Bootstrap Methods for Inference with Cluster-Sample IV Models

Keith Finlay* Leandro M. Magnusson †‡

US Census Bureau University of Western Australia

April 7, 2016

Abstract

Microeconomic data often have within-cluster dependence. This dependence affects standard error estimation and inference in the instrumental variables model. When the number of clusters is small, Wald and weak-instrument tests can be severely oversized. We examine the use of bootstrap methods and find that variants of the wild bootstrap perform well and reduce absolute size bias significantly, independent of instrument strength or cluster size. We also provide guidance in the choice among weak-instrument tests when data have cluster dependence. Two empirical examples illustrate the application of our wild bootstrap methods.

Keywords: two-stage least squares, instrumental variables, hypothesis testing, weak instruments, clustered errors, wild bootstrap.

JEL codes: C12, C15, C31.

*Center for Administrative Records Research and Applications, US Census Bureau, Room 5K132D, 4600 Silver Hill Road, Washington, DC 20233, USA, kfinlay@gmail.com.

†Department of Economics, 35 Stirling Highway M251, Business School, University of Western Australia, Crawley, WA 6009, Australia, phone: +61 (8) 6488-2924, fax: +61 (8) 6488-1016, leandro.magnusson@uwa.edu.au (corresponding author).

‡We would like to thank David Hendry, Jan Ondrich, Maurice Bun, Colin Cameron, participants at the 2012 Econometric Society Australasian Meeting and at the 2015 International Association for Applied Econometrics Conference. We also thank the National Science Foundation (award SMA-1004569) for support. The results and conclusions in this paper are the personal views of the authors and do not necessarily reflect the views of the US Census Bureau.

1 Introduction

Microeconometric data often have a group structure. When regression errors are correlated within these groups or clusters, it is well known that standard error estimates can be biased and hypothesis testing can be misleading. The common solution to this problem is to use cluster-robust standard error estimation methods that requires a large number of clusters. When the number of clusters is small, tests can be oversized even when cluster-robust methods are used (Cameron et al., 2008).

In the linear instrumental variable (IV) model, we show that the Wald and weak-instrument tests, which use the corrected cluster-robust standard errors, are size distorted when the number of clusters is small, under both strong and weak identification scenarios. For the weak instrument tests, we propose bootstrap techniques that perform well when the number of clusters is as few as 20 and the instruments are weak.

Our Monte Carlo simulations provide strong evidence of the benefit of bootstrap techniques in the linear IV model. We find rejection rates level as high as 0.50 with Wald tests in a strong instruments scenario when the nominal level is 5%. Cluster-robust versions of the Wald tests can reduce the rejection rates to 0.15 to 0.20, but never as low as the nominal size. Using our cluster *estimating equations* and *residuals* bootstraps, we get rejection rates that are very close to 0.05.

Recent work has highlighted the use of the bootstrap to improve inference when there is intra-cluster dependence. In the linear model with only exogenous covariates, Cameron et al. (2008) show that a variant of the wild bootstrap (Wu, 1986) with cluster-based sampling performs well in a variety of cases, and bootstrap tests dominate the asymptotic tests in terms of size. Using Edgeworth expansions, Kleibergen (2011) show that the bootstrap decreases the size distortion of weak instrument tests. Davidson and MacKinnon (2008) develop bootstrap techniques for linear IV models assuming that residuals are homoskedastic. Later, they extend the bootstrap by allowing residual heteroskedasticity but only at the individual level (Davidson and MacKinnon, 2010).

Gelbach et al. (2007) implement a variant of the wild cluster bootstrap of Cameron et al. (2008) for the Wald test in an instrumental variables setting. They examine its performance in Monte Carlo simulations and find that it performs well. But they assume

that instruments are strong and do not investigate the performance of weak instrument tests. In fact, in small samples our simulations indicate that the numerical accuracy of the bootstrapped Wald test may be even worse than the asymptotic weak instrument tests.

It is well known that bootstrap techniques cannot improve performance of the Wald test when instruments are weak (Moreira et al., 2009; Davidson and MacKinnon, 2008; Zhan, 2010). Our results show that weak-instrument-robust tests outperform the Wald test when instruments are undoubtedly strong (i.e., the concentration parameter is greater than 200). Thus, we recommend the use of weak instrument tests whether or not instruments are strong. The use of our bootstrap methods with weak-instrument tests provides a comprehensive and practical alternative for testing parameters in the linear IV model when data have cluster dependence.

We also investigate the performance of the first-stage F-test and the conservative version of effective F-test proposed by Olea and Pflueger (2013). Both tests test the null assumption that instruments are weak. Our simulation results shows that the first-stage F overrejects the null while the effective F-test underrejects it. However, the bootstrap version of these tests give similar rejection rates.

The paper proceeds as follows. First, we introduce our versions of weak-instrument-robust tests suitable for clustered residuals. Then, we describe our bootstrap techniques and the Monte Carlo experiments that illustrate the performance of these techniques. Two empirical applications of the bootstrap methods, one about civil conflict in Africa (Miguel et al., 2004) and the another about the role of institutions on economic performance (Acemoglu et al., 2001) end the paper. Some derivations and technical details are in the Appendix.

2 Cluster-robust inference

We consider the following limited-information cluster model with G clusters, indexed by g:

$$\begin{cases} \mathbf{y}_{1,g} = & \mathbf{y}_{2,g}\theta + \mathbf{x}_g\gamma + \mathbf{u}_g \\ \mathbf{y}_{2,g} = & \mathbf{w}_g\Pi_w + \mathbf{v}_g \end{cases} \quad \text{for } g = 1, \ldots, G, \tag{1}$$

where $\mathbf{y}_{1,g}$ is a $n_g \times 1$ vector, $\mathbf{y}_{2,g}$ is a $n_g \times p$ matrix of endogenous explanatory variables, $\mathbf{x}_g$ is a $n_g \times k_x$ vector of included instruments, $\mathbf{w}_g = [\mathbf{z}_g \colon \mathbf{x}_g]$ is a $n_g \times k_w$ matrix of instruments, $\mathbf{z}_g$ and $\mathbf{x}_g$ are $n_g \times k_z$ and $n_g \times k_x$ matrices of excluded and included instruments with $k_w = k_z + k_x$, and $\Pi_w = [\Pi_z' \; \Pi_x']'$ is a $k_w \times p$ matrix of first-stage, reduced-form parameters. We assume that

$$\mathrm{E}\left[(\mathbf{u}_g, \mathrm{vec}\,(\mathbf{v}_g))\,(\mathbf{u}_g, \mathrm{vec}\,(\mathbf{v}_g))'\right] = \Sigma_g = \begin{bmatrix} \Sigma_{u_g u_g} & \Sigma_{u_g v_g} \\ \Sigma_{v_g u_g} & \Sigma_{v_g v_g} \end{bmatrix}$$

and $(\mathbf{u}_g, \mathbf{v}_g)$ are independent across clusters. The equations in (1) have the following general form representation:

$$\mathbf{y}_1 = \mathbf{y}_2\theta + \mathbf{X}\gamma + \mathbf{u} \tag{2}$$

$$\mathbf{y}_2 = \mathbf{W}\Pi_w + \mathbf{V}, \tag{3}$$

where $\mathbf{y}_1$ is a $n \times 1$ vector, $\mathbf{y}_2$ is a $n \times p$ matrix of endogenous explanatory variables, $\mathbf{W} = [\mathbf{Z} : \mathbf{X}]$ is a $n \times k_w$ matrix of instruments, and $\mathbf{Z}$ and $\mathbf{X}$ are $n \times k_z$ and $n \times k_x$ matrices of excluded and included instruments respectively, with $k_w = k_x + k_z$ and $n = \sum_{g=1}^{G} n_g$.

We are interested in making inference about the structural vector parameter θ. For example, we may want to test the following hypotheses:

$$H_0^\theta : \theta = \theta_0 \text{ against } H_1^\theta : \theta \neq \theta_0.$$

The usual procedure for inference is the Wald test, defined as:

$$\left(\hat{\theta}_{\mathrm{IV}} - \theta_0\right)' \left(\widehat{\mathrm{Var}}(\hat{\theta}_{\mathrm{IV}})\right)^{-1} \left(\hat{\theta}_{\mathrm{IV}} - \theta_0\right), \tag{4}$$

where $\hat{\theta}_{\mathrm{IV}} = \left(\mathbf{y}_2'\mathrm{P}_{\mathrm{M}_\mathbf{X}\mathbf{Z}}\mathbf{y}_2\right)^{-1}\mathbf{y}_2'\mathrm{P}_{\mathrm{M}_\mathbf{X}\mathbf{Z}}\mathbf{y}_1$ is the two-stage least squares (TSLS) estimator, $\mathrm{P_A}$ $=\mathrm{A}\left(\mathrm{A}'\mathrm{A}\right)^{-1}\mathrm{A}'$, $\mathrm{M_A}$ $=\mathrm{I}-\mathrm{P_A}$, and $\widehat{\mathrm{Var}}(\hat{\theta}_{\mathrm{IV}})$ is an estimator of $\mathrm{Var}(\hat{\theta}_{\mathrm{IV}})$, the variance of $\hat{\theta}_{\mathrm{IV}}$.

When the errors are assumed to be independent and identically distributed (iid), the estimator of the variance is $\widehat{\mathrm{Var}}_h(\hat{\theta}_{\mathrm{IV}}) = \hat{\sigma}_u^2\left(\mathbf{y}_2'\mathrm{P}_{\mathrm{M}_\mathbf{X}\mathbf{Z}}\mathbf{y}_2\right)^{-1}$, where $\hat{\sigma}_u^2 = \frac{1}{n}\hat{\mathbf{u}}(\hat{\theta}_{\mathrm{IV}})'\hat{\mathbf{u}}(\hat{\theta}_{\mathrm{IV}})$, and $\hat{\mathbf{u}}(\hat{\theta}_{\mathrm{IV}})$ $=\mathrm{M}_\mathbf{X}(\mathbf{y}_1 - \mathbf{y}_2\hat{\theta}_{\mathrm{IV}})$. However, in the presence of intra-cluster dependence, even

if this dependence is negligible, we can use the arguments in Moulton (1990) to show that $\widehat{\mathrm{Var}}_h(\hat{\theta}_{\mathrm{IV}})$ underestimates the variance of $\hat{\theta}_{\mathrm{IV}}$.

The most commonly used estimator of $\mathrm{Var}(\hat{\theta}_{\mathrm{IV}})$ is an adaptation of the Huber-White heteroskedasticity robust sandwich estimator (White, 1980; Arellano, 1987), which does not impose any structure on the variance of error term:

$$\widehat{\mathrm{Var}}\left(\hat{\theta}_{\mathrm{IV}}\right) = \left(\mathbf{y}_2' \mathrm{P}_{\mathrm{M}_\mathbf{X}\mathbf{Z}} \mathbf{y}_2\right)^{-1} \left[\sum_{g=1}^{G} \left(\mathrm{P}_{\mathrm{M}_\mathbf{X}\mathbf{Z}}\mathbf{y}_2\right)_g' \widehat{\mathbf{\Sigma}}_g(\hat{\theta}_{\mathrm{IV}}) \left(\mathrm{P}_{\mathrm{M}_\mathbf{X}\mathbf{Z}}\mathbf{y}_2\right)_g\right] \left(\mathbf{y}_2' \mathrm{P}_{\mathrm{M}_\mathbf{X}\mathbf{Z}} \mathbf{y}_2\right)^{-1}, \quad (5)$$

where $\left(\mathrm{P}_{\mathrm{M}_\mathbf{X}\mathbf{Z}}\mathbf{y}_2\right)_g$ is the $n_g \times p$ submatrix $\mathrm{P}_{\mathrm{M}_\mathbf{X}\mathbf{Z}}\mathbf{y}_2$ associated to the g^{th} cluster, $\widehat{\mathbf{\Sigma}}_g(\hat{\theta}_{\mathrm{IV}}) = \widehat{\mathbf{u}}_g(\hat{\theta}_{\mathrm{IV}})\widehat{\mathbf{u}}_g'(\hat{\theta}_{\mathrm{IV}})$, and $\widehat{\mathbf{u}}_g(\hat{\theta}_{\mathrm{IV}})$ is the two-stage least squares (TSLS) residual of the g^{th} cluster. The sandwich estimator does not suffer from the underestimation described above and is general enough to accommodate different residual structures. The distributions of statistical tests based on the cluster-robust variance estimator, however, can differ considerably from their asymptotic distributions when the number of clusters is small.[1]

The consistency of the estimator $\hat{\theta}_{\mathrm{IV}}$ depends on whether instruments $\mathbf{Z}$ are sufficiently correlated with the explanatory endogenous variables $\mathbf{y}_2$ (i.e., $\|\Pi_z\| \neq 0$). Tests based on the first-stage F-statistic for detecting weak instruments, such as those proposed by Stock and Yogo (2005) and Sanderson and Windmeijer (2015), assume that residuals are homoskedastic. Bun and de Haan (2010) show that, with nonscalar error covariance structure, the standard and the cluster-robust versions of the first-stage F-test can overestimate the strength of instruments. The test for weak instruments proposed by Olea and Pflueger (2013) allows clustered residuals, but only one endogenous variable. In Section 4, our simulation results show that the overestimation of cluster-robust first-stage F-test is more severe when the number of clusters is small, while the conservative version of the Olea-Pflueger test underestimates instrument strength.

There are a number of statistical tests which have asymptotic and nominal size equality, independent of the presence of weak instruments, such as the AR-test (Anderson and Rubin, 1949), the score or KLM-test (Kleibergen, 2002, 2007), and the CLR-test (Moreira, 2003). These tests were originally developed under the assumption that the distribution of the errors is iid, but have been adapted to allow for arbitrary heteroskedasticity or

[1] See Cameron et al. (2008) for the simple linear regression model, and simulations in Section 4.

cluster dependence of the residuals (Chernozhukov and Hansen, 2008; Finlay and Magnusson, 2009).

We start by redefining equations (2) and (3) as:

$$\mathbf{Y}(\theta_0) = \mathbf{W}\delta_w(\theta_0) + \mathbf{e}(\theta_0) \tag{6}$$

$$\mathbf{y}_2 = \mathbf{W}\Pi_w + \mathbf{V}, \tag{7}$$

where $\mathbf{Y}(\theta_0) = \mathbf{y}_1 - \mathbf{y}_2\theta_0$, $\mathbf{e}(\theta_0) = \mathbf{u} + \mathbf{V}\mathrm{d}(\theta_0)$, $\delta_w(\theta_0) = \left[\delta_z(\theta_0)', \delta_x(\theta_0)'\right]' = \Pi_w \mathrm{d}(\theta_0) + \mathrm{H}\gamma$, $\Pi_w = [\Pi_z', \Pi_x']'$, $\mathrm{d}(\theta_0) = (\theta - \theta_0)$ and $\mathrm{H} = [0, \mathrm{I}_{k_x}]'$. Equations (6) and (7) can be further rewritten as:

$$\hat{\delta}_w(\theta_0) = \underbrace{\Pi_w \mathrm{d}(\theta_0) + \mathrm{H}\gamma}_{\delta_w(\theta_0)} + \left(\mathbf{W}'\mathbf{W}\right)^{-1}\mathbf{W}'\mathbf{e}(\theta_0) \tag{8}$$

$$\widehat{\Pi}_w = \Pi_w + \left(\mathbf{W}'\mathbf{W}\right)^{-1}\mathbf{W}'\mathbf{V},$$

where $\hat{\delta}_w(\theta_0) = [\hat{\delta}_z(\theta_0)', \hat{\delta}_x(\theta_0)']' = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Y}(\theta_0)$ and $\widehat{\Pi}_w = [\widehat{\Pi}_z', \widehat{\Pi}_x']' = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{y}_2$ are respectively the OLS estimators of the reduced form parameters in equations (6) and (7). The $k_w(p+1) \times k_w(p+1)$ "sandwich" matrix that corresponds to the cluster-robust estimator of variance of $[\hat{\delta}_w(\theta_0)', \hat{\pi}_w']$, where $\hat{\pi}_w = \mathrm{vec}(\widehat{\Pi}_w)$, has the form

$$\widehat{\Omega}(\theta_0) = \left(\mathrm{I}_{p+1} \otimes \mathbf{W}'\mathbf{W}\right)^{-1}\widehat{\Xi}(\theta_0)\left(\mathrm{I}_{p+1} \otimes \mathbf{W}'\mathbf{W}\right)^{-1}, \tag{9}$$

where $\widehat{\Xi}(\theta_0)$ is the estimator of the $k_w(p+1) \times k_w(p+1)$ variance matrix of $\mathrm{vec}(\mathbf{W}'\mathbf{e}(\theta_0), \mathbf{W}'\mathbf{V})$.[2]

Let us introduce four more statistics before presenting the cluster version of the weak-instrument tests:

$$\tilde{\lambda}_{\mathrm{KLM}}(\theta_0) = \widetilde{\Pi}_z(\theta_0)'\left[\widehat{\Omega}_{\delta_z\delta_z}(\theta_0)\right]^{-1}\hat{\delta}_z(\theta_0),$$

$$\widetilde{\Pi}_z(\theta_0) = \mathrm{mat}\left(\hat{\pi}_z - \widehat{\Omega}_{\pi_z\delta_z}(\theta_0)\left[\widehat{\Omega}_{\delta_z\delta_z}(\theta_0)\right]^{-1}\hat{\delta}_z(\theta_0)\right),$$

$$\widehat{\mathrm{Var}}\left(\tilde{\lambda}_{\mathrm{KLM}}(\theta_0)\right) = \widetilde{\Pi}_z(\theta_0)'\left[\widehat{\Omega}_{\delta_z\delta_z}(\theta_0)\right]^{-1}\widetilde{\Pi}_z(\theta_0), \quad \text{and}$$

[2] Details on Appendix A.

$$\widehat{\mathrm{Var}}\left(\tilde{\pi}_z\left(\theta_0\right)\right) = \widehat{\Omega}_{\pi_z\pi_z} - \widehat{\Omega}_{\pi_z\delta_z}\left(\theta_0\right)\left[\widehat{\Omega}_{\delta_z\delta_z}\left(\theta_0\right)\right]^{-1}\widehat{\Omega}_{\delta_z\pi_z}\left(\theta_0\right),$$

where $\tilde{\lambda}_{\mathrm{KLM}}\left(\theta_0\right)$ is the Lagrange multiplier of a constrained minimum-distance minimization problem[3], $\hat{\pi}_z = \mathrm{vec}(\widehat{\Pi}_z)$ is a $k_z p \times 1$ vector, mat is the rematricizing operator that maps the $k_z p \times 1$ vector $\mathrm{vec}(\widehat{\Pi}_z)$ into the $k_z \times p$ matrix $\widehat{\Pi}_z$, and $\widehat{\Omega}_{\pi_z\delta_z}\left(\theta_0\right)$ is the submatrix of $\widehat{\Omega}\left(\theta_0\right)$ associated to the covariance estimator of $(\hat{\pi}_z, \hat{\delta}_z\left(\theta_0\right))$. The estimators of the variances of $\tilde{\lambda}_{\mathrm{KLM}}\left(\theta_0\right)$ and $\tilde{\pi}_z\left(\theta_0\right)$, where $\tilde{\pi}_z\left(\theta_0\right) = \mathrm{vec}(\widetilde{\Pi}_z\left(\theta_0\right))$ are $\widehat{\mathrm{Var}}(\tilde{\lambda}_{\mathrm{KLM}}\left(\theta_0\right))$ and $\widehat{\mathrm{Var}}\left(\tilde{\pi}_z\left(\theta_0\right)\right)$, respectively.

We define the weak-instrument tests for the cluster-sample model as follows:

Definition 1. (Weak-instrument Tests with clustered residuals). The AR, KLM, and CRL statistics for testing the null hypothesis $H_0 : \mathrm{d}\left(\theta_0\right) = 0$ are, respectively:

$$\begin{aligned}
\Lambda_{\mathrm{AR}}\left(\theta_0\right) &\equiv \hat{\delta}_z\left(\theta_0\right)'\left[\widehat{\Omega}_{\delta_z\delta_z}\left(\theta_0\right)\right]^{-1}\hat{\delta}_z\left(\theta_0\right) \xrightarrow{d} \chi^2\left(k_z\right), \\
\Lambda_{\mathrm{KLM}}\left(\theta_0\right) &\equiv \tilde{\lambda}_{\mathrm{KLM}}\left(\theta_0\right)'\left[\widehat{\mathrm{Var}}\left(\tilde{\lambda}_{\mathrm{KLM}}\left(\theta_0\right)\right)\right]^{-1}\tilde{\lambda}_{\mathrm{KLM}}\left(\theta_0\right) \xrightarrow{d} \chi^2\left(p\right), \text{ and} \\
\Lambda_{\mathrm{CLR}}\left(\theta_0\right) &\equiv \left\{\frac{1}{2}\Lambda_{\mathrm{AR}}\left(\theta_0\right) - \mathrm{rk}\left(\theta_0\right) + \right. \\
&\qquad \left.\sqrt{\left[\Lambda_{\mathrm{AR}}\left(\theta_0\right) + \mathrm{rk}\left(\theta_0\right)\right]^2 - 4\left[\Lambda_{\mathrm{AR}}\left(\theta_0\right) - \Lambda_{\mathrm{KLM}}\left(\theta_0\right)\right] \times \mathrm{rk}\left(\theta_0\right)}\right\},
\end{aligned}$$

where $\mathrm{rk}\left(\theta_0\right)$ is a statistic for testing the rank of $\widetilde{\Pi}_z\left(\theta_0\right)$.

The symbol "$\xrightarrow{d}$"stands for convergence in distribution and $\chi^2\left(s\right)$ is the chi-squared distribution with s degrees of freedom. The CLR-statistic converges to a nonpivotal distribution; however, its critical values, for a given value of $\mathrm{rk}\left(\theta_0\right)$, can be simulated from independent $\chi^2(k_z - p)$ and $\chi^2(p)$ distributions. The tests converge independently of instrument strength.

The above statistics have the correct size asymptotically even when the structural parameter θ is not identified; however, they tests are inconsistent if $\Pi_z = 0$.[4] The $\Lambda_{\mathrm{AR}}\left(\theta_0\right)$ statistic tests if $\mathrm{d}\left(\theta_0\right) = 0$ indirectly by testing the assumption $H_0^{\delta} : \delta_z\left(\theta_0\right) = 0$ against $H_1^{\delta} : \delta_z\left(\theta_0\right) \neq 0$. The degrees of freedom of the $\Lambda_{\mathrm{AR}}\left(\theta_0\right)$-test's asymptotic distribution depends on k_z, the number of excluded instruments, which can be larger than p, the

[3] See derivation in Appendix B.

