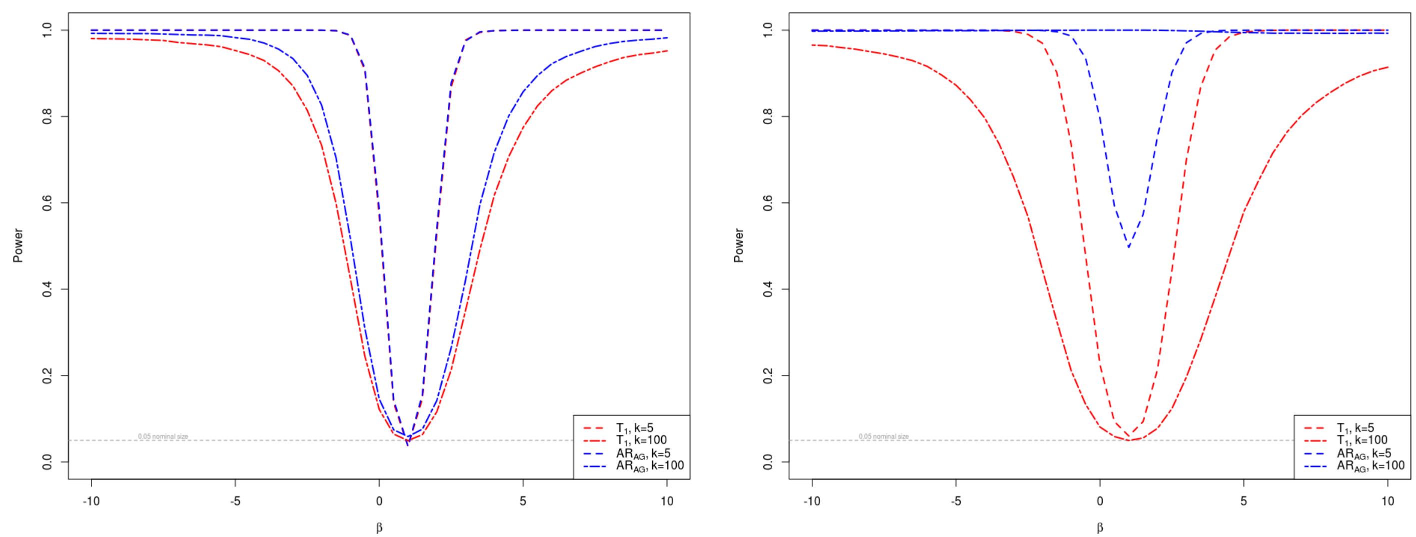

Figure 5: Power of T_2 and AR_{AG} with homoskedasticity (left) and heteroskedasticity (right), $H_0 : \beta_1 = \beta_{10}$.