[4] The tests will not reject $H_0 : \mathrm{d}\left(\theta_0\right) = 0$ when $H_1 : \mathrm{d}\left(\theta_0\right) \neq 0$ is true, because the estimated value of δ_z will be close to 0, independent if $\|\theta - \theta_0\| > 0$.

number of tested parameters. The larger is the difference $k_z - p$, the less powerful is the $\Lambda_{\text{AR}}(\theta_0)$-test. The $\Lambda_{\text{KLM}}(\theta_0)$-test degrees of freedom is equal to the number of tested structural parameters, independent of the number of excluded instruments. Nevertheless, the $\Lambda_{\text{KLM}}(\theta_0)$-test, as a LM type of test, loses power at local extremum and inflection points of the $\Lambda_{\text{AR}}(\theta_0)$-test. The $\Lambda_{\text{CLR}}(\theta_0)$-test, because it is a function of the $\Lambda_{\text{AR}}(\theta_0)$-test, does not show spurious decline of power experienced by the $\Lambda_{\text{KLM}}(\theta_0)$-test.

Remark 2.1. The $\Lambda_{\text{AR}}(\theta_0)$-test also has a Lagrange-multiplier interpretation. It can be rewritten as

$$\Lambda_{\text{AR}}(\theta_0) \equiv \tilde{\lambda}_{\text{AR}}(\theta_0)' \left[\widehat{\text{Var}}(\tilde{\lambda}_{\text{AR}}(\theta_0))\right]^{-1} \tilde{\lambda}_{\text{AR}}(\theta_0),$$

where $\tilde{\lambda}_{\text{AR}}(\theta_0) = [\widehat{\Omega}_{\delta_z\delta_z}(\theta_0)]^{-1}\hat{\delta}_z(\theta_0)$ is the Lagrange multiplier of constrained-minimization problem (A-3), and $\widehat{\text{Var}}\left(\tilde{\lambda}_{\text{AR}}(\theta_0)\right)$ is the estimated variance of $\tilde{\lambda}_{\text{AR}}(\theta_0)$.

Remark 2.2. The *KLM*-test is originally derived from the continuously updating estimator (CUE) objective function. The first-order condition of that problem includes the derivative of the variance with respect to the parameters. The $\Lambda_{\text{KLM}}(\theta_0)$ is derived from the two-step minimum-distance estimator objective function, which does not require the derivative of the variance with respect to the parameters.[5] Because of the regression model is linear, the minimum-distance estimates of the untested well-identified parameters are the same as the GMM-CUE estimator under the null assumption (Goldberger and Olkin, 1971).

Remark 2.3. If θ is scalar, the rank statistic $\text{rk}(\theta_0)$ is defined as:

$$\text{rk}(\theta_0) \equiv \widetilde{\Pi}_z(\theta_0)' \left[\widehat{\text{Var}}\left(\widetilde{\Pi}_z(\theta_0)\right)\right]^{-1} \widetilde{\Pi}_z(\theta_0).$$

If θ is not scalar, then the rank statistics proposed by Kleibergen and Paap (2006) or Kleibergen and Mavroeidis (2009) should be used.

3 Bootstrap methods for the cluster-sample IV model

In many microeconometric applications, data have intra-cluster dependence in which the number of clusters are small and, consequently, the asymptotic results are a poor

[5] See the derivation of $\hat{\lambda}_{\text{KLM}}(\theta_0)$ in Appendix B.

approximation of the true distributions of the test statistics. For example, many papers in labor economics use research designs that rely on policy changes at the state level, in which the number of clusters is at most 51 in USA and 8 in Australia. Our simulations show that asymptotic tests that use cluster-robust variance estimators may under- or overreject with as many as 160 clusters. Therefore, bootstrapping them accordingly can also improve their performance in terms of size, when the number of clusters are small. We next discuss two classes of bootstrap methods for weak instrument tests in linear IV cluster model represented by system (1): the estimating equations and residual bootstraps.

Estimating equations (score) bootstrap

We begin the exposition by rewriting equation (8) as:

$$\hat{\delta}_w(\theta_0) = \delta_w(\theta_0) + \left(\mathbf{W}'\mathbf{W}\right)^{-1} \sum_{g=1}^{G} \underbrace{\mathbf{w}_g' \mathbf{e}_g(\theta_0)}_{\mathbf{h}_g(\theta_0)}.$$

A simple idea about bootstrapping the distributions of $\hat{\delta}_w(\theta_0)$ is based on perturbing the empirical distribution of the scores $\{\mathbf{h}_g(\theta_0)\}_{g=1}^{G}$, but keeping the Hessian $(\mathbf{W}'\mathbf{W})^{-1}$ fixed. Hu and Zidek (1995) denote this type of bootstrap the estimating equations (EE) bootstrap.[6]

Under $H_0 : \mathrm{d}(\theta_0) = 0$, a candidate bootstrap estimator for $\delta_w(\theta_0)$ is:

$$\tilde{\delta}_w^*(\theta_0) = \tilde{\delta}_w(\theta_0) + \left(\mathbf{W}'\mathbf{W}\right)^{-1} \sum_{g=1}^{G} \tilde{\mathbf{h}}_g^*(\theta_0), \tag{10}$$

where $\tilde{\delta}_w(\theta_0) = (0, \tilde{\delta}_x(\theta_0))$, and $\tilde{\delta}_x(\theta_0) = \widehat{\delta}_x(\theta_0) - \widehat{\Omega}_{\delta_x\delta_z}(\theta_0)\left[\widehat{\Omega}_{\delta_z\delta_z}\right]^{-1}\widehat{\delta}_z(\theta_0)$. The $\tilde{\delta}_w(\theta_0)$ is the estimator of $\delta_w(\theta_0)$ derived from equation (A-3). The sequence of bootstrap scores $\left\{\tilde{\mathbf{h}}_g^*(\theta_0)\right\}_{g=1}^{G}$ is sampled with replacement from the recentered scores $\{\tilde{\mathbf{h}}_g^r(\theta_0)\}_{g=1}^{G}$, defined as:

$$\tilde{\mathbf{h}}_g^r(\theta_0) = \tilde{\mathbf{h}}_g(\theta_0) - \frac{n_g}{n}\sum_{g=1}^{G} \tilde{\mathbf{h}}_g(\theta_0),$$

[6] See also Hu and Kalbfleisch (2000) and Kline and Santos (2012).

where $\tilde{\mathbf{h}}_g(\theta_0) = \mathbf{w}_g'\tilde{\mathbf{e}}_g(\theta_0)$, and $\tilde{\mathbf{e}}_g(\theta_0) = \mathbf{Y}_g(\theta_0) - \mathbf{w}_g\tilde{\delta}_w(\theta_0)$.[7]

The estimator of the variance of $\tilde{\delta}_w^*(\theta_0)$, denoted by $\widetilde{\Omega}^*_{\delta_w\delta_w}(\theta_0)$, is a function of $\left\{\tilde{\mathbf{h}}_g^*(\theta_0)\right\}_{g=1}^G$ and does not depend on $\tilde{\delta}_w^*(\theta_0)$ itself. This implies a computational gain of the EE bootstrap over the residual-type bootstraps discussed below.

We define the bootstrap estimator of $\tilde{\lambda}_{\text{KLM}}(\theta_0)$ conditional on $\widetilde{\Pi}_z(\theta_0)$ as:

$$\tilde{\lambda}^*_{\text{KLM}}(\theta_0) = \widetilde{\Pi}_z(\theta_0)'\left[\widetilde{\Omega}^*_{\delta_z\delta_z}(\theta_0)\right]^{-1}\tilde{\delta}_z^*(\theta_0), \tag{11}$$

where $\widetilde{\Omega}^*_{\delta_z\delta_z}(\theta_0)$ is the block variance of $\widetilde{\Omega}^*_{\delta_w\delta_w}(\theta_0)$ associated with the estimator $\tilde{\delta}_z^*(\theta_0)$ obtained from equation (10). The Λ_{CLR}-test, conditional on $\text{rk}(\theta_0)$, are functions of the Λ_{AR}-, and Λ_{KLM} tests. Therefore, bootstrap realizations of the Λ_{CLR}-test are generated from the bootstrap realizations of the Λ_{AR}- and Λ_{KLM}-tests.

The general algorithm for computing the bootstrap tests are:

1. Compute $\Lambda_{\text{AR}}(\theta_0)$, $\Lambda_{\text{KLM}}(\theta_0)$, and $\Lambda_{\text{CLR}}(\theta_0)$ and save the estimates of $\widetilde{\Pi}_z(\theta_0)$ and $\text{rk}(\theta_0)$.

2. For $b = 1, \ldots, B$ bootstrap simulations:

   (a) Sample $\{\omega_g\}_{g=1}^G$, a sequence of bootstrap weights, and define the bootstrap score realizations as:

   $$\left\{\tilde{\mathbf{h}}_1^*(\theta_0), \ldots, \tilde{\mathbf{h}}_G^*(\theta_0)\right\} = \left\{\omega_1\tilde{\mathbf{h}}_1^r(\theta_0), \ldots, \omega_G\tilde{\mathbf{h}}_G^r(\theta_0)\right\}.$$

   (b) Compute $\tilde{\delta}_w^*(\theta_0)$ and its associated variance $\widetilde{\Omega}^*_{\delta_w\delta_w}(\theta_0)$.

   (c) Compute $\tilde{\lambda}^*_{\text{KLM}}(\theta_0)$, given by equation (11), and its variance $\widehat{\text{Var}}(\tilde{\lambda}^*_{\text{KLM}}(\theta_0))$, which is:

   $$\widehat{\text{Var}}\left(\tilde{\lambda}^*_{\text{KLM}}(\theta_0)\right) = \widetilde{\Pi}_z(\theta_0)'\left[\widetilde{\Omega}^*_{\delta_z\delta_z}(\theta_0)\right]^{-1}\widetilde{\Pi}_z(\theta_0).$$

   (d) The b^{th} bootstrap tests are:

   $$\tilde{\Lambda}^*_{\text{AR},b}(\theta_0) = \tilde{\delta}_z^*(\theta_0)'\left[\widetilde{\Omega}^*_{\hat{\delta}_z\hat{\delta}_z}(\theta_0)\right]^{-1}\tilde{\delta}_z^*(\theta_0)$$
   $$\tilde{\Lambda}^*_{\text{KLM},b}(\theta_0) = \tilde{\lambda}^*_{\text{KLM}}(\theta_0)'\left[\widehat{\text{Var}}\left(\tilde{\lambda}^*_{\text{KLM}}(\theta_0)\right)\right]^{-1}\tilde{\lambda}^*_{\text{KLM}}(\theta_0), \text{ and}$$

[7] If the number of observations per cluster is the same, then $\frac{n_g}{n} = \frac{1}{G}$.

$$\tilde{\Lambda}^*_{\text{CLR},b}(\theta_0, b) = \left\{ \frac{1}{2}\tilde{\Lambda}^*_{\text{AR},b}(\theta_0) - \text{rk}(\theta_0) + \sqrt{\left[\tilde{\Lambda}^*_{\text{AR},b}(\theta_0) + \text{rk}(\theta_0)\right]^2 - 4\left[\tilde{\Lambda}^*_{\text{AR},b}(\theta_0) - \tilde{\Lambda}^*_{\text{KLM},b}(\theta_0, b)\right] \times \text{rk}(\theta_0)} \right\}.$$

3. The bootstrap p-value for a test is:

$$\tilde{p}^*\text{-value} = \frac{1}{B}\sum_{b=1}^{B} I\left(\tilde{\Lambda}^*_{[\cdot],b}(\theta_0) > \Lambda_{[\cdot]}(\theta_0)\right),$$

where $I(\cdot)$ is the indicator function and $\Lambda_{[\cdot]}$ represents the Λ_{AR}-, Λ_{KLM}-, or Λ_{CLR}-test. Reject the assumption if $\tilde{p}^*$-value is smaller the desired significance level of the test.

Next we discuss two types of estimating equation bootstraps.

Definition 2. (Estimating Equation (EE) Bootstrap) Let $\{\omega_g\}_{g=1}^G$ be a sequence of bootstrap weights. Conditional on $\widetilde{\Pi}_z(\theta_0)$ and $\text{rk}(\theta_0)$, the EE bootstrap weak instrument tests are computed from the bootstrap score sequence $\left\{\tilde{\mathbf{h}}^*_g(\theta_0)\right\}_{g=1}^G = \{\omega_g \tilde{\mathbf{h}}^r_g(\theta_0)\}_{g=1}^G$. We consider two bootstrap weights:

1. $\{\omega_g\}_{g=1}^G$ are sampled from a multinomial distribution, so that

$$\Pr\left(\tilde{\mathbf{h}}^*_g(\theta_0) = \tilde{\mathbf{h}}_j(\theta_0)\right) = \frac{1}{G}, \quad j = 1, \ldots, G.$$

2. $\{\omega_g\}_{g=1}^G$ be an iid sequence sampled from a distribution satisfying $E[\omega_g] = 0$ and $\text{Var}(\omega_g) = 1$. We discuss the specific distributions below.

The EE bootstrap with multinomial weights is the same as the bootstrap algorithm 1 of Kleibergen (2011) for GMM models. The second bootstrap is similar to the wild score bootstrap proposed by Kline and Santos (2012). They assume, however, that the tested parameter is identified and consistently estimated. Therefore, the empirical score is obtained by replacing the tested parameters by their two-step GMM estimates. Clearly, the 2-step GMM estimator is biased when instruments are weak.

Remark 3.1. Sampling the score from $\left\{\omega_g \tilde{\mathbf{h}}^r_g(\theta_0)\right\}_{g=1}^G$ corresponds to sampling the residuals from $\{\bar{\omega}_g \tilde{\mathbf{e}}_g(\theta_0)\}_{g=1}^G$ where $\bar{\omega}_g = \omega_g - \bar{\omega}$ and $\bar{\omega} = \left(\sum_{g=1}^G \frac{\omega_g n_g}{n}\right)$. We can interpret

$\{\bar{\omega}_g\}_{g=1}^G$ as the sequence of adjusted bootstrap weights.

Remark 3.2. Define $\acute{\delta}_w(\theta_0) = (0, \acute{\delta}_x(\theta_0))$, where $\acute{\delta}_x(\theta_0) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}(\theta_0)$, and $\hat{\delta}_w(\theta_0) = (\hat{\delta}_z(\theta_0), \hat{\delta}_x(\theta_0))$, where $\hat{\delta}_w(\theta_0) = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Y}(\theta_0)$. The estimators $\acute{\delta}_w(\theta_0)$ and $\hat{\delta}_w(\theta_0)$ could replace $\tilde{\delta}_w(\theta_0)$ in equation (10). The bootstrap scores are generated as before, and recentering is unnecessary if a constant is included in $\mathbf{x}_g$. In the case of using $\hat{\delta}_w(\theta_0)$, $\hat{\delta}_z(\theta_0)$ is the mean of the distribution of the bootstrap estimator $\hat{\delta}_z^*(\theta_0)$. Therefore, in computing the bootstrap version of the tests, $\tilde{\delta}_z^*(\theta_0)$ should be replaced by $\hat{\delta}_z^*(\theta_0) - \hat{\delta}_z(\theta_0)$.

Residual bootstraps

By resampling the estimated residuals, we can generate bootstrap samples. The bootstrap weak instrument tests are computed in the same way as the asymptotic ones, using the bootstrap sample in place of the original data. We consider two types of residual bootstraps. In the first version, we bootstrap the residuals only from the auxiliary regression, see equation (6). In the second version, we sample residuals from equations (6) and (7) simultaneously.

Single-equation residual bootstrap

Let $\{\acute{\mathbf{e}}_g(\theta_0)\}_{g=1}^G$ be a sequence of residuals, where $\acute{\mathbf{e}}_g(\theta_0) = \mathbf{Y}_g(\theta_0) - \mathbf{w}_g\acute{\delta}_w(\theta_0)$ is the $n_g \times 1$ vector associated to the g^{th} cluster, and $\acute{\delta}_w(\theta_0)$ is defined in Remark 3.2. We define $\acute{\mathbf{Y}}^*(\theta_0)$, the bootstrap realization of $\mathbf{Y}(\theta_0)$, as:

$$\acute{\mathbf{Y}}^*(\theta_0) = \mathbf{W}\acute{\delta}_w(\theta_0) + \acute{\mathbf{e}}^*(\theta_0),$$

where $\acute{\mathbf{e}}^*(\theta_0) = (\acute{\mathbf{e}}_1^*(\theta_0)', \ldots, \acute{\mathbf{e}}_G^*(\theta_0)')'$, $\{\acute{\mathbf{e}}_g^*(\theta_0)\}_{g=1}^G = \{\omega_g\acute{\mathbf{e}}_g(\theta_0)\}_{g=1}^G$. We define the bootstrap estimates of $\delta_w(\theta_0)$ and $\lambda_{\text{KLM}}(\theta_0)$ as:

$$\acute{\delta}_w^*(\theta_0) = \begin{bmatrix} \acute{\delta}_z^*(\theta_0)' & \acute{\delta}_x^*(\theta_0)' \end{bmatrix}' = \left(\mathbf{W}'\mathbf{W}\right)^{-1}\mathbf{W}'\acute{\mathbf{Y}}^*(\theta_0), \text{ and}$$
$$\acute{\lambda}_{\text{KLM}}^*(\theta_0) = \widetilde{\Pi}_z(\theta_0)'\left[\acute{\Omega}_{\delta_z^*\delta_z^*}^*(\theta_0)\right]^{-1}\acute{\delta}_z^*(\theta_0),$$

where $\acute{\Omega}^*_{\delta^*_z\delta^*_z}(\theta_0)$ is an estimator of variance matrix $\acute{\delta}^*_z(\theta_0)$.[8] This differs from the EE bootstrap in that $\acute{\Omega}^*_{\delta^*_z\delta^*_z}(\theta_0)$ is a function of $\acute{\delta}^*_z(\theta_0)$.

The steps for implementing the *residual* bootstrap are similar to those for the EE bootstrap. In step 2a, we use a sequence of $\left\{\acute{\mathbf{e}}^*_g(\theta_0)\right\}_{g=1}^G = \{\omega_g \acute{\mathbf{e}}_g(\theta_0)\}_{g=1}^G$. In steps 2b and 2c, we compute $\acute{\delta}^*_w(\theta_0)$, $\acute{\lambda}^*_{\text{KLM}}(\theta_0)$, $\acute{\Omega}^*_{\delta^*_z\delta^*_z}(\theta_0)$, and

$$\widehat{\text{Var}}\left(\acute{\lambda}^*(\theta_0)\right) = \widetilde{\Pi}_z(\theta_0)' \left[\acute{\Omega}^*_{\delta^*_z\delta^*_z}(\theta_0)\right]^{-1} \widetilde{\Pi}_z(\theta_0).$$

The b^{th} bootstrap $\acute{\Lambda}^*_{\text{AR},b}(\theta_0)$, $\acute{\Lambda}^*_{\text{KLM},b}(\theta_0)$ and $\acute{\Lambda}^*_{\text{CLR},b}(\theta_0)$ tests are obtained by replacing $\tilde{\delta}^*_z(\theta_0)$, $\widetilde{\Omega}^*_{\hat{\delta}_z\hat{\delta}_z}(\theta_0)$, $\tilde{\lambda}^*_{\text{KLM}}(\theta_0)$, and $\widehat{\text{Var}}\left(\tilde{\lambda}^*_{\text{KLM}}(\theta_0)\right)$ with $\acute{\delta}^*_w(\theta_0)$, $\acute{\Omega}^*_{\delta^*_z\delta^*_z}(\theta_0)$, $\acute{\lambda}^*_{\text{KLM}}(\theta_0)$, and $\widehat{\text{Var}}\left(\acute{\lambda}^*(\theta_0)\right)$ in the formulas of step 2d, respectively.

Definition 3. (Single-equation residual (SE) bootstrap): Let $\{\omega_g\}_{g=1}^G$ be a sequence of bootstrap weights satisfying $E[\omega_g] = 0$ and $\text{Var}(\omega_g) = 1$. Conditional on $\widetilde{\Pi}_z(\theta_0)$ and $\text{rk}(\theta_0)$, the bootstrap data generating process (dgp) is:

1. Inefficient SE (SE-in):

   $\left\{\acute{\mathbf{Y}}^*_g(\theta_0)\right\}_{g=1}^G = \left\{\mathbf{w}_g \acute{\delta}_w(\theta_0) + \acute{\mathbf{e}}^*_g(\theta_0)\right\}_{g=1}^G$, where $\left\{\acute{\mathbf{e}}^*_g(\theta_0)\right\}_{g=1}^G = \{\omega_g \acute{\mathbf{e}}_g(\theta_0)\}_{g=1}^G$; and

2. New efficient SE (SE-neff):

   $\left\{\tilde{\mathbf{Y}}^*_g(\theta_0)\right\}_{g=1}^G = \left\{\mathbf{w}_g \tilde{\delta}_w(\theta_0) + \tilde{\mathbf{e}}^*_g(\theta_0)\right\}_{g=1}^G$, where $\left\{\tilde{\mathbf{e}}^*_g(\theta_0)\right\}_{g=1}^G = \{\omega_g \tilde{\mathbf{e}}_g(\theta_0)\}_{g=1}^G$.

Remark 3.3. If a constant is not included in $\mathbf{x}_g$, then the fitted residuals $\{\acute{\mathbf{e}}_g(\theta_0)\}_{g=1}^G$ should be recentered.

Remark 3.4. As in Remark 3.2, we could use $\left\{\hat{\mathbf{e}}^*_g(\theta_0)\right\}_{g=1}^G = \{\omega_g \hat{\mathbf{e}}_g(\theta_0)\}_{g=1}^G$, where $\hat{\mathbf{e}}_g = \mathbf{Y}_g(\theta_0) - \mathbf{w}_g \hat{\delta}_w(\theta_0)$ to generate bootstrap realizations of $\mathbf{Y}(\theta_0)$. Then, when computing the bootstrap weak instrument tests, $\hat{\delta}^*_z(\theta_0) - \hat{\delta}_z(\theta_0)$ should be in place of $\acute{\delta}^*_z(\theta_0)$, where $\hat{\delta}^*_z(\theta_0)$ is the bootstrap estimator. In this case, the only difference between the EE and SE bootstrap weak instrument test is the bootstrap estimator of the variance of $\delta^*_z(\theta_0)$. For the EE bootstrap, we use $\left\{\hat{\mathbf{e}}^*_g(\theta_0)\right\}_{g=1}^G$, while for the SE bootstrap, we use use $\{\hat{\mathbf{e}}^*_{\text{b},g}(\theta_0)\}_{g=1}^G$, where $\hat{\mathbf{e}}^*_{\text{b},g}(\theta_0) = \hat{\mathbf{Y}}^*_g(\theta_0) - \mathbf{w}_g \hat{\delta}^*_w(\theta_0)$ to estimate the variance.

[8] See definition in Appendix A.

Multi-equation residual bootstraps

The bootstrap dgp for the limited-information model system (1) can be set as:

$$\begin{cases} \acute{\mathbf{Y}}_g^*(\theta_0) = & \mathbf{w}_g \acute{\delta}_w(\theta_0) + \acute{\mathbf{e}}_g^*(\theta_0) \\ \hat{\mathbf{y}}_{2,g}^* = & \mathbf{w}_g \widehat{\Pi}_w + \hat{\mathbf{v}}_g^*, \end{cases}$$

where $\left\{\acute{\mathbf{e}}_g^*(\theta_0), \hat{\mathbf{v}}_g^*\right\}_{g=1}^G = \left\{\omega_g \acute{\mathbf{e}}_g(\theta_0), \omega_g \hat{\mathbf{v}}_g\right\}_{g=1}^G$, $\acute{\mathbf{e}}_g(\theta_0) = \mathbf{Y}_g(\theta_0) - \mathbf{w}_g \acute{\delta}_w(\theta_0)$ and $\hat{\mathbf{v}}_g = \mathbf{y}_{2,g} - \mathbf{w}_g \widehat{\Pi}_w$.

Davidson and MacKinnon (2010) propose a more efficient bootstrap procedure that incorporates information about the correlation structure between $\mathbf{e}_g$ and $\mathbf{v}_g$ when estimating the first-stage residuals. They first estimate Π_w using the following auxiliary regression model

$$\mathbf{y}_2 = \mathbf{W}\Pi_w + \acute{\mathbf{e}}(\theta_0)\Gamma + \text{residuals}, \tag{12}$$

to obtain $\acute{\mathbf{V}}(\theta_0) = \mathbf{y}_2 - \mathbf{W}\acute{\Pi}_w(\theta_0)$, where $\acute{\Pi}_w(\theta_0)$ is the OLS estimator of Π_w. If the residuals are homoskedastic, $\acute{\Pi}_w(\theta_0)$ would be equivalent to the three-stage least squares (3SLS) or limited-information maximum likelihood estimator (under residual normality). Under cluster residuals, however, $\widetilde{\Pi}_w(\theta_0)$, the estimator of Π_w derived from equation (A-1), incorporates information about the cluster nature of the residual covariance matrix. Therefore, the cluster residual for the bootstrap dgp is

$$\left\{\acute{\mathbf{e}}_g^*(\theta_0), \tilde{\mathbf{v}}_g^*(\theta_0)\right\}_{g=1}^G = \left\{\omega_g \acute{\mathbf{e}}_g(\theta_0), \omega_g \tilde{\mathbf{v}}_g(\theta_0)\right\}_{g=1}^G,$$

where $\tilde{\mathbf{v}}_g(\theta_0) = \mathbf{y}_{2,g} - \mathbf{w}_g \widetilde{\Pi}_w(\theta_0)$.

Let $\tilde{\gamma}(\theta_0)$ be the continuous updating estimator derived from equation (A-1). A new bootstrap procedure uses $\tilde{\mathbf{e}}_g(\theta_0) = \mathbf{Y}_g(\theta_0) - \mathbf{x}_g \tilde{\gamma}(\theta_0)$ in place $\mathbf{e}_g(\theta_0)$ for the bootstrap dgp:

$$\left\{\tilde{\mathbf{e}}_g^*(\theta_0), \tilde{\mathbf{v}}_g^*(\theta_0)\right\}_{g=1}^G = \left\{\omega_g \tilde{\mathbf{e}}_g(\theta_0), \omega_g \tilde{\mathbf{v}}_g(\theta_0)\right\}_{g=1}^G.$$

The multi-equation residual bootstrap weak instrument tests are computed using the same formulas as the asymptotic tests described in Section 2 using the bootstrap sample in place of the original one. In contrast with the EE and SE bootstraps, the estimated of

values of $\widetilde{\Pi}_z(\theta_0)$ in step 1 do not need to be retained.

The three discussed multi-equation residual bootstraps are:

Definition 4. (Multi-equation residual (ME) bootstrap): Let $\{\omega_g\}_{g=1}^G$ be a sequence of bootstrap weights satisfying $E[\omega_g] = 0$ and $\mathrm{Var}(\omega_g) = 1$. The bootstrap dgp (dgp) for $\{\mathbf{Y}_g(\theta_0), \mathbf{y}_{2,g}\}_{g=1}^G$ is:

1. Inefficient ME (ME-in):

$$\left\{\acute{\mathbf{Y}}_g^*(\theta_0), \hat{\mathbf{y}}_{2,g}^*\right\}_{g=1}^G = \left\{\mathbf{w}_g\acute{\delta}_w(\theta_0) + \acute{\mathbf{e}}_g^*(\theta_0), \mathbf{w}_g\widehat{\Pi}_w + \hat{\mathbf{v}}_g^*\right\}_{g=1}^G, \text{ where}$$

$$\left\{\acute{\mathbf{e}}_g^*(\theta_0), \hat{\mathbf{v}}_g^*\right\}_{g=1}^G = \{\omega_g\acute{\mathbf{e}}_g(\theta_0), \omega_g\hat{\mathbf{v}}_g\}_{g=1}^G;$$

2. Efficient ME (ME-eff):

$$\left\{\acute{\mathbf{Y}}_g^*(\theta_0), \tilde{\mathbf{y}}_{2,g}^*(\theta_0)\right\}_{g=1}^G = \left\{\mathbf{w}_g\acute{\delta}_w(\theta_0) + \acute{\mathbf{e}}_g^*(\theta_0), \mathbf{w}_g\widetilde{\Pi}_w(\theta_0) + \tilde{\mathbf{v}}_g^*(\theta_0)\right\}_{g=1}^G, \text{ where}$$

$$\left\{\acute{\mathbf{e}}_g^*(\theta_0), \tilde{\mathbf{v}}_g^*(\theta_0)\right\}_{g=1}^G = \{\omega_g\acute{\mathbf{e}}_g(\theta_0), \omega_g\tilde{\mathbf{v}}_g(\theta_0)\}_{g=1}^G; \text{ and}$$

3. New efficient ME (ME-neff):

$$\left\{\tilde{\mathbf{Y}}_g^*(\theta_0), \hat{\mathbf{y}}_{2,g}^*(\theta_0)\right\}_{g=1}^G = \left\{\mathbf{x}_g\tilde{\gamma}(\theta_0) + \tilde{\mathbf{e}}_g^*(\theta_0), \mathbf{w}_g\widehat{\Pi}_w(\theta_0) + \tilde{\mathbf{v}}_g^*(\theta_0)\right\}_{g=1}^G, \text{ where}$$

$$\left\{\tilde{\mathbf{e}}_g^*(\theta_0), \tilde{\mathbf{v}}_g^*(\theta_0)\right\}_{g=1}^G = \{\omega_g\tilde{\mathbf{e}}_g(\theta_0), \omega_g\tilde{\mathbf{v}}_g(\theta_0)\}_{g=1}^G.$$

Remark 3.5. Since the $\Lambda_{\mathrm{AR}}(\theta_0)$-test does not depend on the first-stage equation, the SE-in, ME-in, and ME-eff bootstraps are equal. For the same reason, the SE-neff and ME-neff bootstraps are also equal.

Remark 3.6. The previous bootstrap procedures take the rank statistic as given. In this case, the ME bootstrap version of the CLR-test as explained in Davidson and MacKinnon (2008) requires a double bootstrap procedure: one bootstrap for the rank statistic and a second bootstrap conditional on the bootstrapped rank statistic. Therefore, we have not computed the ME version of this statistic.

Other bootstrap methods

Davidson and MacKinnon bootstrap for weak instruments

In Section 4, we compare the performance of the proposed bootstraps to the *wild restricted efficient (WRE) residual bootstrap* proposed by Davidson and MacKinnon (2010). The WRE bootstrap simulates the distribution of the original AR- and KLM-tests, which are defined respectively as:

$$\text{AR}(\theta_0) = \frac{n-k_w}{k_z} \frac{\mathbf{Y}(\theta_0)' \text{M}_{\mathbf{X}} \mathbf{Z} (\mathbf{Z}' \text{M}_{\mathbf{X}} \mathbf{Z})^{-1} \mathbf{Z}' \text{M}_{\mathbf{X}} \mathbf{Y}(\theta_0)}{\mathbf{Y}(\theta_0)' \text{M}_{\mathbf{W}} \mathbf{Y}(\theta_0)}, \quad \text{and} \tag{13}$$

$$\text{KLM}(\theta_0) = (n-k_w) \frac{\mathbf{Y}(\theta_0)' \text{P}_{\text{M}_{\mathbf{X}} \mathbf{Z} \acute{\Pi}_w(\theta_0)} \mathbf{Y}(\theta_0)}{\mathbf{Y}(\theta_0)' \text{M}_{\mathbf{W}} \mathbf{Y}(\theta_0)}. \tag{14}$$

The WRE bootstrap dgp are generated from residuals sampled at individual level. We define as Davidson and MacKinnon bootstrap (DM bootstrap) the WRE bootstrap method with dgp are generated by residuals sampled at cluster instead, i.e.

$$\left\{ \acute{\mathbf{Y}}_g^*(\theta_0), \acute{\mathbf{y}}_{2,g}^* \right\}_{g=1}^G = \left\{ \mathbf{w}_g \acute{\delta}_w(\theta_0) + \acute{\mathbf{e}}_g^*(\theta_0), \mathbf{w}_g \acute{\Pi}_w(\theta_0) + \acute{\mathbf{v}}_g^*(\theta_0) \right\}_{g=1}^G,$$

where $\left\{ \acute{\mathbf{e}}_g^*(\theta_0), \acute{\mathbf{v}}_g^*(\theta_0) \right\}_{g=1}^G = \left\{ \omega_g \acute{\mathbf{e}}_g(\theta_0), \omega_g \acute{\mathbf{v}}_g(\theta_0) \right\}_{g=1}^G$, with $\acute{\mathbf{v}}_g(\theta_0) = \mathbf{y}_{2,g} - \mathbf{w}_g \acute{\Pi}_w(\theta_0)$.

If the residuals are homoskedastic, then the AR and KLM of equations (13) and (14) are distributed asymptotically as $F_{(\infty,k_z)}$ and $\chi^2(p)$, respectively. Although the AR-test is not pivotal if the residuals are heteroskedastic, Davidson and MacKinnon (2010) show that the limiting distributions of $n^{-\frac{1}{2}} \mathbf{Z}' \text{M}_{\mathbf{X}} \mathbf{Y}(\theta_0)$ and $n^{-\frac{1}{2}} \mathbf{Z}' \text{M}_{\mathbf{X}} \acute{\mathbf{Y}}^*(\theta_0)$ are equal. The same is true for the probability limits of $n^{-1} \mathbf{Y}(\theta_0)' \text{M}_{\mathbf{W}} \mathbf{Y}(\theta_0)$ and $n^{-1} \acute{\mathbf{Y}}^*(\theta_0)' \text{M}_{\mathbf{W}} \acute{\mathbf{Y}}^*(\theta_0)$. Therefore, since the AR-statistic and its wild bootstrap counterpart converge to the same limit distribution, their bootstrap method is valid for the AR-test in the presence of heteroskedastic errors. They also show that their bootstrap for the KLM-test is correct if the concentration parameter is high. We can use the same arguments in Davidson and MacKinnon (2010) to show that the DM bootstrap for the AR-test with cluster residuals is consistent, assuming that the number of clusters is increasing as the sample size increases and the number of observations within clusters is constant. However, the same argument cannot be applied in the number of observations within each cluster increases at the same

rate as the number of clusters.

Bootstrap for Wald tests

For the Wald test, we consider three types of bootstrap methods. Two of them are similar to the multiequation residual bootstraps, in which the bootstrap residuals are sampled from:

$$\left\{\hat{\mathbf{u}}_g^*, \hat{\mathbf{v}}_g^*\right\}_{g=1}^G = \{\omega_g \hat{\mathbf{u}}_g, \omega_g \widehat{\mathbf{v}}_g\}_{g=1}^G \quad \text{and} \quad \left\{\acute{\mathbf{e}}_g^*(\theta_0), \tilde{\mathbf{v}}_g^*(\theta_0)\right\}_{g=1}^G = \{\omega_g \acute{\mathbf{e}}_g(\theta_0), \omega_g \tilde{\mathbf{v}}_g(\theta_0)\}_{g=1}^G,$$

where $\hat{\mathbf{u}}_g$ is the TSLS residual. Let $\hat{\theta}_{\text{IV}}^*$ be the bootstrap TSLS estimator obtained from the bootstrap sample generated by the residual sequence $\left\{\hat{\mathbf{u}}_g^*, \hat{\mathbf{v}}_g^*\right\}_{g=1}^G$. The Wald multiequation instrumental variable (ME-IV) bootstrap is the Wald statistic computed as in equation (4), replacing $(\hat{\theta}_{\text{IV}} - \theta_0)$ with $(\hat{\theta}_{\text{IV}}^* - \hat{\theta}_{\text{IV}})$, and $\widehat{\text{Var}}(\hat{\theta}_{\text{IV}})$ by the bootstrap variance estimate of $\hat{\theta}_{\text{IV}}^*$. When the bootstrap sample is based on $\left\{\acute{\mathbf{e}}_g^*(\theta_0), \tilde{\mathbf{v}}_g^*(\theta_0)\right\}_{g=1}^G$, the bootstrap residuals is the same as in the ME-eff bootstrap. In this case the Wald bootstrap statistic is compute with $(\acute{\theta}_{\text{IV}}^* - \theta_0)$ in place of $(\hat{\theta}_{\text{IV}} - \theta_0)$, where

$$\acute{\theta}_{\text{IV}}^* = \left(\tilde{\mathbf{y}}_2^*(\theta_0)' \, \text{P}_{\text{M}_\mathbf{X}\mathbf{Z}} \tilde{\mathbf{y}}_2^*(\theta_0)\right)^{-1} \tilde{\mathbf{y}}_2^*(\theta_0)' \, \text{P}_{\text{M}_\mathbf{X}\mathbf{Z}} \acute{\mathbf{y}}_1^*(\theta_0),$$

with $\acute{\mathbf{y}}_1^*(\theta_0) = \tilde{\mathbf{y}}_{2,g}^*(\theta_0) \times \theta_0 + \mathbf{x}\tilde{\gamma}(\theta_0) + \acute{\mathbf{e}}^*(\theta_0)$. For both Wald bootstrap tests, the variance of the test is computed using the bootstrap sample and estimated parameters in equation (5).

The third Wald bootstrap method is the classical *pairs bootstrap.* This method is a completely nonparametric one. In the pairs bootstrap, the bootstrap sample is generated as:

$$\Pr\left(\left\{\mathbf{y}_{1,g}^*, \mathbf{y}_{2,g}^*, \mathbf{w}_g^*\right\} = \{\mathbf{y}_{1,j}, \mathbf{y}_{2,j}, \mathbf{w}_j\}\right) = \frac{1}{G}, \quad j = 1, \ldots, G.$$

Since the pairs bootstrap dgp does not impose the null under hypothesis for generating the bootstrap samples, the computation of the Wald bootstrap test is centered at the TSLS. The test is computed by substituting the bootstrap estimates of θ and residuals into the equations (4) and (5).

Bootstrap first-stage F- and efficient F-tests

If there is one endogenous variable, the first-stage F-statistic for the null hypothesis H_0 : $\Pi_z = 0$ is defined as:

$$\hat{\mathrm{F}} \equiv \frac{\hat{\Pi}_z' \left(\widehat{\mathrm{Var}}(\hat{\Pi}_z)\right)^{-1} \hat{\Pi}_z}{k_z}, \tag{15}$$

where $\widehat{\mathrm{Var}}(\hat{\Pi}_z)$ is the cluster-robust variance matrix. A more recent test for testing instrument weakness suitable for clusters residuals is the *effective F-statistic* proposed by Olea and Pflueger (2013), defined as

$$\hat{\mathrm{F}}_{eff} \equiv \frac{\mathbf{y}_2' \mathbf{Z}^{\perp} \mathbf{Z}^{\perp\prime} \mathbf{y}_2}{n \operatorname{tr}\left(\hat{\Xi}_{\mathbf{Z}^{\perp}\mathbf{V}}\right)},$$

with *effective degrees of freedom* equal to:

$$\hat{K}_{eff} \equiv \frac{\left[\operatorname{tr}\left(\hat{\Xi}_{\mathbf{Z}^{\perp}\mathbf{V}}\right)\right]^2 (1+2f)}{\operatorname{tr}\left(\hat{\Xi}'_{\mathbf{Z}^{\perp}\mathbf{V}} \hat{\Xi}_{\mathbf{Z}^{\perp}\mathbf{V}}\right) + 2f \operatorname{tr}\left(\hat{\Xi}_{\mathbf{Z}^{\perp}\mathbf{V}}\right) \max \operatorname{eval}\left(\hat{\Xi}_{\mathbf{Z}^{\perp}\mathbf{V}}\right)},$$

where $\mathbf{Z}^{\perp} = (\mathbf{Z}\mathrm{M}_{\mathbf{X}}\mathbf{Z})^{-\frac{1}{2}}\mathrm{M}_{\mathbf{X}}\mathbf{Z}$, $\hat{\Xi}_{\mathbf{Z}^{\perp}\mathbf{V}}$ is the estimator of the variance of $n^{-\frac{1}{2}}\mathbf{Z}^{\perp\prime}\mathbf{V}$, and $\operatorname{tr}(\cdot)$ and $\max \operatorname{eval}(\cdot)$ are the trace and the maximum eigenvalue operators, respectively. The parameter f in the formula of $\hat{K}_{eff}$ is a function of the maximum Nagar bias relative to the benchmark, and require a numerical routine. We use the simplified and conservative version of the test which sets $f = 10$. In this case, the critical values for $\hat{\mathrm{F}}_{eff}$ are on Table 1, page 360, of Olea and Pflueger (2013).

The computation of the F- and (conservative) effective F-tests are computed based only on the first stage regression residuals (SE-1st). Under the assumption that the null hypothesis is true, $\Pi_z = 0$. Therefore, we generate bootstrap samples as

$$\left\{\acute{\mathbf{y}}^*_{2,g}\right\}_{g=1}^{G} = \left\{\mathbf{w}_g \acute{\Pi}_w + \acute{\mathbf{v}}^*_g\right\}_{g=1}^{G},$$

where $\left\{\acute{\mathbf{v}}^*_g\right\}_{g=1}^{G} = \{\omega_g \acute{\mathbf{v}}_g\}_{g=1}^{G}$, $\acute{\mathbf{v}}_g = \mathbf{y}_{2,g} - \mathbf{w}_g \acute{\Pi}_w$, and $\acute{\Pi}_w = \left(\acute{\Pi}_z, \acute{\Pi}_x\right) = \left(0, (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_2\right)$. The bootstrap tests are computed exactly as the original tests, and their respective p-values are obtained as previously described. We also compute the *pairs bootstrap* version of the first-stage F-test, which has the bootstrap statistic centered at $\hat{\Pi}_z$.

Notes on the bootstrap algorithm

Residual weights. Apart from the pairs and EE bootstraps with multinomial (M) weights, the remaining weights used for the proposed bootstraps satisfy $E[\omega_g] = 0$, and $E[\omega_g^2] = 1$. This ensures that the distribution of the resampled scores or residuals have the same the first and second moments of their underlying empirical distributions. Matching higher moments of the bootstrap and empirical distributions yields the asymptotic refinement. Many residual weights satisfy this property for the wild bootstrap. Liu (1988) proposes weights defined as $\omega_g = \zeta_g - E(\zeta_g)$, where ζ_g is a gamma random variable with shape parameter 4 and scale parameter $\frac{1}{2}$. The gamma (Γ) weights also satisfy $E[\omega_g^3] = 1$, and therefore match the first three moments. Davidson and MacKinnon (2010) suggest sampling the weights from the Rademacher distribution, which is defined as

$$\omega_g = \begin{cases} 1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2 \end{cases}.$$

The Rademacher (R) weights match the first four moments if the underlying distribution is symmetric.[9]

A summary of the bootstrap methods used in the paper is presented in Table 1.

[Insert Table 1 here]

4 Monte Carlo simulations

We now evaluate the performance of the proposed cluster bootstraps using Monte Carlo simulations, and we experiment with a variety of dgps. The baseline model has a structure that resembles the panel data random-effects regression. We repeat the system (1)

[9] Liu also proposes a continuous weight based on the normal distribution, defined as $\omega_g = w_g z_g - E(w_g)E(z_g)$, where w_g and z_g are independent normal random variables with mean $\frac{1}{2}(\sqrt{17/6} + \sqrt{1/6})$ and variance $\frac{1}{2}$. These will maintain the first three moments of the empirical distribution of residuals. Mammen (1993) weights are alternative discrete weights defined as:

$$\omega_g = \begin{cases} (1-\sqrt{5})/2 & \text{with probability } \frac{1+\sqrt{5}}{2\sqrt{5}}, \\ 1-(1-\sqrt{5})/2 & \text{with probability } 1 - \frac{1+\sqrt{5}}{2\sqrt{5}}. \end{cases}$$

for convenience:

$$\begin{cases} \mathbf{y}_{1,g} = & \mathbf{y}_{2,g}\theta + \mathbf{x}_g\gamma + \mathbf{u}_g \\ \mathbf{y}_{2,g} = & \mathbf{z}_g\Pi_z + \mathbf{x}_g\Pi_x + \mathbf{v}_g \end{cases} \quad \text{for } g = 1, \ldots, G.$$

We assume that $\theta = 0$ is scalar, and set $\Pi_z' = (c_z, 0, \ldots, 0)'$, so that only the first instrument is relevant. The included instruments are $\mathbf{x}_g = [\boldsymbol{\iota}_g \; \boldsymbol{\epsilon}_{n_g}]$, where $\boldsymbol{\iota}_g$ is a $n_g \times 1$ vector of ones, $\text{vec}(\boldsymbol{\epsilon}_g) \sim N(0, \text{I}_{n_{g\times(k_x-1)}})$, and $\gamma = \Pi_x = (1, \ldots, 1)'$. The excluded instrument $\mathbf{z}_g$ is set as $\mathbf{z}_g = \iota_{n_g}\mathbf{d}_g' + \boldsymbol{\vartheta}_g$, where $\mathbf{d}_g$ is a $k_z \times 1$ vector and $\boldsymbol{\vartheta}_g$ is a $n_g \times k_z$ matrix. Both $\mathbf{d}_g$ and $\boldsymbol{\vartheta}_g$ are sampled from independent multivariate standard normal distributions and adjusted such that $\sum_{g=1}^{G} n_g \left(\mathbf{d}_g - \overline{\mathbf{d}}\right)' \left(\mathbf{d}_g - \overline{\mathbf{d}}\right) = (1-\lambda)\, n\text{I}_{k_z}$, where $\overline{\mathbf{d}} = \frac{1}{G}\sum_{g=1}^{G}\mathbf{d}_g$, and $\left(\sum_{g=1}^{G}\boldsymbol{\vartheta}_{n_g}'\boldsymbol{\vartheta}_{n_g}\right) = \lambda\; n\text{I}_{k_z}$, with $\frac{1}{n_g}\boldsymbol{\iota}_g'\boldsymbol{\vartheta}_g = 0$, $\frac{1}{n}\sum_{g=1}^{G}\boldsymbol{\iota}_g'\boldsymbol{\vartheta}_g = 0$, and $0 \leq \lambda \leq 1$. These adjustments allow us to have $n^{-1}\mathbf{Z}'\text{M}_{\mathbf{X}}\mathbf{Z} = \text{I}_{k_z}$, where n is the total number of observations. If $\lambda = 0$, the excluded instruments are the same within groups. We keep the instruments $\mathbf{W} = [\mathbf{Z} : \mathbf{X}]$ fixed in all simulations.