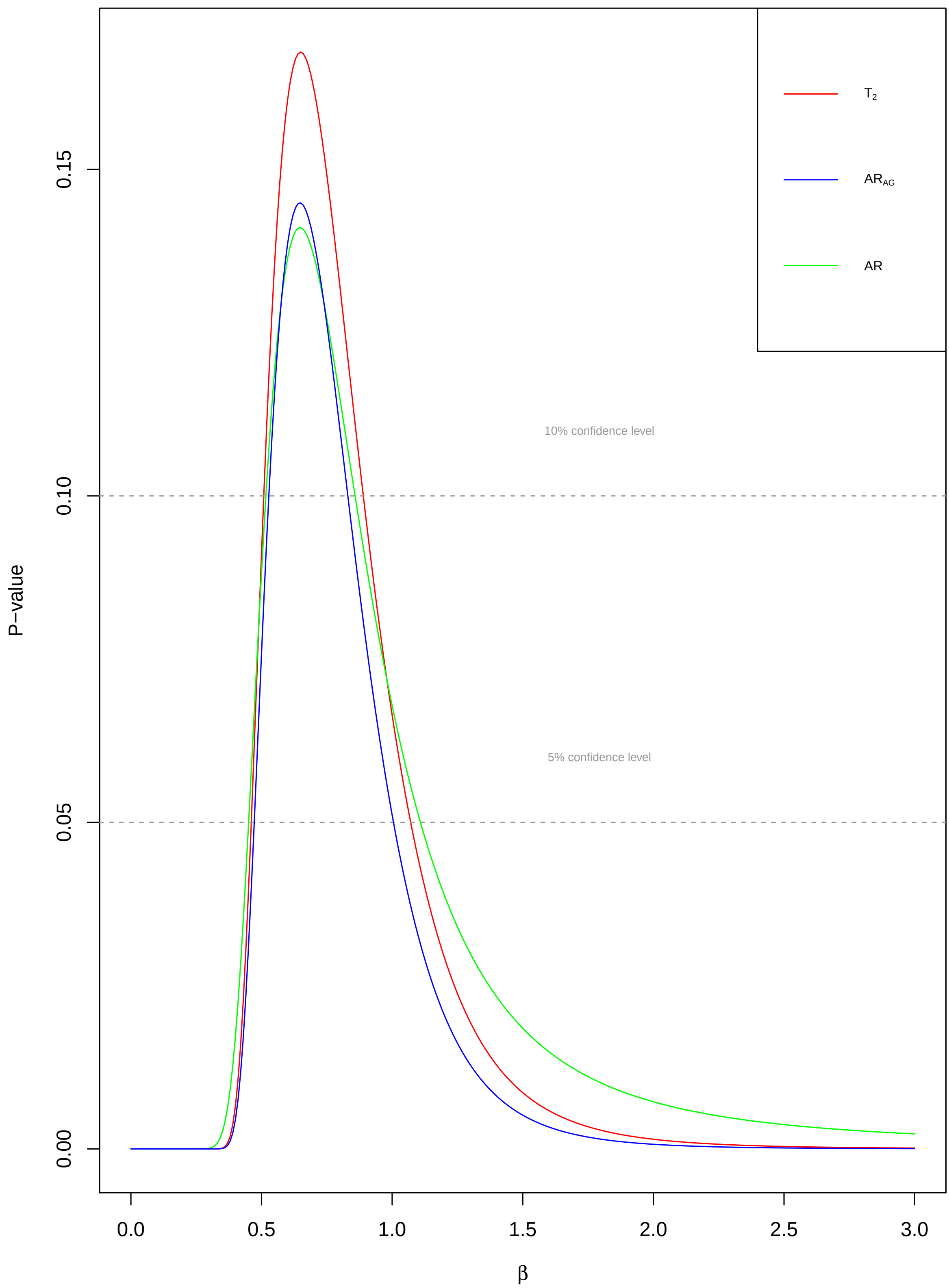

Figure 6: T_2, AR_{AG} and AR P-values for different values of β for the model with four instruments.

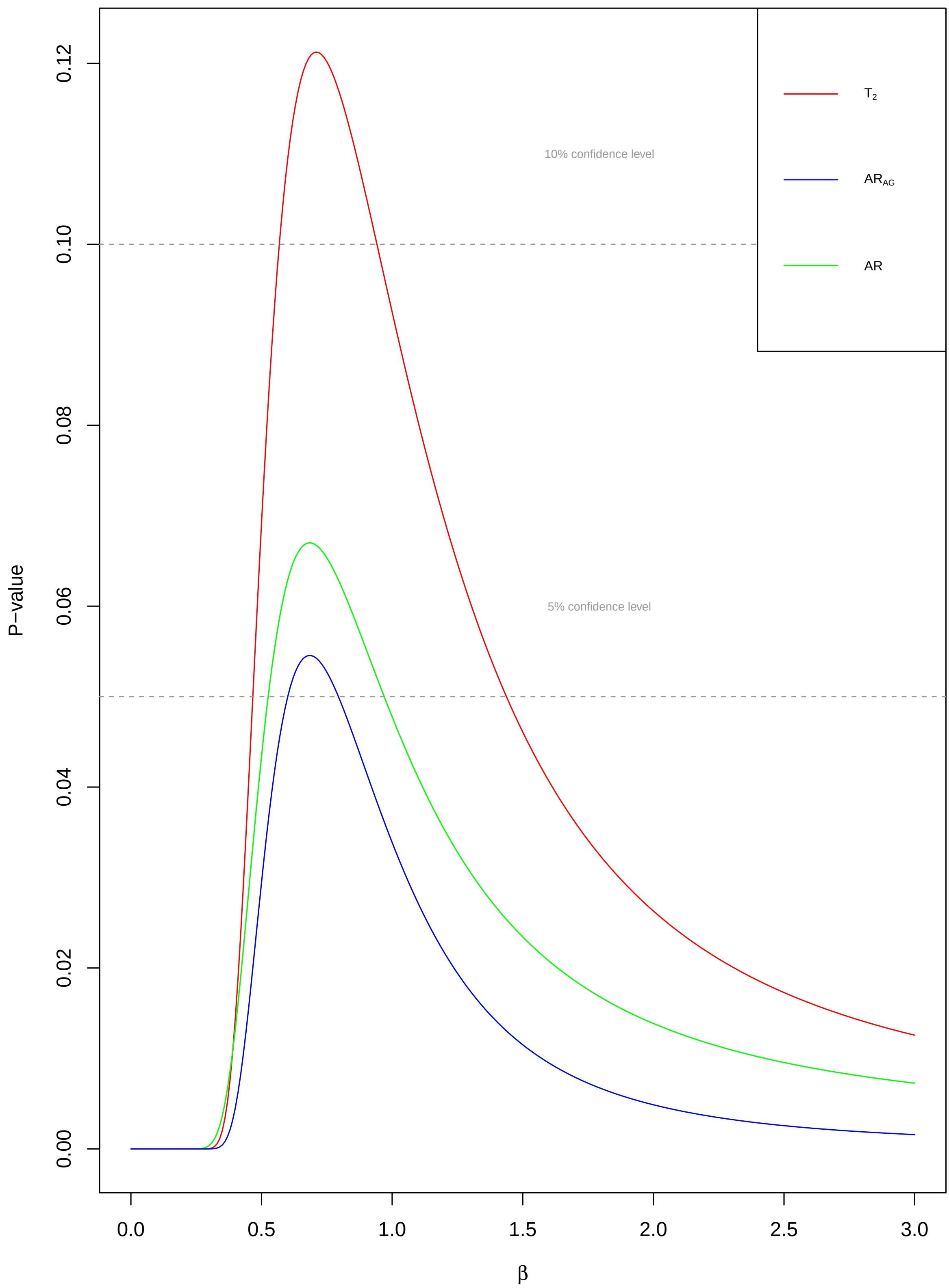

Figure 7: T_2, AR_{AG} and AR P-values for different values of β for the model with fourteen instruments.