The cluster-robust first-stage F-statistic for testing $H_0 : \Pi_z = 0$ defined in equation (15) is asymptotically distributed as

$$\underbrace{n\frac{\Pi_z'\left[\text{Var}_\infty\left(\hat{\Pi}_z\right)\right]^{-1}\Pi_z}{k_z}}_{\mu_{k_z}} + F(k_z, +\infty),$$

where $\text{Var}_\infty(\hat{\Pi}_z) = \lim_{n\to+\infty}\frac{1}{n}\text{E}[\mathbf{Z}'\text{M}_{\mathbf{X}}\mathbf{V}\mathbf{V}'\text{M}_{\mathbf{X}}\mathbf{Z}]$ and $F(k_z, +\infty)$ represents the asymptotic F-distribution. If the residulas are non-spherical, then the parameter μ_{k_z} can be interpreted as the "concentration parameter" divided by the number of exogenous instruments. Since $\Pi_z = (\pi_z, 0, ..., 0)$, the noncentrality parameter becomes:

$$\mu_{k_z} = n\frac{c_z^2}{k_z}\left[\text{Var}_\infty\left(\hat{\Pi}_z\right)\right]_{11}^{-1}, \tag{16}$$

where $\left[\text{Var}_\infty(\hat{\Pi}_z)\right]_{11}^{-1}$ indicates the first diagonal entry of $\left[\text{Var}_\infty(\hat{\Pi}_z)\right]^{-1}$. We fix the value for c_z as:

$$c_z = \sqrt{\frac{k_z}{\left[\text{Var}_\infty\left(\hat{\Pi}_z\right)\right]_{11}^{-1}\mu_{k_z}}}.$$

We replace $\mathrm{Var}_\infty(\hat{\Pi}_z)$ in Equation 16 by $\frac{1}{n}\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathrm{E}[\mathbf{V}\mathbf{V}'|\,\mathbf{W}]\,\mathrm{M}_{\mathbf{X}}\mathbf{Z}$, where $\mathrm{E}[\mathbf{V}\mathbf{V}'\,\mathbf{W}|]$ is explained below. We show in Appendix D that, under our dgp, μ_{k_z} is closely related to μ^2 parameter of Olea and Pflueger (2013) which captures the bias of the TSLS.[10]

We generate errors of the model as:

$$\begin{cases} \mathbf{u}_g &= \left(\mathbf{u}_g^c + \mathbf{u}_g^i\right)|z_{11,g}|^\kappa \\ \mathbf{v}_g &= \left(\rho\mathbf{u}_g^c + \varrho\mathbf{u}_g^i\right)|z_{11,g}|^\kappa + \left(1-\rho^2\right)^{\frac{1}{2}}\mathbf{v}_g^c + \left(1-\varrho^2\right)^{\frac{1}{2}}\mathbf{v}_g^i \end{cases} \quad \text{for} \quad g=1,\ldots,G,$$

where $\left(\mathbf{u}_g^c, \mathbf{v}_g^c\right) = \sqrt{\phi}\,(\varepsilon_{1,g}, \varepsilon_{2,g}) \otimes \boldsymbol{\iota}_g$, $\left(\mathbf{u}_g^i, \mathbf{v}_g^i\right) = \sqrt{1-\phi}(\boldsymbol{\psi}_{1,g}, \boldsymbol{\psi}_{2,g})$, $\boldsymbol{\varepsilon}_g = (\varepsilon_{1,g}, \varepsilon_{2,g}) \sim N\left(0, \mathrm{I}_2\right)$, $\boldsymbol{\psi}_g = (\boldsymbol{\psi}_{1,g}', \boldsymbol{\psi}_{2,g}')' \sim N(0, \mathrm{I}_{2n_g})$, and $\boldsymbol{\varepsilon}_g$ and $\boldsymbol{\psi}_g$ are independent. The parameters ρ and ϱ are scalars that capture the intra-cluster and the idiosyncratic correlations, respectively, and $z_{11,g}$ is a scalar that corresponds to the first element of $\mathbf{z}_g$ matrix. The parameter κ depicts the degree of heterogeneity in the model. When $\kappa = 0$, the residuals are akin to the random-effects panel-data model. The scalars ϕ and φ are weights for the cluster and idiosyncratic components of the variance, which satisfy $0 \leq \phi, \varphi \leq 1$ and $\phi + \varphi = 1$. Therefore, the joint distribution of $\left(\mathbf{u}_g', \mathbf{v}_g{}'\right)'$ is approximately:

$$\begin{pmatrix} \mathbf{u}_g \\ \mathbf{v}_g \end{pmatrix} \sim N\left(0, \begin{bmatrix} \mathsf{W}_g + \overline{\mathsf{W}}_g & \rho\mathsf{W}_g + \varrho\overline{\mathsf{W}}_g \\ & \mathsf{W}_g + \overline{\mathsf{W}}_g \end{bmatrix}\right).$$

where $\mathsf{W}_g = \phi\ \boldsymbol{\iota}_g\boldsymbol{\iota}_g'\,(z_{11,g})^{2\kappa}$, and $\overline{\mathsf{W}}_g = \varphi\mathrm{I}_{n_g}\,(z_{11,g})^{2\kappa}$. Therefore, we find $\mathrm{E}[\mathbf{V}\mathbf{V}'|\,\mathbf{W}] = \mathsf{W}+\overline{\mathsf{W}}$, where $\mathsf{W} = \mathrm{diag}(\{\mathsf{W}_g\}_{g=1}^G)$ and $\overline{\mathsf{W}} = \mathrm{diag}(\{\overline{\mathsf{W}}_g\}_{g=1}^G)$. Due to the normalization, $\mathrm{E}[(z_{11,g})^2] \approx 1$ by construction. Moreover, assuming that the number of observations per cluster is the same ($n_g = \bar{n}$ for all g, which implies that $n = \bar{n} \times G$), the asymptotic variance $\mathrm{Var}_\infty\left(\hat{\Pi}_z\right)$ simplifies to

$$\mathrm{Var}_\infty\left(\hat{\Pi}_z\right) = \mathrm{I}_{k_z}\left(\phi\bar{n}\,(1-\lambda) + \varphi\right),$$

and the Nagar (1959) approximation for computing the bias of the TSLS is:

$$\mathrm{E}\left[\hat{\theta}_{\mathrm{IV}} - \theta \middle| \mathbf{W}\right] \approx (k_z - 2)\left(nc_z^2\right)^{-1}\left(\phi\bar{n}\,(1-\lambda)\,\rho + \varphi\varrho\right).^{11} \tag{17}$$

[10] The bias of TLS is derived in Appendix C.

[11] See derivation in Appendix C.

If $\phi = \varphi$, it is clear from equation (17) that the within-cluster correlation ρ counts for $\bar{n}$ times more than the correlation of the idiosyncratic term ϱ for the bias of the two-stage least squares (TSLS) estimator. By setting $\rho = \varrho$, and using the concentration parameter divided by the number of exogenous instruments as defined in equation (16), the approximate bias of the IV estimator can be rewritten as

$$\mathrm{E}\left[\hat{\theta}_{\mathrm{IV}} - \theta \middle| \mathbf{W}\right] \approx \frac{(k_z - 2)}{k_z} \left(\mu_{k_z}\right)^{-1} \rho.$$

This is similar to the bias for $\hat{\theta}_{\mathrm{IV}}$ derived in Bun and de Haan (2010).

Simulation results

Our results are based on 10,000 Monte Carlo experiments. In the experiments, we set $(k_x, k_z) = (2, 5)$, $\lambda = 0.1$, $\phi = \varphi = 0.5$, and $\rho = \varrho$. We investigate the cases where $\mu_{k_z} = 0.1, 1,$ or 9, indicating very weak, weak, or strong instruments, respectively. For each μ_{k_z}, we report also report $\mu^h_{k_z} = \frac{\Pi_z'(\mathbf{Z}\mathrm{M}_{\mathbf{X}}\mathbf{Z})\Pi_z}{k_z}$, which is the standard concentration parameter divided by k_z. We choose κ from $\{0, 1, 2\}$, meaning no heterogeneity, heterogeneity and strong heterogeneity, respectively. We set the endogeneity degree as $\rho = 0.20, 0.95$. Finally, we consider cases where the number of observations per cluster are 20, or the number of observations differs across cluster but average approximately 20 per cluster. For the latter case of nonidentical cluster size, we first set the total number of clusters G, and then choose the number of observations for each cluster from $\{16, 17, \ldots, 25\}$. The number of clusters with different sample sizes is equal. For example, when $G = 40$, we have four clusters with 16, 17, and up to 25 observations each. To save space, we only report results for the case where the number of observations differs across clusters. For the same reason, we do not report results of the SE-neff bootstrap Λ_{KLM}-test either because the results are very close to the ones obtained from ME-neff bootstrap.[12]

In our experiments, we use 199 and 499 bootstrap replications for size and power results, respectively. In repeated Monte Carlo experiments, the sampling error from a small number of bootstrap replications should cancel out. In practice, at least 999 replications should be used.

[12]The full set of results are available upon request with the authors.

Size results

Tables 2A and 2B contain the rejection rates for the asymptotic tests with their respective bootstrap counterparts for $\mu_{k_z} = 1$ and 9, with $G = 20$ clusters and different heterogeneity levels (κ), instruments strengthes (μ_{k_z}) and endogeneity degrees (ρ). The significance level for the tests is 5%.

[Insert Table 2A here]

[Insert Table 2B here]

The rejections rates of the asymptotic Wald test differ considerably from the nominal level even with strong instruments ($\mu_{k_z} = 9$), and this difference is increasing in the degree of endogeneity (ρ) and residual heterogeneity (κ). The bootstrap Wald tests also have rejection rates far from the nominal size even when the instruments are strong. Nevertheless, the performance of the Wald ME-eff bootstrap, which imposes the null $H_0 : \theta = \theta_0$ when generating the bootstrap samples, is superior to the Wald ME-IV bootstrap, which does not impose the null. All the asymptotic weak instrument tests are oversized when residuals heteroskedasticity is at $\kappa = 0$, or 1, and the Λ_{AR}-test is undersized under strong heteroskedasticy ($\kappa = 2$); however, their rejections rates are closer to the nominal level than are those of the Wald tests. Except for the EE bootstrap with multinomial weights, which are severely undersized in all scenarios, the remaining proposed bootstraps present rejection rates closer to the nominal level. The SE and ME bootstraps outperforms the EE bootstraps, specially when the degree of heterogeneity is high, which is similar to the Kline and Santos (2012) results. The SE-neff bootstrap for the Λ_{AR} and Λ_{CLR}-tests and the ME-neff bootstrap for the Λ_{KLM}-test with Rademacher (R) weights are the bootstrap tests with rejection rates closer to the nominal level in almost all cases. In particular, they outperform the remaining bootstrap procedures when heteroskedascity is very strong ($\kappa = 2$). The DM bootstrap tests performs well under homoskedastic residuals ($\kappa = 0$). In the presence of heteroskedastic residuals, however, their performance worsens with the distortion increasing with heteregeneity degree. This results is not surprising since the number of within cluster observations is high vis a vis the number of clusters.

The asymptotic cluster robust first-stage F-test overrejects the weakness of the instruments, and this overrjection can be high-above 84% when the concentration parameter is small. On the other hand, the effective F-test underrejects the same hypothesis, which is an expected result since we are using a conservative version of the test. The bootstrapped F and effective F-tests present, respectively, lower and higher rejection rates compared to their asymptotic counterparts. Interestingly, the SE-1st F- and effective F-tests give similar rejection rates under Rademacher weights.

In Tables 3A and 3B, we study the performance of the tests when the number of clusters increases, but the number of observations within the clusters remains constant. The results are based on the dgp using heteroskedastic errors ($\kappa = 1$) with endogenity degree at $\rho = 0.95$.

[Insert Table 3A here]

[Insert Table 3B here]

Even even when the instruments are strong ($\mu_{k_z} = 9$), the asymptotic and bootstrap Wald test remains oversized in all experiments, but with rejection rates approaching the nominal size as the number of clusters G increases. The asymptotic weak instrument tests rejection probabilities also approach the nominal level as the sample size increases. The differences between the weak instrument tests rejection probabilities and nominal level are smaller compared to difference obtained from the Wald test. When the instruments are weak and strong, the weak instruments bootstrap tests also converge to the nominal level, although their convergence is not monotonic; however, when the instruments are very weak ($\mu_{k_z} = 0.1$) only the bootstrap Λ_{AR} test has rejection probabilities close to nominal size. As expected, the DM bootstrap AR-test rejection rates approach the nominal size as the number of clusters increases, but the convergence is slower when residuals are highly heteroskedastic ($\kappa = 2$). The EE bootstrap tests with multinomial weights remain severely undersized and slowly converging to the nominal size. Finally, the asymptotic cluster robust first-stage F- and effective F-tests have different rejection probabilities for testing $H_0 : \Pi_z = 0$. Remarkably, the bootstrap SE-1st for F- and effective F-tests give similar rejection rates.

Power comparison

Next, we compare power across the alternative bootstrap tests, using the same dgp as in Tables 2A and 2B. The bootstrap results based on gamma weights are similar to those obtained with Rademacher weights, therefore we only report the later. The power curves of the bootstrap Λ_{AR}- and Λ_{CLR}-tests are very similar for SE-in and SE-neff bootstraps . Therefore, only the SE-in results are reported. The same for ME-eff and ME-neff bootstrap Λ_{KLM}-tests, and, thence, only the power curves of ME-neff are reported.

[Insert Figure 1 here]

[Insert Figure 2 here]

Figures 1 and 2 reveal the great size distortion of the asymptotic tests and of the Wald bootstrap tests, although the distortion is lower when the concentration parameter is higher. The figures also reveal that the asymptotic Λ_{AR}, Λ_{KLM}, Λ_{CLR} tests are not consistent when instruments are weak ($\mu_{k_z} = 1$), as expected. Surprisingly, even the Λ_{AR}-test with strong instruments ($\mu_{k_z} = 9$) are also not consistent with small number of clusters. For the Λ_{AR}, and Λ_{CLR} cases, the SE-in bootstraps power dominate the EE bootstraps, and the difference is more pronounce when endogeneity is high ($\rho = 0.95$). In the Λ_{KLM} case, the SE-in bootstrap test also power dominates the other bootstraps methods. There is no clear evidence if the EE method power dominates the ME-in and ME-neff bootstraps, but the ME bootstrap Λ_{KLM}-tests, however, present better size performance than the EE bootstrap Λ_{KLM}-test.

The simulation results suggest that SE bootstrap methods have better power perfomance than the other bootstrap methods, and their performance in terms of size is at least as good as the other bootstraps methods.

5 Empirical Applications

We now use our bootstrap methods in two empirical applications that fit system (1). We construct confidence regions for the structural parameter of interest by inverting the bootstrap tests. The $1 - \alpha$ confidence set is formed by the points in the parameter space that

do not reject the null hypothesis at significance level α. We use 1999 bootstrap samples to compute the bootstrap p-values. In both empirical applications, the unboundedness of the confidence sets prevents us from using less intensive computational methods than the ones suggested in Davidson and MacKinnon (2010).

"Economic Shocks and Civil Conflict: An Instrumental Variables Approach"

Miguel et al. (2004) investigate the relationship between economic conditions and civil war in sub-Saharan Africa. The authors are interested in how the deterioration of the economic environment affects the probability of a civil conflict. The endogeneity problem arises from several channels. One channel is the government institution quality, which is not observed by the econometricians and drives both economic growth and the probability of civil wars. As an instrument for income growth they use variation in rainfalls. This choice is motivated by the fact that those economies rely on subsistence agriculture. Their data consists of an unbalanced panel of 41 African countries from 1981 to 1999, with 743 total observations, averaging 18.6 observations per country. The structural equation, which captures the impact of economic fluctuations on civil conflict, is:

$$\mathrm{C}_{i,t} = \Delta\mathrm{GDP}_{i,t}\theta_1 + \Delta\mathrm{GDP}_{i,t-1}\theta_2 + X_{i,t}\gamma + u_{i,t}, \tag{18}$$

where $\mathrm{C}_{i,t}$ is an indicator variable equal to one if there is a civil conflict with at least of 25 battle deaths per year and zero otherwise, $\Delta\mathrm{GDP}_{i,t}$ is the annual growth rate, $X_{i,t}$ is a set of control variables including country effects and country-specific time trends. They consider a linear probability model and estimate (18) by TSLS with current and lagged rainfall growth as instruments. Using country as the cluster unit allows them to treat the errors within each country as serially correlated; however, they treat each country as independent units.

In Table 4, we reproduce the same estimates of the Table 4 columns (5) and (6), page 739 obtained by Miguel et al. (2004). We found that the negative shocks on the economy raise the probability of civil conflict, as expected by the theory.

[Insert Table 4 here]

We report the cluster-robust F-tests, which show that the exogenous instruments both $\Delta\text{GDP}_{i,t}$ and $\Delta\text{GDP}_{i,t-1}$are statistically different from 0 at the 5% significance level. Nevertheless, the F-tests are below the Staiger and Stock (1997) rule-of-thumb of 10, and below the critical values in Table 1 in Olea and Pflueger (2013), suggesting that the instruments may be weak. The rank statistic of Kleibergen and Paap (2006), however, points strongly to the joint significance of the instruments.

We examine the above results by comparing the Wald test to the Λ_{AR}-test confidence regions, since the model is just identified. Figures 3A and 3B show the confidence regions for $\theta = (\theta_1, \theta_2)$ derived from the original models (5) and (6) in Miguel et al. (2004) page 739, respectively. The asymptotic confidence regions are constructed from the inversion of the asymptotic Wald and Λ_{AR}-tests, while the bootstrap confidence regions are obtained from the Wald ME-eff and Λ_{AR} SE-neff bootstrap tests with Rademacher weights.[13]

[Insert Figure 3A here]

[Insert Figure 3B here]

The asymptotic Wald confidence regions are elliptic and indicate a positive correlation among the estimated structural parameters. The Wald bootstrap confidence regions are very different from their asymptotic ones. Zhan (2010) establishes that this difference between the asymptotic and bootstrap confidence regions is not surprising if instruments are weak. Both the asymptotic Λ_{AR} and Λ_{AR} SE-neff bootstrap confidence regions are non-convex and unbounded, very different from the bounded asymptotic Wald confidence set. Therefore, even though the rank-test suggests strong correlation between instruments and endogenous variables, the Λ_{AR} and Λ_{AR} SE-neff bootstrap confidence regions points to weak instrument presence.

Interestingly, in Figure 3A the effect of economic fluctuations on civil war conflict is statistically insignificant at 10% level according the Wald asymptotic test. At the same significant level, however, this hypothesis is rejected by both Λ_{AR} asymptotic and SE-neff bootstrap tests. The same is true in Figure 3B at 5% significant level.

[13]The results are similar with gamma weights.

"The Colonial Origins of Comparative Development: An Empirical Investigation"

The seminal article of Acemoglu et al. (2001) shows how institutions, as measured by the protection of risk expropriation, affect economic performance. Capturing this effect is difficult because economic growth also shapes institutions. Moreover, there are potential omitted variables that influence both institutions and economic performance. The authors argue that the mortality rates faced by Europeans affect their willingness to establish settlements and choice of colonization strategy. Places where mortality rates are high are likely to have "extractive" institutions, whereas healthy places are prone to receive better economic and political institutions. Therefore, the mortality rate would be a good instrument for the institution variable. Their proposed regression model is the following two stage model:

$$\begin{cases} y_{i,g} = & r_{i,g}\theta + x_{i,g}\gamma + u_{i,g} \\ r_{i,g} = & m_{i,g}\pi_m + x_{i,g}\pi_x + u_{i,g} \end{cases}$$

where y is the log of GDP per capita in 1995, r is the protection of risk appropriation, and m is log settler mortality. Acemoglu et al. (2001) consider several specifications in which x could include latitude, continent dummies, percentage of European descent in 1975, and malaria, measure by the 1994 Falciparum malaria index. The index i refers to colonial country, and the index g refers to countries which share the same mortality rates. For example, due to the difficulty of obtaining historical data, several Latin American countries are assigned the same mortality rates. In Africa, the mortality rate of a country are inferred from the mortality rates of a neighboring country. Therefore, as pointed by Albouy (2012), the errors of the regression specification should be clustered and not treated an independent as originally done in Acemoglu et al. (2001).

Albouy (2012) also raises other criticism of Acemoglu et al. (2001) concerning measurement of the mortality rate. In particular, he argues that mortality rates during peacetime and "campaign" episodes are not the same. He also argues that some data for West and Central African countries are unreliable. By adding a campaign dummy and discarding contested observations, he finds that, because of weak instruments, the effect

of institutions on growth becomes less clear. Acemoglu et al. (2012) rebuff each of Albouy (2012) criticism in the same issue of the *American Economic Review*.[14] Since the econometric investigation of Acemoglu et al. (2012) and Albouy (2012) are based on samples with at most 62 observations and 35 clusters, we revisited their findings by comparing their asymptotic methods with the bootstrapped methods we propose.

Tables 5A and 5B contain the confidence intervals for some of the specifications in Acemoglu et al. (2012) together with bootstrapped *p*-values of the first-stage F- and effective F-tests. The AR-ARJ refers to AR the confidence intervals obtained by Acemoglu et al. (2012).[15] The remaining confidence intervals are the methods discussed in previous sections.

[Insert Table 5A here]

[Insert Table 5B here]

In the line with the previous empirical example, the Wald ME-eff bootstrap confidence intervals are larger than the Wald asymptotic confidence intervals in several cases, as depicted on Column (3), which indicate weak instruments. The AR-ARJ confidence intervals have in general a smaller range when compared with the Λ_{AR} confidence intervals. In the majority of the cases, the Λ_{AR} bootstrap confidence intervals have a larger range compared to the asymptotic ones. Nevertheless, the Λ_{AR} bootstraps confidence intervals still suggest positive the effect of institutions on growth in the majory of cases. Only in very few cases, for example, as the model that includes continent dummies and latitude as explanatory variables, that the SE-neff bootstrap confidence intervals indicates that

[14] The confidence intervals for θ based on the clustered *AR* tests in Albouy (2012) and Acemoglu et al. (2012) are different, although they use the same regression specification, data and software package. The reason is that Albouy (2012) wrote his own code taking the advantage of the built-in functions of *Stata* while Acemoglu et al. (2012) use the Stata package `rivtest` developed by Finlay and Magnusson (2009) which compute the minimum distance version of the *AR*-test. The algorithms have different estimators of the covariance matrix $\widehat{\Omega}(\theta_0)$, which, due to small sample size, results in very different confidence sets. That is the reason why Acemoglu et al. (2012) could not match Albouy (2012) Albouy's (2012) results as mentioned on footnote 28 of their paper.

[15] We use the `weakiv` package developed by Finlay et al. (2013) to compute the AR-AJR confidence intervals. The `weakiv` package, which incorporated the `rivtest`, has the cluster correction variance term $\frac{G}{(G-1)}\frac{(n-1)}{(n-k_w)}$ which is in the default of the Stata cluster routine to compute the variance matrix. This correction makes the intervals approximately 0.01 larger than the ones reported in Acemoglu et al. (2012). The correction term is also used when computing the AR and AR bootstrapped tests.

the institutional effect on growth is statistically insignificant. We also note that the length of the bootstrap Λ_{AR} SE confidence intervals of are smaller than the Λ_{AR} EE bootstrap ones in most cases, as suggested by the power curve simulations. For the case of the F and efficient F statistics, their bootstrap *p*-values are very close to each other with the asymptotic F and effective F *p*-values are higher and lower than their respective bootstrap counterparts, as also indicated by the Monte Carlo size simulations.

Albouy (2012) discards countries with conjectured mortality rate, which reduces the sample to only 28 countries. Acemoglu et al. (2012) argues that the results using this small sample are mainly driven by Gambia, which is an outlier. In Tables 6A and 6B repeat a similar exercise as the previous tables using Albouy (2012) preferred sample of 28 countries, and with the same sample without Gambia. By reducing the sample, the cluster dimension is lost; however, our proposed methods remains valid in the individual heteroskedastic case. We also include the DM bootstrap confidence intervals which are valid in the presence of heteroskedastic errors.

[Insert Table 6A here]

[Insert Table 6B here]

We find that using Albouy's preferred sample, the Λ_{AR} asymptotic and bootstrap confidence intervals, but not the AR-AJR, cover the entire real line in all specifications except the one with no covariates. By excluding Gambia, the confidence intervals become smaller; however, in some specifications, we cannot ruled out that the institutional effect is not statistically significant. For the remaining specifications, the bootstrap confidence intervals are generally larger then the asymptotic ones, specially the DM bootstrap, suggesting that a potential smaller effect of institutions on economic performance in comparison to Acemoglu et al. (2012).

6 Conclusion

We propose bootstrap methods for Wald, weak-instrument-robust, F- and effective F-tests in the linear IV framework with clustered residuals. Our simulations show that

asymptotic tests are size distorted even when instruments are strong and the sample size is relatively large. The same simulations shows that even with small sample size, residual bootstraps of weak-instrument-robust tests present rejection probabilities close to nominal size. From all the proposed bootstrap methods for the weak-instrument-robust tests, the single-equation residual bootstrap power dominates the other methods. We also find that the single-equation residual bootstrap for F- and effective F-tests are very close to each other, and their asymptotic tests over- and underreject the null hypothesis, respectively.

We use the proposed methods in two empirical applications: the impact of economy on civil conflits of Miguel et al. (2004), and the study of how institutions affect economic growth discussed by Albouy (2012) and Acemoglu et al. (2012). In the first application, the Wald asymptotic and bootstrap confidence regions are very different. The AR asymptotic and bootstrap confidence regions are larger than their Wald counterpart, and they indicate no evidence that a bad economic performance affects the probability starting a civil conflict in sub-saharan countries. In the second application, although the AR single-equation residual bootstrap confidence regions are larger than the AR asymptotic confidence region, the majority of the replication results support the claims in Acemoglu et al. (2012).

A Cluster variance-covariance matrix

Define $\mathbf{h}_g(\theta_0) = \sum_{i=1}^{n_g} \mathbf{h}_{i,g}(\theta_0) = \mathbf{w}_g' \mathbf{e}_g(\theta_0)$ with $\mathbf{h}_{i,g} = \mathbf{w}_{i,g}' e_{i,g}(\theta_0)$, and $\mathbf{q}_g = \sum_{i=1}^{n_g} \mathbf{q}_{i,g} = (\mathrm{I}_p \otimes \mathbf{w}_g')\mathbf{v}_g$ with $\mathbf{q}_{i,g} = (\mathrm{I}_p \otimes \mathbf{w}_{i,g}')v_{i,g}$, where $\mathbf{w}_g = (\mathbf{z}_g : \mathbf{x}_g)$, $\mathbf{v}_g = \mathrm{vec}(\mathbf{y}_{2,g} - \mathbf{w}_g \Pi_w)$, $\mathbf{e}_g = \mathbf{Y}_g(\theta_0) - \mathbf{w}_g \delta_w$. Let us partition $\Xi(\theta_0)$, the variance matrix in equation (9) as $\Xi(\theta_0) = [\Xi_{\mathbf{hh}}(\theta_0), \Xi_{\mathbf{hq}}(\theta_0) : \Xi_{\mathbf{qh}}(\theta_0), \Xi_{\mathbf{qq}}(\theta_0)]$. Each component of $\Xi(\theta_0)$ is estimated as:

$$\Xi_{\mathbf{ds}}(\theta_0) = \frac{1}{n} \sum_{g=1}^{G} \left(\mathbf{d}_g(\theta_0) - n_g \overline{\mathbf{d}}(\theta_0)\right) \left(\mathbf{s}_g(\theta_0) - n_g \overline{\mathbf{s}}_g(\theta_0)\right)',$$

where $\mathbf{d}_g(\theta_0) = \mathbf{h}_g(\theta_0)$ or $\mathbf{q}_g(\theta_0)$ with $\overline{\mathbf{d}}(\theta_0) = \frac{1}{n}\sum_{g=1}^{G} \mathbf{d}_g(\theta_0)$, and $\mathbf{s}_g(\theta_0) = \mathbf{h}_g(\theta_0)$ or $\mathbf{q}_g(\theta_0)$ with $\overline{\mathbf{s}}(\theta_0) = \frac{1}{n}\sum_{g=1}^{G} \mathbf{s}_g(\theta_0)$. For computing the variance of the weak instrument tests in Section 2, we replace $(\mathbf{e}_g(\theta_0), \mathbf{v}_g)$ by $(\acute{\mathbf{e}}_g(\theta_0), \hat{\mathbf{v}}_g) = (\mathbf{Y}_g(\theta_0) - \mathbf{w}_g \acute{\delta}_w(\theta_0), \mathbf{y}_{2,g} - \mathbf{w}_g \widehat{\Pi}_w)$. In the case of the EE bootstrap, we use

$$\tilde{\mathbf{h}}_g^*(\theta_0) = \omega_g \tilde{\mathbf{h}}_g^c(\theta_0) = \omega_g \left(\mathbf{w}_g' \tilde{\mathbf{e}}_g(\theta_0) - \frac{n_g}{n} \sum_{g=1}^{G} \mathbf{w}_g' \tilde{\mathbf{e}}_g(\theta_0) \right)$$

in place of $\mathbf{h}_g(\theta_0)$. The remaining variance terms of the EE bootstrap are not computed because we are conditioning on $\widetilde{\Pi}_z(\theta_0)$. For the residual bootstrap cases the covariance matrix is computed by substituting $(\mathbf{e}_g(\theta_0), \mathbf{v}_g)$ by $(\mathbf{e}_{\mathrm{b},,g}^*(\theta_0), \mathbf{v}_{\mathrm{b},,g}^*)$, in $\mathbf{h}_g(\theta_0)$ and $\mathbf{q}_g$ equation defined above, where $\mathbf{e}_{\mathrm{b},g}^* = \mathbf{Y}_g^*(\theta_0) - \mathbf{w}_g \delta_w^*(\theta_0)$, $\mathbf{v}_{\mathrm{b},g}^* = \mathbf{y}_{2,g}^*(\theta_0) - \mathbf{w}_g \Pi_w^*(\theta_0)$, and $(\delta_w^*(\theta_0), \Pi_w^*(\theta_0))$ are the bootstrap estimates values.

B Derivation of the Lagrange multiplier estimator

The Kleibergen test is the Lagrange multiplier test derived from the following restricted minimization problem under $H_0 : \mathrm{d}(\theta_0) = 0$:

$$\begin{aligned} \min_{\pi_w, \gamma} \quad & \frac{1}{2} \begin{pmatrix} \hat{\delta}_w(\theta_0) - \Pi_w \mathbf{d}(\theta_0) - \mathrm{H}\gamma \\ \hat{\pi}_w - \pi_w \end{pmatrix}' \left[\widehat{\Omega}(\theta_0)\right]^{-1} \begin{pmatrix} \hat{\delta}_w(\theta_0) - \Pi_w \mathbf{d}(\theta_0) - \mathrm{H}\gamma \\ \hat{\pi}_w - \pi_w \end{pmatrix}, \quad \text{(A-1)} \\ \text{s.t. } & \mathrm{d}(\theta_0) = 0 \end{aligned}$$

where $\mathrm{H} = [\ 0' \quad \mathrm{I}_{k_x}'\ ]'$ and the estimator $\widehat{\Omega}(\theta_0)$ is defined as in equation (9). In the following, we omit (θ_0) from $\delta_w(\theta_0)$ and $\mathrm{d}(\theta_0)$ and $\widehat{\Omega}(\theta_0)$ to facilitate the exposition. From the FOC condition we obtain:

$$\mathrm{H}' \left(\widehat{\Omega}_{\delta_w \delta_w}\right)^{-1} \left(\hat{\delta}_w - \mathrm{H}\tilde{\gamma}(\theta_0)\right) = 0, \text{ and } \widetilde{\Pi}_w(\theta_0)' \left(\widehat{\Omega}_{\delta_w \delta_w}\right)^{-1} \left(\hat{\delta}_w - \mathrm{H}\tilde{\gamma}(\theta_0)\right) = \tilde{\lambda}(\theta_0), \quad \text{(A-2)}$$

from where we derive $\tilde{\gamma}(\theta_0) = [\mathrm{H}'(\widehat{\Omega}_{\delta_w \delta_w})^{-1}\mathrm{H}]^{-1}\mathrm{H}'(\widehat{\Omega}_{\delta_w \delta_w})^{-1}\hat{\delta}_w$, and $\tilde{\lambda}(\theta_0) = \widetilde{\Pi}_w(\theta_0)'$ $(\widehat{\Omega}_{\delta_w \delta_w})^{-1} \mathrm{M}_{\mathrm{H}}^{(\widehat{\Omega}_{\delta_w \delta_w})^{-1}} \hat{\delta}_w$, where $\mathrm{M}_{\mathrm{H}}^{(\widehat{\Omega}_{\delta_w \delta_w})^{-1}} = \mathrm{I} - \mathrm{P}_{\mathrm{H}}^{(\widehat{\Omega}_{\delta_w \delta_w})^{-1}}$ and $\mathrm{P}_{\mathrm{H}}^{(\widehat{\Omega}_{\delta_w \delta_w})^{-1}} = \mathrm{H}\ [\mathrm{H}'(\widehat{\Omega}_{\delta_w \delta_w})^{-1}$

$\mathrm{H}]^{-1}\,\mathrm{H}'\,(\widehat{\Omega}_{\delta_w\delta_w})^{-1}$ is an oblique projection.[16] Using the fact that $(\widehat{\Omega}_{\delta_w\delta_w})^{-1}\mathrm{M}_{\mathrm{H}}^{(\widehat{\Omega}_{\delta_w\delta_w})^{-1}} = [(\widehat{\Omega}_{\delta_z\delta_z})^{-1},\ 0:\ 0\,,0]$, further simplification allow us to write the Lagrange multiplier estimator as $\tilde{\lambda}\,(\theta_0) = \widetilde{\Pi}_z\,(\theta_0)'\ (\widehat{\Omega}_{\delta_z\delta_z})^{-1}\,\hat{\delta}_z$. An estimator of the variance of $\tilde{\lambda}\,(\theta_0)$ conditional on $\widetilde{\Pi}_z\,(\theta_0)$ is $\widehat{\mathrm{Var}}(\tilde{\lambda}\,(\theta_0)) = \widetilde{\Pi}_z\,(\theta_0)'\,(\widehat{\Omega}_{\delta_z\delta_z})^{-1}\,\widetilde{\Pi}_z\,(\theta_0)$. Finally, we have that the estimator for Π_w derived from equation (A-2):

$$\tilde{\pi}_w\,(\theta_0) = \mathrm{vec}\left(\widehat{\Pi}_w\right) - \widehat{\Omega}_{\pi_w\delta_z}\left(\widehat{\Omega}_{\delta_z\delta_z}\right)^{-1}\hat{\delta}_z.$$

The estimator of the variance of $\tilde{\pi}\,(\theta_0)$ is $\widehat{\mathrm{Var}}\,(\tilde{\pi}\,(\theta_0)) = \widehat{\Omega}_{\pi_w\pi_w} - \widehat{\Omega}_{\pi_w\delta_z}\left(\widehat{\Omega}_{\delta_z\delta_z}\right)^{-1}\widehat{\Omega}_{\delta_z\pi_w}$.

The Anderson and Rubin test is the Lagrange multiplier test derived from the following restricted minimization problem:

$$\min_{\delta_w,\text{s.t. }\delta_z=0} \quad \frac{1}{2}\left(\hat{\delta}_w-\delta_w\right)'\left[\widehat{\Omega}_{\delta_w\delta_w}\right]^{-1}\left(\hat{\delta}_w-\delta_w\right). \tag{A-3}$$

From the FOC conditions with respect to δ_x and the Lagrange multiplier λ we find $\tilde{\lambda}_{\mathrm{AR}}\,(\theta_0) = (\widehat{\Omega}_{\delta_z\delta_z})^{-1}\hat{\delta}_z$, and $\tilde{\delta}_x\,(\theta_0) = \hat{\delta}_x - \widehat{\Omega}_{\delta_x\delta_z}\tilde{\lambda}_{\mathrm{AR}}\,(\theta_0)$, which is the same as the minimum-distance estimator for γ in equation (A-2).

C Bias of the cluster IV estimator for θ

In matrix notation, the cluster residual model in system (1) is

$$\left\{\begin{array}{ll}\mathbf{y}_1 = & \mathbf{y}_2\theta+\mathbf{X}\gamma+\mathbf{u}\\ \mathbf{y}_2 = & \mathbf{Z}\Pi_z+\mathbf{X}\Pi_x+\mathbf{V}\end{array}\right., \qquad \mathrm{E}\left[\left.\begin{array}{cc}\mathbf{u}\mathbf{u}' & \mathbf{u}\underline{\mathbf{v}}'\\ \underline{\mathbf{v}}\mathbf{u}' & \underline{\mathbf{v}}\underline{\mathbf{v}}'\end{array}\right|\mathbf{W}\right] = \left[\begin{array}{cc}\Sigma_{\mathbf{uu}} & \Sigma_{\mathbf{uv}}\\ \Sigma_{\mathbf{vu}} & \Sigma_{\mathbf{vv}}\end{array}\right],$$

where, $\underline{\mathbf{v}} = \mathrm{vec}\,(\mathbf{V})$, $\Sigma_{\mathbf{uu}} = \mathrm{diag}[\left\{\Sigma_{\mathbf{u}_g\mathbf{u}_g}\right\}_{g=1}^{G}]$, $\Sigma_{\mathbf{uv}} = [\Sigma_{\mathbf{uv}_1},\ldots,\Sigma_{\mathbf{uv}_p}]'$, with $\Sigma_{\mathbf{uv}_j} = \mathrm{diag}[\left\{\Sigma_{\mathbf{v}_{j,g}\mathbf{u}_g}\right\}_{g=1}^{G}]$, for $j=1,\ldots,p$, and $\Sigma_{\mathbf{v}_j\mathbf{v}_m} = \mathrm{diag}\left[\left\{\Sigma_{\mathbf{v}_{j,g}\mathbf{v}_{m,g}}\right\}_{g=1}^{G}\right]$, for $j,m=1,\ldots,p$. Let $n=\sum_{g=1}^{G}n_g$ and assume that $\mathrm{rank}\,(\Pi_z)=p$. The TSLS estimator $\hat{\theta}_{\mathrm{IV}}$ can be written as:

$$\hat{\theta}_{\mathrm{IV}}-\theta = \left(\mathbf{y}_2'\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}}\mathbf{y}_2\right)^{-1}\mathbf{y}_2'\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}}\mathbf{u} = \left(\mathrm{I}+\mathbf{Q}^{-1}\Delta\right)^{-1}\mathbf{Q}^{-1}\mathbf{C}, \tag{A-4}$$

where $\mathbf{Q} = \Pi_z'\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z$, $\Delta = \Pi_z'\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{V} + \mathbf{V}'\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z + \mathbf{V}'\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}}\mathbf{V}$, and $\mathbf{C} = \Pi_z'\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{u} + \mathbf{V}'\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}}\mathbf{u}$. We have $\mathbf{Q}^{-1} = n^{-1}\times O_p\,(1)$ and $\Delta = \sqrt{n}\times O_p\,(1)$, which implies that, as $n\longrightarrow+\infty$, $\left(\mathrm{I}+\mathbf{Q}^{-1}\Delta\right)^{-1} = O_p\,(1)$. Using a Taylor expansion (see Nagar (1959)) we derive $\left(\mathrm{I}+\mathbf{Q}^{-1}\Delta\right)^{-1}\approx\mathrm{I}-\mathbf{Q}^{-1}\Delta$. Equation (A-4) can be simplified to:

$$\hat{\theta}_{\mathrm{IV}}-\theta = \mathbf{Q}^{-1}\left\{\mathbf{C}_{\mathbf{Vu}}-\left(\Delta_{\mathbf{V}}+\Delta_{\mathbf{V}}'\right)\mathbf{Q}^{-1}\mathbf{C}_{\mathbf{u}}\right\}+\mathbf{H},$$

where $\mathbf{C}_{\mathbf{Vu}} = \mathbf{V}'\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}}\mathbf{u}$, $\Delta_{\mathbf{V}} = \Pi_z'\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{V}$, $\mathbf{C}_{\mathbf{u}} = \Pi_z'\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{u}$, and $\mathbf{H}$ has terms related to odd moments of the joint distribution of $(\mathbf{u}',\underline{\mathbf{v}}')'$ and terms which are of small order.

[16] The oblique projection $\mathrm{P}_{\mathrm{H}}^{(\widehat{\Omega}_{\delta_w\delta_w})^{-1}}$ satisfies the properties $\mathrm{H}'(\widehat{\Omega}_{\delta_w\delta_w})^{-1}\mathrm{P}_{\mathrm{H}}^{(\widehat{\Omega}_{\delta_w\delta_w})^{-1}} = \mathrm{H}'(\widehat{\Omega}_{\delta_w\delta_w})^{-1}$ and $\mathrm{P}_{\mathrm{H}}^{(\widehat{\Omega}_{\delta_w\delta_w})^{-1}}\mathrm{H} = \mathrm{H}$.

Assuming that the first and third moments of the joint distribution are zeros,[17] the bias of the IV estimator is approximately:

$$\mathrm{E}\left[\hat{\theta}_{\mathrm{IV}}-\theta\right] \approx \mathrm{E}\left\{\mathbf{Q^{-1}}\left[\mathbf{V}'\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z^{\perp}}\mathbf{u}-\mathbf{\Delta_V}\mathbf{Q}^{-1}\mathbf{C_u}\right]\right\},$$

where $\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z^{\perp}} \equiv \mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}}-\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z}$. The first element of $\mathbf{V}'\ \mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z^{\perp}}\mathbf{u}$ is $\mathbf{v}_1'\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z^{\perp}}\mathbf{u}$, whose expectation is $\mathrm{E}[\mathbf{v}_1'\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z^{\perp}}\mathbf{u}] = \mathrm{trace}(\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z^{\perp}}\Sigma_{\mathbf{uv}_1})$. So, $\mathrm{E}[\mathbf{V}'\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z^{\perp}}\mathbf{u}] = [\mathrm{trace}(\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z^{\perp}}\Sigma_{\mathbf{uv}_1}), \ldots, \mathrm{trace}(\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z^{\perp}}\Sigma_{\mathbf{uv}_p})]'$. Partition $\mathbf{Q}^{-1}$ as $\mathbf{Q}^{-1} =[\mathbf{Q}^{\cdot 1}, \ldots, \mathbf{Q}^{\cdot p}]$. Therefore,

$$\mathrm{E}\left[\mathbf{Q^{-1}}\mathbf{V}'\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z^{\perp}}\mathbf{u}\right] = \sum_{j=1}^{p}\mathrm{trace}\left(\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z^{\perp}}\Sigma_{\mathbf{uv}_j}\right)\mathbf{Q}^{\cdot j}. \tag{A-5}$$

To study $\mathrm{E}[\mathbf{Q^{-1}}\mathbf{\Delta_V}\mathbf{Q}^{-1}\mathbf{C_u}]$, we rewrite $\mathbf{\Delta_V}\mathbf{Q}^{-1}\mathbf{C_u}$ as $\mathrm{vec}(\mathbf{C_u'}\mathbf{Q}^{-1}\mathbf{\Delta_V'}) = \Pi_z'\mathbf{Z}'\mathrm{M}_{\mathbf{X}}(\mathbf{V}\otimes\mathbf{u}')(\mathrm{I}_p\otimes\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z)\,\mathrm{vec}(\mathbf{Q}^{-1})$. Since $\mathrm{E}(\mathbf{V}\otimes\mathbf{u}')=[\Sigma_{\mathbf{v}_1\mathbf{u}}, \ldots, \Sigma_{\mathbf{v}_p\mathbf{u}}]$, we have $\mathrm{E}[\mathbf{C_u'}\mathbf{Q}^{-1}\mathbf{\Delta_V'}] = \sum_{j=1}^{p}\Pi_z'\ \mathbf{Z}'\mathrm{M}_{\mathbf{X}}\ \Sigma_{\mathbf{v}_j\mathbf{u}}\ \mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z\mathbf{Q}^{\cdot j}$, and, consequently,

$$\mathrm{E}\left[\mathbf{Q}^{-1}\mathbf{\Delta_V}\mathbf{Q}^{-1}\mathbf{C_u}\right] = \sum_{j=1}^{p}\mathbf{Q}^{-1}\Pi_z'\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\Sigma_{\mathbf{v}_j\mathbf{u}}\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z\mathbf{Q}^{\cdot j}. \tag{A-6}$$

Substituting Equations A-5 and A-6 we have:

$$\mathrm{E}\left[\hat{\theta}_{\mathrm{IV}}-\theta\right] \approx \sum_{j=1}^{p}\mathrm{trace}\left(\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z^{\perp}}\Sigma_{\mathbf{uv}_j}\right)\mathrm{I}_p\mathbf{Q}^{\cdot j} - \sum_{j=1}^{p}\mathbf{Q}^{-1}\Pi_z'\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\Sigma_{\mathbf{v}_j\mathbf{u}}\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z\mathbf{Q}^{\cdot j}. \tag{A-7}$$

When $p=1$, equation (A-7) simplifies to $\mathrm{trace}(\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}}\Sigma_{\mathbf{uv}})\mathbf{Q^{-1}}-2(\Pi_z'\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\Sigma_{\mathbf{uv}}\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z)\mathbf{Q}^{-2}$, which is the same as the bias derived by Bun and de Haan (2010).

Assuming that $(\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{Z}) = n\mathrm{I}_{k_z}$ and $\Pi_z = \|\Pi_z\|\,\Pi_0$, we find:

$$\mathrm{E}\left[\hat{\theta}_{\mathrm{IV}}-\theta\right] \approx \mu^{-2}\frac{\mathrm{trace}\,(\mathrm{S}_{12})}{\mathrm{trace}\,(\mathrm{S}_{22})}\left\{1-2\frac{(\Pi_0'\mathrm{S}_{12}\Pi_0)}{\mathrm{trace}\,(\mathrm{S}_{12})}\right\}$$

where $\mathrm{S}_{12} = n^{-1}\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\Sigma_{\mathbf{uv}}\mathrm{M}_{\mathbf{X}}\mathbf{Z}$, $\mathrm{S}_{22} = n^{-1}\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\Sigma_{\mathbf{vv}}\mathrm{M}_{\mathbf{X}}\mathbf{Z}$, and $\mu^2 = [\mathrm{trace}\,(\mathrm{S}_{22})]^{-1}\,n\,\|\Pi_z\|^2$ represents the concentration parameter for the case of non spherical errors derived by Olea and Pflueger (2013) in Theorem 1. If the residuals are homoskedastic, then $\Sigma_{\mathbf{uv}_j} = \tau_j\mathrm{I}_n$, and, Equation A-7 is simplified to $(k_z-p-1)\,\mathbf{Q}^{-1}\boldsymbol{\tau}$ where $\tau'=[\tau_1, ..., \tau_p]$.

Consider that the errors of the model are generated as $\mathbf{u}_g = \iota_{n_g}\upsilon_g + \varepsilon_g$, and $\mathrm{vec}(\mathbf{v}_g) = \left[\boldsymbol{\nu}_g\otimes\iota_{n_g}\right] + \boldsymbol{\epsilon}_g$, for $g=1,\ldots,G$, where υ_g is scalar, and $\boldsymbol{\nu}_g$ and $\boldsymbol{\epsilon}_g$ are $p\times 1$ and $(n_g p)\times 1$ vectors with $(\upsilon_g, \boldsymbol{\nu}_g)' \sim \sqrt{\phi_g}N(0,\ [1,\ \boldsymbol{\rho}' : \boldsymbol{\rho}],\ \mathrm{I}_p)$, $(\varepsilon_g, \boldsymbol{\epsilon}_g)' \sim \sqrt{\varphi_g}N(0, [1, \boldsymbol{\varrho}' : \boldsymbol{\varrho}, \mathrm{I}_p]\otimes\mathrm{I}_{n_g})$, where $\boldsymbol{\rho}' = [\rho_1, \ldots, \rho_p]$ and $\boldsymbol{\varrho}' = [\varrho_1, \ldots, \varrho_p]$ are $p\times 1$ vectors capturing the intra-cluster effect and the idiosyncratic term correlations and the scalar ϕ_g, $1 \geq \phi_g \geq 0$. So the joint distribution of $\left(\mathbf{u}_g', \mathrm{vec}(\mathbf{v}_g)'\right)'$ is:

[17] If $\mathbf{u}, \mathrm{vec}\,(\mathbf{V})$ follows a multivariate distribution, we can invoke Isselis' or Wick's theorem, which says that the expected value of odd moments of a centered multivariate normal distribution are 0.

$$\begin{pmatrix} \mathbf{u}_g \\ \text{vec}(\mathbf{v}_g) \end{pmatrix} \sim N\left(0, \begin{bmatrix} \mathrm{W}_g + \overline{\mathrm{W}}_g & \left(\boldsymbol{\rho}' \otimes \mathrm{I}_{n_g}\right)\mathrm{W}_g + \left(\boldsymbol{\varrho}' \otimes \mathrm{I}_{n_g}\right)\overline{\mathrm{W}}_g \\ \cdot & \mathrm{I}_p \otimes \left(\mathrm{W}_g + \overline{\mathrm{W}}_g\right) \end{bmatrix}\right)$$

where $\mathrm{W}_g = \phi_g\ \iota_{n_g}\iota'_{n_g}$, $\overline{\mathrm{W}}_g = (1-\phi_g)\ \mathrm{I}_{n_g}$. We interpret ϕ_g and $(1-\phi_g)$ as the weights due to the cluster and idiosyncratic effects in the correlation.

Under this assumption, $\Sigma_{\mathbf{uv}_j} = \text{diag}\{\ \rho_j\mathrm{W}_g\ +\ \varrho_j\overline{\mathrm{W}}_g\}_{g=1}^{G}$ and $\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\Sigma_{\mathbf{uv}_j}\mathrm{M}_{\mathbf{X}}\mathbf{Z} = \mathbf{Z}'\mathrm{M}_{\mathbf{X}}\left(\rho_j\mathrm{W}+\ \varrho_j\overline{\mathrm{W}}\right)\mathrm{M}_{\mathbf{X}}\mathbf{Z}$, for $j = 1,\ldots,p$, where $\mathrm{W} = \text{diag}\left\{\mathrm{W}_g\right\}_{g=1}^{G}$, and $\overline{\mathrm{W}} = \text{diag}\left\{\overline{\mathrm{W}}_g\right\}_{g=1}^{G}$. Then, after further simplifications in equation (A-7), the bias of IV estimator turns out to be

$$\begin{aligned}\mathrm{E}\left[\hat{\theta}_{\mathrm{IV}} - \theta\right] \approx & \left\{\text{trace}\left[\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z^{\perp}}\mathrm{W}\right]\mathrm{I}_p - \mathbf{Q}^{-1}\Pi'_z\left(\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathrm{W}\mathrm{M}_{\mathbf{X}}\mathbf{Z}\right)\Pi_z\right\}\mathbf{Q}^{-1}\boldsymbol{\rho} \\ & + \left\{\text{trace}\left[\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z^{\perp}}\overline{\mathrm{W}}\right]\mathrm{I}_p - \mathbf{Q}^{-1}\Pi'_z\left(\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\overline{\mathrm{W}}\mathrm{M}_{\mathbf{X}}\mathbf{Z}\right)\Pi_z\right\}\mathbf{Q}^{-1}\boldsymbol{\varrho}\end{aligned}$$

The first term captures the bias of the IV estimator due to the cluster effect while the second term is function of the within cluster correlations.

Let us define $\mathbf{Z}$ and $\mathrm{M}_{\mathbf{X}}\mathbf{Z}$ as $\mathbf{Z} = \left[\mathbf{d}'_1\iota'_{n_1} + \vartheta'_1, \ldots, \mathbf{d}'_G\iota'_{n_G} + \vartheta'_G\right]'$ and $\mathrm{M}_{\mathbf{X}}\mathbf{Z} = \left[\mathbf{z}_{\mathbf{1}}^{\perp\prime}, \ldots, \mathbf{z}_{\mathbf{G}}^{\perp\prime}\right] = [\left(\mathbf{d}_1 - \overline{\mathbf{d}}\right)'\iota'_{n_1} + \left(\vartheta_1 - \iota_{n_1}\overline{\vartheta}\right)',\ \ldots,\ \left(\mathbf{d}_G - \overline{\mathbf{d}}\right)'\iota'_{n_G} + \left(\vartheta_G - \iota_{n_G}\overline{\vartheta}\right)']'$ where $\overline{\mathbf{d}} = \left(n^{-1}\sum_{g=1}^{G} n_g\mathbf{d}_g\right)$and $\overline{\vartheta} = \left(n^{-1}\sum_{g=1}^{G}\iota'_{n_g}\vartheta_g\right)$. We interpret $\mathbf{d}_g$ and the part of instruments which is common to all observations in cluster g, while that ϑ_g captures the part which is idiosyncratic for each observation.

Define $\overline{\vartheta}_g = n_g^{-1}\iota'_{n_g}\vartheta_g$. If we further impose in the data generate process that $\overline{\vartheta}_g = \overline{\vartheta} = 0$, $\phi_g = \phi$, $n_g = \bar{n}$ for all g, then we have $\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{Z} = \bar{n}\sum_{g=1}^{G}[\left(\mathbf{d}_g - \overline{\mathbf{d}}\right)'\left(\mathbf{d}_g - \overline{\mathbf{d}}\right) + \vartheta'_g\vartheta_g]$, $\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\overline{\mathrm{W}}\mathrm{M}_{\mathbf{X}}\mathbf{Z} = (1-\phi)\,\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{Z}$, and $\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathrm{W}\mathrm{M}_{\mathbf{X}}\mathbf{Z} = \phi\bar{n}\left(\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{Z} - \sum_{g=1}^{G}\vartheta'_g\vartheta_g\right)$. When generating the data, we reescale the values of $\mathbf{d}_g$ such that $\sum_{g=1}^{G} n_g\left(\mathbf{d}_g - \overline{\mathbf{d}}\right)'\left(\mathbf{d}_g - \overline{\mathbf{d}}\right) = (1-\lambda)\,n\mathrm{I}_{k_z}$ and $(\sum_{g=1}^{G}\vartheta'_g\vartheta_g) = \lambda n\mathrm{I}_{k_z}$, so $\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{Z} = n\mathrm{I}_{k_z}$ for $0 \leq \lambda \leq 1$. In the simulations, we set $\lambda = 0.1$. Then, the bias of TSLS becomes:

$$\begin{aligned}\mathrm{E}\left[\hat{\theta}_{\mathrm{IV}} - \theta\right] \approx & \quad \phi\bar{n}\,(1-\lambda)\,(k_z - p - 1)\,\mathbf{Q}^{-1}\boldsymbol{\rho} + (1-\phi)\,(k_z - p - 1)\,\mathbf{Q}^{-1}\boldsymbol{\varrho} \\ \approx & \quad (k_z - p - 1)\,\mathbf{Q}^{-1}\left[\phi\bar{n}\,(1-\lambda)\,\boldsymbol{\rho} + (1-\phi)\,\boldsymbol{\varrho}\right]\end{aligned}$$

The last equation is the same as equation (17).

D The concentration parameter

When $p = 1$, the cluster robust F-test for testing $H_0 : \Pi = 0$ is $\hat{\mathrm{F}} = k_z^{-1}\widehat{\Pi}_z'[\widehat{\Xi}_{\Pi_z\Pi_z}]^{-1}\widehat{\Pi}_z$, where $\widehat{\Xi}_{\Pi_z\Pi_z}$ is the cluster robust variance estimator. Let $\Xi_{\Pi_z\Pi_z}$ be the true (conditional) variance of $\widehat{\Pi}_z$. The above statistic can be asymptotically approximated by $\mu_{k_z} + F(k_z, +\infty)$ where $\mu_{k_z} = \Pi_z' \left[k_z \Xi_{\Pi_z\Pi_z}\right]^{-1} \Pi_z$, and $F(k_z, +\infty)$ represents the F-distribution. Since only the first instrument is validity, i.e. $\Pi_z = (c_z, 0, ..., 0)$, the noncentrality parameter becomes $\mu_{k_z} = c_z^2 k_z^{-1} \left[\Xi_{\Pi_z\Pi_z}\right]_{11}^{-1}$, where $\left[\Xi_{\Pi_z\Pi_z}\right]_{11}^{-1}$ indicates the first diagonal entry of $[\Xi_{\Pi_z\Pi_z}]^{-1}$. In our simulation experiment, we set $\mu_{k_z} = 0.1$, 1 and 9 to indicate weak and strong instruments, respectively. We fix c_z as

$$c_z = \sqrt{\frac{k_z}{\left[\Xi_{\Pi_z\Pi_z}\right]_{11}^{-1}}\mu_{k_z}}$$

The noncentrality parameter derived above is closely related to the measure proposed by Olea and Pflueger. Their test is

$$\hat{\mathrm{F}}_{eff} = \frac{\mathbf{y}_2'\mathrm{M}_{\mathbf{X}}\mathbf{Z}\left(\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{y}_2}{\operatorname{trace}\left(\hat{\mathrm{S}}_{22}\right)},$$

where $\hat{\mathrm{S}}_{22}$ is an estimator of $n^{-1}\mathrm{E}[\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{V}\mathbf{V}'\mathrm{M}_{\mathbf{X}}\mathbf{Z}]$, which is, under our assumptions $n^{-1}\xi\left(\phi, \bar{n}, \lambda\right)\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{Z}$, where $\xi\left(\phi, \bar{n}, \lambda\right) = \phi\bar{n}\left(1-\lambda\right) + \left(1-\phi\right)$. The $\hat{\mathrm{F}}_{eff}$ statistic can be approximated to

$$\frac{\Pi_z'\left(\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{Z}\right)\Pi_z}{\xi\left(\phi, \bar{n}, \lambda\right)\operatorname{trace}\left(n^{-1}\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{Z}\right)} + \frac{\mathbf{V}'\mathrm{M}_{\mathbf{X}}\mathbf{Z}\left(\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{Z}\right)^{-1}\left(\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{V}\right)}{\xi\left(\phi, \bar{n}, \lambda\right)\operatorname{trace}\left(n^{-1}\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{Z}\right)}$$

or, using the fact that $\Xi_{\Pi\Pi} = \xi\left(\phi, \bar{n}, \lambda\right)\left(\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{Z}\right)^{-1}$ and $n^{-1}\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{Z} = \mathrm{I}_{k_z}$, the later term becomes $\mu_{k_z} + {k_z}^{-1}\mathbf{V}'\mathrm{M}_{\mathbf{X}}\mathbf{Z}\left[\Xi_{\Pi_z\Pi_z}\right]^{-1}\left(\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{V}\right)$. The second term of the above expression is asymptotically distributed as $F(k_z, +\infty)$.

The effective degrees of freedom is defined as:

$$K_{eff} \equiv \frac{\left[\operatorname{trace}\left(\frac{\mathrm{E}[\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{V}\mathbf{V}'\mathrm{M}_{\mathbf{X}}\mathbf{Z}]}{n}\right)\right]^2(1+2x)}{\operatorname{trace}\left(\frac{\mathrm{E}[\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{V}\mathbf{V}'\mathrm{M}_{\mathbf{X}}\mathbf{Z}]}{n}\frac{\mathrm{E}[\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{V}\mathbf{V}'\mathrm{M}_{\mathbf{X}}\mathbf{Z}]}{n}\right) + 2\operatorname{trace}\left(\frac{\mathrm{E}[\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{V}\mathbf{V}'\mathrm{M}_{\mathbf{X}}\mathbf{Z}]}{n}\right)\max\operatorname{eval}\left(\frac{\mathrm{E}[\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{V}\mathbf{V}'\mathrm{M}_{\mathbf{X}}\mathbf{Z}]}{n}\right)x}$$

which, in our case, is simplified to

$$K_{eff} \equiv \frac{\left[\operatorname{trace}\left(\xi\left(\phi, \bar{n}, \lambda\right)n\mathrm{I}_{k_z}\right)\right]^2(1+2x)}{\operatorname{trace}\left(\left[\xi\left(\phi, \bar{n}, \lambda\right)\right]^2 n^2\mathrm{I}_{k_z}\right) + 2\operatorname{trace}\left(\xi\left(\phi, \bar{n}, \lambda\right)n\mathrm{I}_{k_z}\right)\max\operatorname{eval}\left(\xi\left(\phi, \bar{n}, \lambda\right)n\mathrm{I}_{k_z}\right)x}$$

or, $\frac{(\xi(\phi,\bar{n},\lambda)k_z)^2(1+2x)}{(\xi(\phi,\bar{n},\lambda))^2k_z+2(\xi(\phi,\bar{n},\lambda))^2k_zx} = \frac{k_z(1+2x)}{(1+2x)} = k_z.$

References

Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2001. The Colonial Origins of Comparative Development: An Empirical Investigation. *American Economic Review* 91(5): 1369–1401.

———. 2012. The Colonial Origins of Comparative Development: An Empirical Investigation: Reply. *American Economic Review* 102(6): 3077–3110.
http://www.jstor.org/stable/41724682

Albouy, David Y. 2012. The Colonial Origins of Comparative Development: An Empirical Investigation: Comment. *American Economic Review* 102(6): 3059–3076.
http://www.jstor.org/stable/41724681

Anderson, Theodore W. and Herman Rubin. 1949. Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations. *Annals of Mathematical Statistics* 20(1): 46–63.
http://www.jstor.org/stable/2236803

Arellano, Manuel. 1987. Computing Robust Standard Errors for Within-groups Estimators. *Oxford Bulletin of Economics and Statistics* 49(4): 431–434.
http://dx.doi.org/10.1111/j.1468-0084.1987.mp49004006.x

Bun, Maurice and Monique de Haan. 2010. Weak Instruments and the First Stage F-Statistic in IV Models With a Nonscalar Error Covariance Structure. Unpublished manuscript.
http://aseri.uva.nl/binaries/content/assets/subsites/amsterdam-school-of-economics-research-institute/uva-econometrics/dp-2010/1002.pdf

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. Bootstrap-Based Improvements for Inference with Clustered Errors. *Review of Economics and Statistics* 90(3): 414–27.
http://dx.doi.org/10.1162/rest.90.3.414

Chernozhukov, Victor and Christian Hansen. 2008. The Reduced Form: A Simple Approach to Inference with Weak Instruments. *Economics Letters* 100(1): 68–71.
http://www.sciencedirect.com/science/article/pii/S0165176507004107

Davidson, Russell and James G. MacKinnon. 2008. Bootstrap Inference in a Linear Equation Estimated by Instrumental Variables. *Econometrics Journal* 11(3): 443–77.
http://dx.doi.org/10.1111/j.1368-423X.2008.00247.x

———. 2010. Wild Bootstrap Tests for IV Regression. *Journal of Business and Economic Statistics* 28(1): 128–44.
http://pubs.amstat.org/doi/abs/10.1198/jbes.2009.07221

Finlay, Keith, Leandro Magnusson, and Mark E Schaffer. 2013. WEAKIV: Stata module to perform weak-instrument-robust tests and confidence intervals for instrumental-variable (IV) estimation of linear, probit and tobit models. Statistical Software Components, Boston College Department of Economics.
https://ideas.repec.org/c/boc/bocode/s457684.html

Finlay, Keith and Leandro M. Magnusson. 2009. Implementing Weak Instrument Robust Tests for a General Class of Instrumental Variables Models. *Stata Journal* 9(3): 398–421.
http://ideas.repec.org/a/tsj/stataj/v9y2009i3p398-421.html

Gelbach, Jonah B., Jonathan Klick, and Thomas Stratmann. 2007. Cheap Donuts and Expensive Broccoli: The Effect of Relative Prices on Obesity. Working paper.
http://papers.ssrn.com/sol3/papers.cfm?abstract_id=976484

Goldberger, Arthur and Ingram Olkin. 1971. A Minimum-Distance Interpretation of Limited-Information Estimation. *Econometrica* 39(3): 635–639.

Hu, Feifang and John D. Kalbfleisch. 2000. The Estimating Function Bootstrap. *Canadian Journal of Statistics* 28(3): 449–81.
http://dx.doi.org/10.2307/3315958

Hu, Feifang and James V. Zidek. 1995. A Bootstrap Based on the Estimating Equations of the Linear Model. *Biometrika* 82(2): 263–75.
http://biomet.oxfordjournals.org/content/82/2/263.abstract

Kleibergen, Frank. 2002. Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression. *Econometrica* 70(5): 1781–1803.
http://www.jstor.org/stable/3082020

———. 2007. Generalizing Weak Instrument Robust IV Statistics towards Multiple Parameters, Unrestricted Covariance Matrices, and Identification Statistics. *Journal of Econometrics* 139(1): 181–216.
http://www.sciencedirect.com/science/article/pii/S0304407606001084

———. 2011. Improved Accuracy of Weak Instrument Robust GMM Statistics through Bootstrap and Edgeworth Approximations. Unpublished manuscript.
http://www.econ.brown.edu/fac/Frank_Kleibergen/improv_ac.pdf

Kleibergen, Frank and Sophocles Mavroeidis. 2009. Weak Instrument Robust Tests in GMM and the New Keynesian Phillips Curve. *Journal of Business and Economic Statistics* 27(3): 293–311.
http://dx.doi.org/10.1198/jbes.2009.08280

Kleibergen, Frank and Richard Paap. 2006. Generalized Reduced Rank Tests Using the Singular Value Decomposition. *Journal of Econometrics* 133(1): 97–126.
http://www.sciencedirect.com/science/article/pii/S0304407605000850

Kline, Patrick and Andres Santos. 2012. A Score Based Approach to Wild Bootstrap Inference. *Journal of Econometric Methods* 1(1): 23–41.
http://dx.doi.org/10.1515/2156-6674.1006

Liu, Regina Y. 1988. Bootstrap Procedures under Some Non-I.I.D. Models. *Annals of Statistics* 16(4): 1696–1708.
http://www.jstor.org/stable/2241788

Mammen, Enno. 1993. Bootstrap and Wild Bootstrap for High Dimensional Linear Models. *Annals of Statistics* 21(1): 255–285.
http://www.jstor.org/stable/3035590

Miguel, Edward, Shanker Satyanath, and Ernest Sergenti. 2004. Economic Shocks and Civil Conflict: An Instrumental Variables Approach. *Journal of Political Economy* 112(4): 725–53.
http://www.jstor.org/stable/10.1086/421174

Moreira, Marcelo J. 2003. A Conditional Likelihood Ratio Test for Structural Models. *Econometrica* 71(4): 1027–48.
http://www.jstor.org/stable/1555489

Moreira, Marcelo J., Jack R. Porter, and Gustavo A. Suarez. 2009. Bootstrap Validity for the Score Test when Instruments May Be Weak. *Journal of Econometrics* 149(1): 52–64.
http://www.sciencedirect.com/science/article/pii/S0304407608002030

Moulton, Brent R. 1990. An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units. *Review of Economics and Statistics* 72(2): 334–38.
http://www.jstor.org/stable/2109724

Nagar, A. L. 1959. The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations. *Econometrica* 27(4): 575–95.
http://www.jstor.org/stable/1909352

Olea, José Luis Montiel and Carolin Pflueger. 2013. A Robust Test for Weak Instruments. *Journal of Business & Economic Statistics* 31(3): 358–369.
http://dx.doi.org/10.1080/00401706.2013.806694

Sanderson, Eleanor and Frank Windmeijer. 2015. A weak instrument -test in linear {IV} models with multiple endogenous variables. *Journal of Econometrics* pp. –.
http://www.sciencedirect.com/science/article/pii/S0304407615001736

Staiger, Douglas and James H. Stock. 1997. Instrumental Variables Regression with Weak Instruments. *Econometrica* 65(3): 557–86.
http://www.jstor.org/stable/2171753

Stock, James H. and Motohiro Yogo. 2005. Testing for Weak Instruments in Linear IV Regression. In Donald W.K. Andrews and James H. Stock, editors, *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*. Cambridge: Cambridge University Press.
http://www.nber.org/papers/t0284

White, Halbert. 1980. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica* 48(4): 817–38.
http://www.jstor.org/stable/1912934

Wu, C. F. Jeff. 1986. Jackknife, Bootstrap, and Other Resampling Methods in Regression Analysis. *Annals of Statistics* 14(4): 1261–95.
http://www.jstor.org/stable/2241454

Zhan, Zhaoguo. 2010. Detecting Weak Identification by Bootstrap. Mimeo.
http://www.econ.brown.edu/econ/events/jmp_zhaoguo_zhan.pdf

Table 1: Different Bootstrap Methods for Cluster IV

Method	Tests	Weights	Estimator for $\delta_w(\theta_0)$	fixed $\widetilde{\Pi}_z(\theta_0)$?	H_0 imposed?
Estimating Equations (EE)	AR, KLM, CLR	M, Γ, R	$\tilde{\delta}_w(\theta_0)$	Yes	Yes
Residuals Single Equation					
inefficient (SE-in)	AR, KLM, CLR	Γ, R	$\acute{\delta}_w(\theta_0)$	Yes	Yes
new-efficient (SE-neff)	AR, CLR	Γ, R	$\tilde{\delta}_w(\theta_0)$	Yes	Yes
First-stage (SE-1st)	F, F_{eff}	Γ, R	–	Yes	Yes ($\Pi_z = 0$)
Residuals Multiple Equation					
IV (ME-IV)	Wald	Γ, R	–	No	No
inefficient (ME-in)	AR, KLM	Γ, R	$\acute{\delta}_w(\theta_0)$	No	Yes
efficient (ME-eff)	AR, KLM, Wald	Γ, R	$\acute{\delta}_w(\theta_0)$	No	Yes
new-efficient (ME-eff)	KLM	Γ, R	$\tilde{\delta}_w(\theta_0)$	No	Yes
Davidson-MacKinnon (DM)	AR, KLM	Γ, R	$\acute{\delta}_w(\theta_0)$	No	Yes
Pairs	Wald, F	M	–	–	No

Notes: The weights M, Γ and R correspond to the multinomial, gamma, and Rademacher weights, respectively.

Table 2A: Size (in percent) for testing $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$ at the 5% significance level, dgp with group-level random errors, 20 clusters with different number of observations in each cluster

			$\kappa = 0$				$\kappa = 1$				$\kappa = 2$			
			$\mu_{k_z} = 1$ (9.96)		$\mu_{k_z} = 9$ (89.64)		$\mu_{k_z} = 1$ (24.17)		$\mu_{k_z} = 9$ (217.53)		$\mu_{k_z} = 1$ (39.11)		$\mu_{k_z} = 9$ (352.02)	
Test		$\rho =$	0.20	0.95	0.20	0.95	0.20	0.95	0.20	0.95	0.20	0.95	0.20	0.95
Wald	Asymp.		13.68	61.09	17.94	25.04	30.44	55.25	41.53	40.88	43.75	79.94	52.33	53.67
	ME-IV	Γ	8.34	40.87	14.69	19.03	11.24	39.16	16.86	18.45	13.90	67.00	10.99	23.05
		R	8.58	37.81	14.25	17.94	10.39	42.86	15.34	17.43	12.79	70.65	9.22	23.90
	ME-eff	Γ	4.06	21.79	5.15	10.57	7.05	25.51	7.07	10.70	9.56	55.87	4.97	21.11
		R	4.37	14.00	4.14	6.47	6.91	22.53	5.41	8.47	12.06	56.49	6.77	25.22
	Pairs		4.38	23.18	8.71	11.69	3.80	30.53	7.04	11.11	1.15	44.10	2.06	3.56
AR	Asymp.		19.79	19.79	19.79	19.79	12.12	12.12	12.12	12.12	2.60	2.60	2.60	2.60
	EE	M	0.03	0.03	0.03	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		Γ	7.37	7.37	7.37	7.37	3.57	3.57	3.57	3.57	0.54	0.54	0.54	0.54
		R	4.95	4.95	4.95	4.95	2.67	2.67	2.67	2.67	0.45	0.45	0.45	0.45
	SE-in	Γ	7.41	7.41	7.41	7.41	5.35	5.35	5.35	5.35	1.48	1.48	1.48	1.48
		R	5.23	5.23	5.23	5.23	4.14	4.14	4.14	4.14	1.30	1.30	1.30	1.30
	SE-neff	Γ	7.41	7.41	7.41	7.41	5.78	5.78	5.78	5.78	2.02	2.02	2.02	2.02
		R	5.06	5.06	5.06	5.06	4.64	4.64	4.64	4.64	4.31	4.31	4.31	4.31
	DM	Γ	3.75	3.75	3.75	3.75	2.92	2.92	2.92	2.92	11.72	11.72	11.72	11.72
		R	5.81	5.81	5.81	5.81	8.04	8.04	8.04	8.04	15.20	15.20	15.20	15.20
KLM	Asymp.		16.44	21.39	15.71	16.36	13.60	15.12	12.73	12.62	4.83	6.90	4.75	4.96
	EE	M	0.17	0.33	0.12	0.20	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00
		Γ	5.84	8.44	5.41	5.52	3.82	4.62	3.36	3.43	0.95	1.09	0.98	0.99
		R	5.34	7.57	4.74	5.02	3.44	4.13	3.06	3.11	0.87	1.02	0.85	0.87
	SE-in	Γ	6.05	8.89	5.72	5.89	5.27	6.07	5.02	4.58	1.98	2.37	1.68	1.77
		R	5.57	8.41	5.24	5.52	4.81	6.00	4.62	4.60	2.03	2.76	1.87	1.98
	ME-in	Γ	6.07	8.93	5.69	5.85	5.45	6.19	4.85	4.51	2.04	2.46	1.70	1.79
		R	5.33	7.15	5.04	5.24	4.61	5.41	4.49	4.40	1.89	2.50	1.80	1.85
	ME-eff	Γ	5.90	9.00	5.53	5.88	5.33	6.12	4.87	4.54	2.01	2.48	1.68	1.79
		R	5.29	6.75	5.09	5.14	4.64	5.31	4.54	4.45	1.83	2.26	1.80	1.87
	ME-neff	Γ	5.94	8.98	5.38	5.76	5.63	6.74	5.14	5.03	2.65	4.48	2.58	2.88
		R	5.09	6.40	4.89	4.92	4.96	5.82	4.81	4.75	4.95	7.48	4.77	4.85
	DM	Γ	3.05	3.11	3.05	3.28	7.10	7.76	7.76	7.87	13.03	19.56	19.42	20.93
		R	4.21	4.31	4.30	4.55	7.47	7.52	7.57	7.31	12.18	15.67	15.64	16.24
CLR	Asymp.		17.06	21.73	15.80	16.33	13.56	15.14	12.70	12.52	4.77	6.54	4.68	4.92
	EE	M	0.09	0.27	0.11	0.18	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00
		Γ	5.99	8.53	5.44	5.55	3.90	4.59	3.39	3.43	0.91	1.03	0.97	0.98
		R	5.10	7.59	4.71	4.97	3.17	4.01	3.07	3.10	0.85	0.96	0.83	0.86
	SE-in	Γ	6.09	9.00	5.73	5.93	5.33	6.10	5.03	4.56	1.93	2.27	1.68	1.77
		R	5.43	8.24	5.24	5.52	4.72	5.99	4.61	4.60	2.00	2.65	1.87	1.98
	SE-neff	Γ	5.93	8.78	5.52	5.79	5.69	6.74	5.16	4.96	2.62	4.04	2.49	2.79
		R	5.08	7.96	5.02	5.28	5.11	6.67	5.03	4.97	4.97	7.61	4.76	5.26

(Continued).

Table 2B: Size (in percent) for testing $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$ at the 5% significance level, dgp with group-level random errors, 20 clusters with different number of observations in each cluster

			$\kappa = 0$				$\kappa = 1$				$\kappa = 2$			
			$\mu_{k_z} = 1$ (9.96)		$\mu_{k_z} = 9$ (89.64)		$\mu_{k_z} = 1$ (24.17)		$\mu_{k_z} = 9$ (217.53)		$\mu_{k_z} = 1$ (39.11)		$\mu_{k_z} = 9$ (352.02)	
Test		$\rho =$	0.20	0.95	0.20	0.95	0.20	0.95	0.20	0.95	0.20	0.95	0.20	0.95
F	Asymp.		88.08	88.20	100.00	100.00	96.99	89.46	100.00	100.00	95.46	84.61	100.00	98.87
	SE-1^{st}	Γ	25.80	26.70	93.07	93.09	50.04	48.81	99.37	94.43	58.43	31.60	98.75	75.51
		R	16.25	16.92	83.67	83.93	35.21	45.88	97.02	90.89	47.55	36.54	96.49	75.71
	Pairs		12.11	12.27	79.47	79.06	29.84	25.51	95.45	71.09	34.36	11.76	91.09	29.64
Eff. F	Asymp.		0.74	0.81	53.55	53.67	5.32	25.35	97.35	91.45	17.46	28.24	98.30	70.55
	SE-1^{st}	Γ	13.84	13.94	86.88	86.83	30.68	48.78	99.82	99.34	47.22	46.33	99.97	83.26
		R	16.37	16.82	90.21	90.78	36.45	50.74	99.93	99.25	51.21	45.65	100.00	82.70

Notes: Authors' calculation from 10,000 Monte Carlo simulations. For bootstrap methods, each experiment consists of 199 bootstrap replications. The number in parenthesis is the concentration parameter divided by k_z assuming homoskedastic errors. The weights M, Γ and R correspond to the multinomial, gamma, and Rademacher weights, respectively. Sample size is 410 observations.

Table 3A: Size (in percent) for testing $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$ at the 5% significance level, with strong heteroskedastic group-level random errors, $\rho = 0.95$, and different number of observations in each cluster.

			$\mu_{k_z} = 0.1$					$\mu_{k_z} = 1$					$\mu_{k_z} = 9$				
			(0.70)	(2.42)	(1.66)	(2.16)	(2.13)	(7.05)	(24.17)	(16.59)	(21.56)	(21.34)	(63.42)	(217.53)	(149.28)	(194.06)	(192.08)
Test		$G =$	10	20	40	80	160	10	20	40	80	160	10	20	40	80	160
Wald	Asymp.		95.51	90.40	88.55	86.31	87.46	71.84	55.25	46.82	35.16	35.05	31.35	40.88	17.62	11.04	9.57
	ME-IV	Γ	84.18	81.55	72.83	69.55	69.94	48.69	39.16	30.19	21.01	19.12	15.05	18.45	10.46	8.18	7.63
		R	83.58	82.20	73.79	70.33	70.66	48.76	42.86	28.02	18.81	17.63	11.96	17.43	7.90	6.31	6.68
	ME-eff	Γ	63.12	55.65	29.40	13.09	7.32	33.23	25.51	15.94	9.88	7.76	11.81	10.70	9.36	7.53	7.02
		R	55.65	51.18	26.82	11.80	7.02	27.23	22.53	12.39	7.51	6.33	8.35	8.47	6.36	5.27	5.59
	Pairs		61.07	66.80	63.96	63.54	65.33	26.08	30.53	19.14	15.43	14.32	5.97	11.11	6.52	6.17	6.87
AR	Asymp.		46.00	12.12	7.66	5.28	4.28	46.00	12.12	7.66	5.28	4.28	46.00	12.12	7.66	5.28	4.28
	EE	M	0.00	0.00	0.02	0.19	0.65	0.00	0.00	0.02	0.19	0.65	0.00	0.00	0.02	0.19	0.65
		Γ	4.33	3.57	6.03	5.89	5.72	4.33	3.57	6.03	5.89	5.72	4.33	3.57	6.03	5.89	5.72
		R	1.96	2.67	4.64	4.77	4.67	1.96	2.67	4.64	4.77	4.67	1.96	2.67	4.64	4.77	4.67
	SE-in	Γ	6.06	5.35	6.39	5.84	5.43	6.06	5.35	6.39	5.84	5.43	6.06	5.35	6.39	5.84	5.43
		R	5.33	4.14	5.06	4.87	4.68	5.33	4.14	5.06	4.87	4.68	5.33	4.14	5.06	4.87	4.68
	SE-neff	Γ	6.60	5.78	6.49	5.82	5.37	6.60	5.78	6.49	5.82	5.37	6.60	5.78	6.49	5.82	5.37
		R	5.51	4.64	5.08	4.81	4.66	5.51	4.64	5.08	4.81	4.66	5.51	4.64	5.08	4.81	4.66
	DM	Γ	2.36	2.92	5.77	5.55	5.37	2.36	2.92	5.77	5.55	5.37	2.36	2.92	5.77	5.55	5.37
		R	12.13	8.04	5.57	5.38	4.82	12.13	8.04	5.57	5.38	4.82	12.13	8.04	5.57	5.38	4.82
KLM	Asymp.		54.47	23.14	16.33	10.15	7.68	42.62	15.12	10.77	7.28	5.96	33.18	12.62	9.41	6.88	5.72
	EE	M	0.00	0.00	0.20	0.80	2.24	0.00	0.01	0.08	0.40	1.61	0.00	0.01	0.08	0.31	1.54
		Γ	7.10	7.49	10.65	8.21	6.70	5.42	4.62	6.53	5.64	5.30	4.12	3.43	5.36	5.10	5.01
		R	3.77	6.79	10.10	7.99	6.71	2.81	4.13	6.07	5.65	5.11	2.15	3.11	5.11	4.98	4.92
	SE-in	Γ	10.51	10.32	10.52	7.71	6.48	7.88	6.07	6.57	5.45	5.07	5.98	4.58	5.26	4.91	4.77
		R	10.75	9.46	10.51	7.91	6.77	7.76	6.00	6.21	5.47	5.04	6.13	4.60	5.29	4.90	4.81
	ME-in	Γ	10.37	10.12	10.82	7.91	6.44	7.90	6.19	6.46	5.52	5.23	6.20	4.51	5.44	4.94	4.87
		R	9.68	8.20	8.99	7.35	6.27	7.42	5.41	5.72	5.09	4.90	5.98	4.40	5.12	4.84	4.86
	ME-eff	Γ	10.41	10.40	10.22	7.67	6.18	7.83	6.12	6.60	5.43	5.08	6.06	4.54	5.40	4.86	4.69
		R	8.94	7.47	7.62	6.10	5.38	7.08	5.31	5.46	5.13	4.91	5.82	4.45	5.12	4.98	4.78
	ME-neff	Γ	10.28	11.05	10.30	7.66	6.22	8.15	6.74	6.52	5.47	5.07	6.27	5.03	5.33	4.96	4.61
		R	8.15	8.23	7.70	6.12	5.34	6.67	5.82	5.56	5.06	4.95	5.42	4.75	5.14	4.98	4.85
	DM	Γ	3.94	7.05	7.44	7.18	6.43	8.61	7.76	7.57	6.70	5.83	10.01	7.87	7.57	6.58	5.57
		R	5.64	7.42	6.33	6.57	5.91	7.23	7.52	5.83	5.65	5.10	7.87	7.31	5.54	5.44	4.97
CLR	Asymp.		62.05	22.93	14.99	9.56	6.73	51.61	15.14	10.67	7.21	5.80	43.20	12.52	9.33	6.78	5.69
	EE	M	0.00	0.00	0.07	0.37	1.31	0.00	0.00	0.08	0.36	1.53	0.00	0.01	0.08	0.31	1.51
		Γ	6.36	7.29	9.88	8.21	6.81	5.02	4.59	6.52	5.69	5.35	3.63	3.43	5.35	5.09	5.00
		R	3.19	6.36	9.04	7.65	6.38	2.43	4.01	5.98	5.62	5.21	1.97	3.10	5.09	5.03	4.94
	SE-in	Γ	9.24	10.14	10.01	7.73	6.50	7.06	6.10	6.57	5.46	5.07	5.29	4.56	5.25	4.92	4.77
		R	9.21	9.25	9.53	7.74	6.18	6.86	5.99	6.12	5.44	5.03	5.20	4.60	5.28	4.86	4.82
	SE-eff	Γ	8.86	10.80	9.97	7.80	6.50	7.49	6.74	6.54	5.42	5.07	5.64	4.96	5.39	4.89	4.81
		R	7.91	10.19	9.42	7.73	6.15	6.39	6.67	6.19	5.41	5.04	4.84	4.97	5.18	4.92	4.76

(Continued).

Table 3B: Size (in percent) for testing $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$ at the 5% significance level, with strong heteroskedastic group-level random errors, $\rho = 0.95$, and different number of observations in each cluster.

			$\mu_{k_z} = 0.1$					$\mu_{k_z} = 1$					$\mu_{k_z} = 9$				
			(0.70)	(2.42)	(1.66)	(2.16)	(2.13)	(7.05)	(24.17)	(16.59)	(21.56)	(21.34)	(63.42)	(217.53)	(149.28)	(194.06)	(192.08)
Test		$G =$	10	20	40	80	160	10	20	40	80	160	10	20	40	80	160
F	Asymp.		92.72	79.61	32.30	19.04	13.08	97.17	89.46	68.82	62.06	51.74	99.98	99.99	99.97	99.98	99.99
	SE-1st	Γ	7.92	12.44	12.77	12.69	10.64	16.85	48.81	46.38	53.15	46.60	60.27	94.43	99.55	99.92	99.98
		R	5.98	10.01	7.76	8.42	8.18	13.22	45.88	38.19	44.98	40.86	48.46	90.89	98.58	99.89	99.97
	Pairs		0.23	5.20	2.38	2.72	4.09	0.86	25.51	13.94	25.16	28.68	5.52	71.09	77.58	97.55	99.74
Eff. F	Asymp.		13.91	2.57	0.01	0.00	0.00	28.34	25.35	0.75	0.04	0.03	85.71	91.45	58.86	61.75	48.79
	SE-1st	Γ	9.28	10.50	12.62	11.91	10.42	19.16	48.78	47.64	53.95	50.97	73.15	99.34	97.64	99.93	99.99
		R	7.23	11.67	9.12	9.90	9.23	16.03	50.74	43.21	50.62	48.41	68.12	99.25	97.02	99.83	99.98

Notes: Authors' calculation from 10,000 Monte Carlo simulations. For bootstrap methods, each experiment consists of 199 bootstrap replications. The number in parenthesis is the concentration parameter divided by k_z assuming homoskedastic errors. The weights M, Γ and R correspond to the multinomial, gamma, and Rademacher weights, respectively. The sample sizes are 205, 410, 820, 1640 are 3280 observations for $G = 20$, 40, 80, and 160, respectively.

Table 4: Application: Economic Growth and Civil Conflict

	Dependent Variable: Civil Conflict ≥ 25	
	(1)	(2)
$\Delta\text{GDP}_{i,t}$	-0.412	-1.132
	(1.479)	(1.403)
$\Delta\text{GDP}_{i,t-1}$	-2.249^{**}	-2.547^{**}
	(1.074)	(1.103)
1^{st} F-test (cluster) $\Delta\text{GDP}_{i,t}$	5.739	4.491
p-value	0.003	0.011
1^{st} F-test (cluster) $\Delta\text{GDP}_{i,t-1}$	3.935	3.642
p-value	0.020	0.026
Kleibergen and Paap (2006) rank-test	15.287	16.195
p-value	0.000	0.000
Included instruments	49	82

Notes: Number of observations are 743 with 41 unbalanced clusters. Standard errors, corrected for arbitrary forms of heteroskedasticity and autocorrelation, are in parenthesis. The excluded instruments are growth in rainfall at t and growth in rainfall at $t-1$. In model (1) the included exogenous variables are country specific time trends, log of the GDP per capita in 1979, log of the proportion that a country is mountainous, log of the nation population at $t-1$, an indicator for the countries which are oil-exporters, ethnolinguistic fractionization, religious fractionization, and measures of democracy. In model (2) the included exogenous variables are country specific dummies and country specific trends.

$**$ Significant different from zero at 5% significant level.

Table 5A: 95% Confidence Intervals, Wald, AR, and Bootstrapped Tests

	Original AJR Series	Original AJR series, capped at 250	Original AJR Series, Albouy campaign dummy	Original AJR series, capped at 250, Albouy campaign dummy	Original AJR series, without contested observations in West and Central Africa	Original AJR series, without contested observations in West and Central Africa, mortality capped at 250
	(1)	(2)	(3)	(4)	(5)	(6)
No covariates						
Wald-asymp	$[0.53, 1.32]$	$[0.55, 1.09]$	$[0.34, 1.84]$	$[0.43, 1.30]$	$[0.50, 1.24]$	$[0.51, 1.03]$
Wald-boot ME-eff	$[0.51, 1.64]$	$[0.56, 1.21]$	$[0.35, 3.31]$	$[0.47, 1.53]$	$[0.47, 1.55]$	$[0.53, 1.15]$
AR *AJR*	$[0.67, 1.73]$	$[0.61, 1.20]$	$[0.64, 3.96]$	$[0.52, 1.55]$	$[0.62, 1.62]$	$[0.57, 1.12]$
AR	$[0.65, 2.16]$	$[0.61, 1.49]$	$[0.63, 5.16]$	$[0.55, 2.17]$	$[0.58, 2.09]$	$[0.54, 1.40]$
AR-boot EE	$[0.65, 2.20]$	$[0.61, 1.53]$	$[0.63, 5.29]$	$[0.55, 2.26]$	$[0.59, 2.35]$	$[0.54, 1.47]$
AR-boot SE-neff	$[0.64, 2.10]$	$[0.61, 1.49]$	$[0.61, 6.84]$	$[0.53, 2.24]$	$[0.60, 2.04]$	$[0.56, 1.35]$
F-stat *p-value*	0.000	0.000	0.014	0.000	0.000	0.000
F-boot *p-value*	0.014	0.001	0.053	0.004	0.015	0.001
Eff. F-boot *p-value*	0.000	0.001	0.046	0.003	0.014	0.001
With latitude						
Wald-asymp	$[0.44, 1.48]$	$[0.50, 1.09]$	$[0.16, 2.15]$	$[0.34, 1.36]$	$[0.43, 1.35]$	$[0.47, 1.04]$
Wald-boot ME-eff	$[0.45, 2.35]$	$[0.48, 1.25]$	$[-7.00, -0.13] \cup [0.33, 6.19]$	$[0.35, 1.74]$	$[0.42, 2.02]$	$[0.45, 1.20]$
AR *AJR*	$[0.64, 2.50]$	$[0.55, 1.20]$	$[0.61, 34.78]$	$[0.41, 1.71]$	$[0.59, 2.08]$	$[0.51, 1.14]$
AR	$[0.61, 7.43]$	$[0.53, 1.77]$	$-\infty, -7.10] \cup [0.59, +\infty$	$[0.45, 3.48]$	$[0.54, 4.77]$	$[0.46, 1.64]$
AR-boot EE	$[0.61, 4.99]$	$[0.52, 1.71]$	$[0.57, +\infty$	$[0.41, 3.05]$	$[0.57, 6.13]$	$[0.46, 1.78]$
AR-boot SE-neff	$[0.61, 5.55]$	$[0.49, 1.61]$	$-\infty, -4.26] \cup [0.52, +\infty$	$[0.16, 3.66]$	$[0.57, 4.19]$	$[0.47, 1.55]$
F-stat *p-value*	0.006	0.000	0.053	0.002	0.004	0.000
F-boot *p-value*	0.057	0.006	0.142	0.024	0.061	0.007
Eff. F-boot *p-value*	0.052	0.005	0.130	0.020	0.058	0.006
With continent dummies and latitude						
Wald-asymp	$[0.03, 2.12]$	$[0.33, 1.28]$	$[-0.25, 2.62]$	$[0.24, 1.43]$	$[0.17, 1.82]$	$[0.37, 1.30]$
Wald-boot ME-eff	$[-6.57, -0.10] \cup [0.30, 6.42]$	$[0.24, 1.40]$	$[-11.69, 11.65]$	$[-0.05, 2.01]$	$[0.32, 3.79]$	$[0.30, 1.44]$
AR *AJR*	$-\infty, -4.74] \cup [0.46, +\infty$	$[0.31, 1.53]$	$-\infty, -1.16] \cup [0.37, +\infty$	$[0.13, 2.21]$	$[0.47, 20.55]$	$[0.41, 1.56]$
AR	$-\infty, -1.71] \cup [0.40, +\infty$	$[0.17, 1.81]$	$-\infty, -0.79] \cup [0.43, +\infty$	$[0.09, 7.74]$	$[0.41, +\infty$	$[0.30, 1.93]$
AR-boot EE	$-\infty, -2.00] \cup [0.42, +\infty$	$[0.24, 1.66]$	$-\infty, -0.93] \cup [0.47, +\infty$	$[0.13, 3.49]$	$[0.42, +\infty$	$[0.33, 1.71]$
AR-boot SE-neff	$-\infty, -1.21] \cup [0.31, +\infty$	$[-0.06, 1.78]$	$-\infty, -0.47] \cup [0.36, +\infty$	$[-0.25, 8.64]$	$[0.31, +\infty$	$[0.10, 1.90]$
F-stat *p-value*	0.096	0.005	0.185	0.027	0.056	0.003
F-boot *p-value*	0.175	0.016	0.295	0.082	0.161	0.017
Eff. F-boot *p-value*	0.153	0.011	0.269	0.059	0.140	0.000

(Continued).

Table 5B: 95% Confidence Intervals, Wald, AR, and Bootstrapped Tests *(Continued)*

	Original AJR Series	Original AJR series, capped at 250	Original AJR Series, Albouy campaign dummy	Original AJR series, capped at 250, Albouy campaign dummy	Original AJR series, without contested observations in West and Central Africa	Original AJR series, without contested observations in West and Central Africa, mortality capped at 250
	(1)	(2)	(3)	(4)	(5)	(6)
With percent of European descent in 1975						
Wald-asymp	$[0.30, 1.54]$	$[0.34, 1.08]$	$[-0.24, 2.61]$	$[0.13, 1.34]$	$[0.30, 1.37]$	$[0.31, 1.03]$
Wald-boot ME-eff	$[0.41, 3.25]$	$[0.28, 1.21]$	$[-17.07, -0.03] \cup [0.24, 15.55]$	$[0.13, 1.84]$	$[0.36, 2.31]$	$[0.21, 1.15]$
AR *AJR*	$[0.53, 4.309]$	$[0.36, 1.21]$	$-\infty, -2.30] \cup [0.47, +\infty$	$[0.11, 1.96]$	$[0.48, 2.73]$	$[0.32, 1.13]$
AR	$[0.42, +\infty$	$[0.15, 1.52]$	$-\infty, -1.51] \cup [0.44, +\infty$	$[0.10, 5.22]$	$[0.34, 9.12]$	$[0.09, 1.28]$
AR-boot EE	$[0.41, +\infty$	$[0.18, 1.60]$	$-\infty, -1.52] \cup [0.44, +\infty$	$[0.12, 7.25]$	$[0.32, 20.09]$	$[0.10, 1.26]$
AR-boot SE-neff	$[0.40, +\infty$	$[0.12, 1.50]$	$-\infty, -1.30] \cup [0.41, +\infty$	$[-0.10, 5.34]$	$[0.33, 5.20]$	$[0.06, 1.29]$
F-stat *p-value*	0.025	0.000	0.153	0.012	0.015	0.000
F-boot *p-value*	0.087	0.006	0.243	0.035	0.079	0.005
Eff. F-boot *p-value*	0.082	0.006	0.223	0.030	0.073	0.005

Notes: All variables from (Acemoglu et al., 2001). Dependent variable is log of GDP per capita in 1995. Right hand side variable is protection against expropriation, instruments by log settler mortality. Column 2 uses original settler mortality series, capped at 250 per 1,000 per annum. Column 3 uses original settler mortality series from (Acemoglu et al., 2001) as the instruments includes Albouy's campaign dummy. Column 4 do the same as Column 3 but caps mortality at 250 per 1,000 per annum. Column 5 is the same as Column 1 but drops the contested observations for West and Central Africa, and Column 6 is the same as Column 5 but caps mortality at 250 per 1,000. The number of observations in Columns 1, 2, 3 and 4 are 62, with 35 the number of clusters. In Columns 4 and 5, the number of observations are 51 with 34 clusters. For bootstrap methods, each experiment consists of 1999 bootstrap replications.

Table 6A: 95% Confidence Intervals, Wald, AR, and Bootstrapped Tests

	Albouy preferred sample	Albouy preferred sample, without Gambia	Albouy preferred sample; campaign dummy	Albouy preferred sample, capped at 250; extended correction Albouy campaign dummy; without Gambia
	(1)	(2)	(3)	(4)
No covariates				
Wald-asymp	$[0.45, 1.29]$	$[0.52, 0.97]$	$[0.01, 2.03]$	$[0.49, 1.17]$
Wald-boot ME-eff	$[0.46, 1.79]$	$[0.53, 1.03]$	$[-9.42, 38.02]$	$[0.53, 1.42]$
AR *AJR*	$[0.59, 1.82]$	$[0.55, 1.02]$	$-\infty, -4.20] \cup [0.43, +\infty$	$[0.59, 1.33]$
AR MD	$[0.57, 2.34]$	$[0.54, 1.15]$	$-\infty, -2.19] \cup [0.46, +\infty$	$[0.56, 2.64]$
AR-boot EE	$[0.57, 2.64]$	$[0.54, 1.21]$	$-\infty, -2.10] \cup [0.48, +\infty$	$[0.56, 2.25]$
AR-boot SE-neff	$[0.57, 2.63]$	$[0.53, 1.18]$	$-\infty, -1.54] \cup [0.46, +\infty$	$[0.54, 2.38]$
AR-boot D&M	$[0.56, 2.64]$	$[0.53, 1.20]$	$-\infty, -1.70] \cup [0.46, +\infty$	$[0.53, 2.62]$
F-stat *p-value*	0.002	0.000	0.102	0.000
F-boot *p-value*	0.026	0.003	0.183	0.017
Eff. F-boot *p-value*	0.024	0.002	0.165	0.013
With latitude				
Wald-asymp	$[0.16, 1.48]$	$[0.34, 0.93]$	$[-0.93, 2.73]$	$[0.29, 1.03]$
Wald-boot ME-eff	$[-2.04, 4.16]$	$[0.29, 1.00]$	$[-30.30, 27.55]$	$[0.03, 1.09]$
AR *AJR*	$[0.42, 19.04]$	$[0.35, 0.96]$	$-\infty, +\infty$	$[0.28, 1.08]$
AR	$-\infty, +\infty$	$[0.20, 1.58]$	$-\infty, +\infty$	$[-0.24, 1.63]$
AR-boot EE	$-\infty, +\infty$	$[0.26, 1.40]$	$-\infty, +\infty$	$[0.22, 1.52]$
AR-boot SE-neff	$-\infty, +\infty$	$[0.16, 1.36]$	$-\infty, +\infty$	$[0.04, 1.77]$
AR-boot D&M	$-\infty, +\infty$	$[0.12, 1.54]$	$-\infty, +\infty$	$[0.04, 1.93]$
F-stat *p-value*	0.052	0.000	0.403	0.001
F-boot *p-value*	0.178	0.019	0.498	0.043
Eff. F-boot *p-value*	0.168	0.015	0.472	0.031
With continent dummies and latitude				
Wald-asymp	$[-1.07, 3.56]$	$[0.08, 1.46]$	$[-3.69, 6.56]$	$[0.00, 1.37]$
Wald-boot EME	$[-18.46, 20.01]$	$[-0.02, 1.99]$	$[-49.73, 48.68]$	$[-0.34, 1.34]$
AR *AJR*	$-\infty, -0.343] \cup [0.107, +\infty]$	$[0.09, 1.72]$	$-\infty, +\infty$	$[-0.11, 1.29]$
AR	$-\infty, +\infty$	$-\infty, -70.12] \cup [-0.10, +\infty$	$-\infty, +\infty$	$[-0.94, 2.05]$
AR-boot EE	$-\infty, +\infty$	$[0.04, 4.28]$	$-\infty, +\infty$	$[-0.14, 1.70]$
AR-boot SE-neff	$-\infty, +\infty$	$[-0.21, 22.79]$	$-\infty, +\infty$	$[-0.36, 1.90]$
AR-boot D&M	$-\infty, +\infty$	$-\infty, -7.408] \cup [-0.231, +\infty$	$-\infty, +\infty$	$[-0.48, 4.21]$
F-stat *p-value*	0.330	0.007	0.583	0.003
F-boot *p-value*	0.439	0.056	0.631	0.036
Eff. F-boot *p-value*	0.404	0.037	0.584	0.016

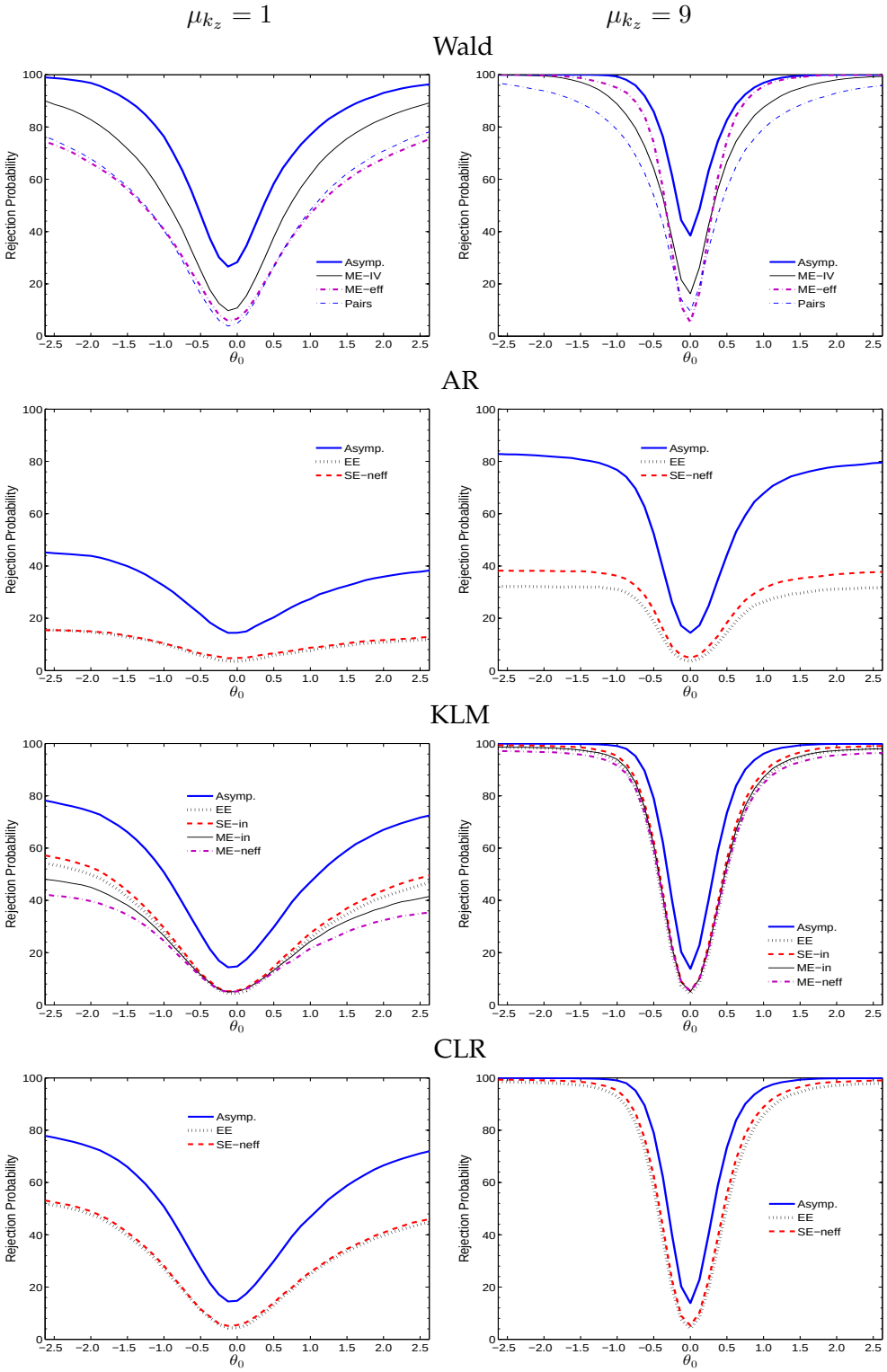
(Continued).

Table 6B: 95% Confidence Intervals, Wald, AR, and Bootstrapped Tests *(Continued)*

	Albouy preferred sample	Albouy preferred sample, without Gambia	Albouy preferred sample; campaign dummy	Albouy preferred sample, capped at 250; extended correction Albouy campaign dummy; without Gambia
	(1)	(2)	(3)	(4)
With percent of European descent in 1975				
Wald-asymp	$[-0.27, 2.16]$	$[0.18, 1.12]$	$[-1.11, 3.38]$	$[0.06, 1.33]$
Wald-boot ME-eff	$[-13.94, 13.38]$	$[0.21, 1.41]$	$[-26.57, 26.53]$	$[-0.20, 1.96]$
AR *AJR*	$-\infty, -1.459] \cup [0.322, +\infty$	$[0.24, 1.37]$	$-\infty, +\infty$	$[-0.02, 1.90]$
AR	$-\infty, +\infty$	$[0.07, 3.08]$	$-\infty, +\infty$	$-\infty, +\infty$
AR-boot EE	$-\infty, +\infty$	$[0.07, 2.52]$	$-\infty, +\infty$	$[-0.66, 4.92]$
AR-boot SE-neff	$-\infty, +\infty$	$[0.03, 2.47]$	$-\infty, +\infty$	$[-1.13, 45.09]$
AR-boot D&M	$-\infty, +\infty$	$[0.03, 3.47]$	$-\infty, +\infty$	$-\infty, +\infty$
F-stat *p-value*	0.168	0.001	0.368	0.022
F-boot *p-value*	0.290	0.026	0.472	0.074
Eff. F-boot *p-value*	0.275	0.021	0.452	0.057

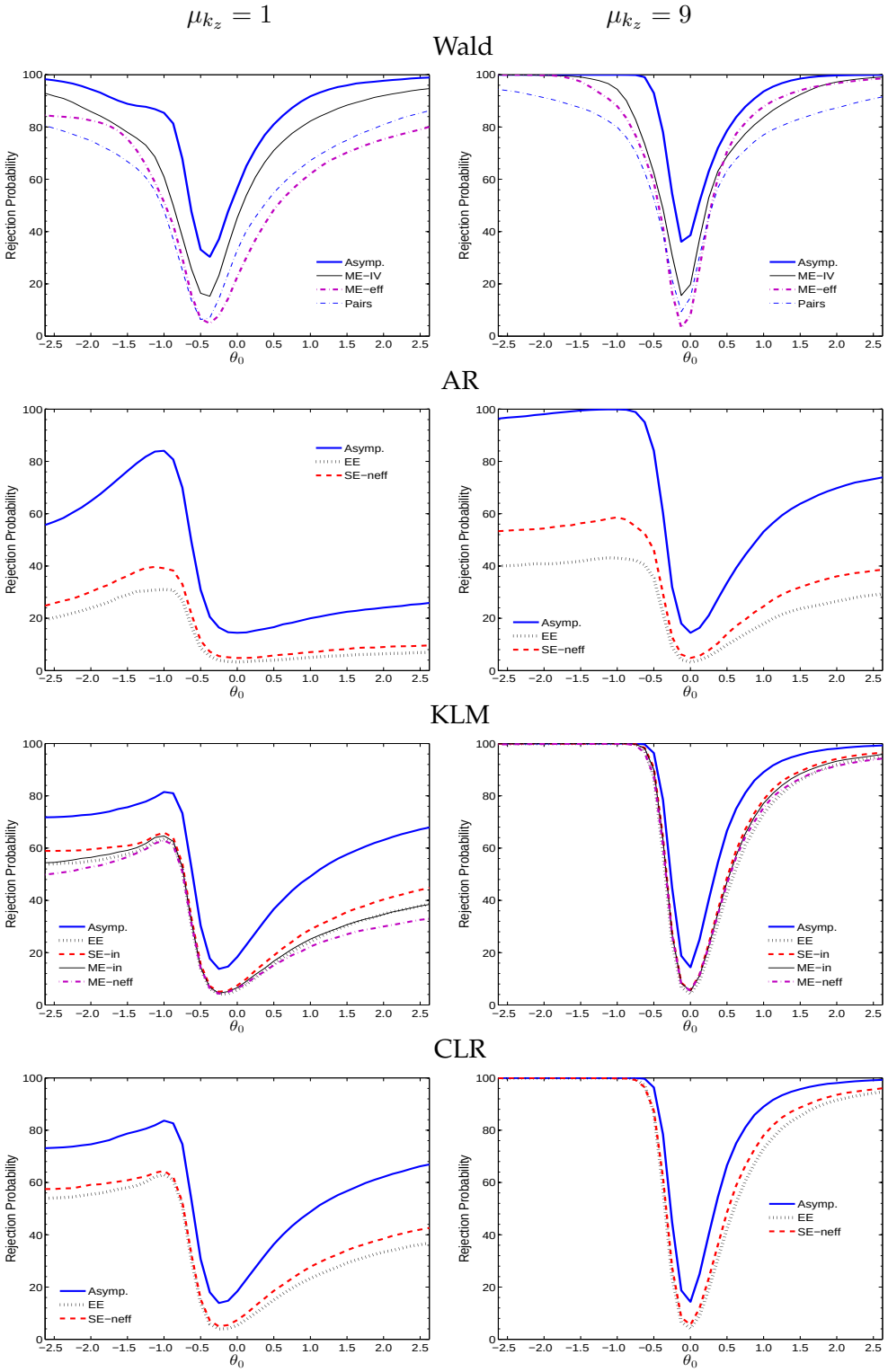
Notes: All variables from (Acemoglu et al., 2001). Dependent variable is log of GDP per capita in 1995. Right hand side variable is protection against expropriation instrumented by log settler mortality. Column 1 uses uses original settler mortality series from (Acemoglu et al., 2001) as the instrument, but Albouy's preferred sample of 28 countries. Column 2 is the same as Column 1 but drops Gambia. Column 3 uses original settler mortality series from (Acemoglu et al., 2001) as the instruments but includes Albouy's campaign dummy. Column 4 is the same as Column 3 uses the extended correction of (Acemoglu et al., 2012) of the campaign dummy, drops Gambia, and caps mortality at 250. The number of observations is 28 in Columns 1 and 3, and 27 observations in Columns 2 and 4. There is no cluster. For bootstrap methods, each experiment consists of 1999 bootstrap replications.

Figure 1: Power Curve for Testing $H_0 : \theta = \theta_0$ at 5% significance level, using $\rho = 0.20$



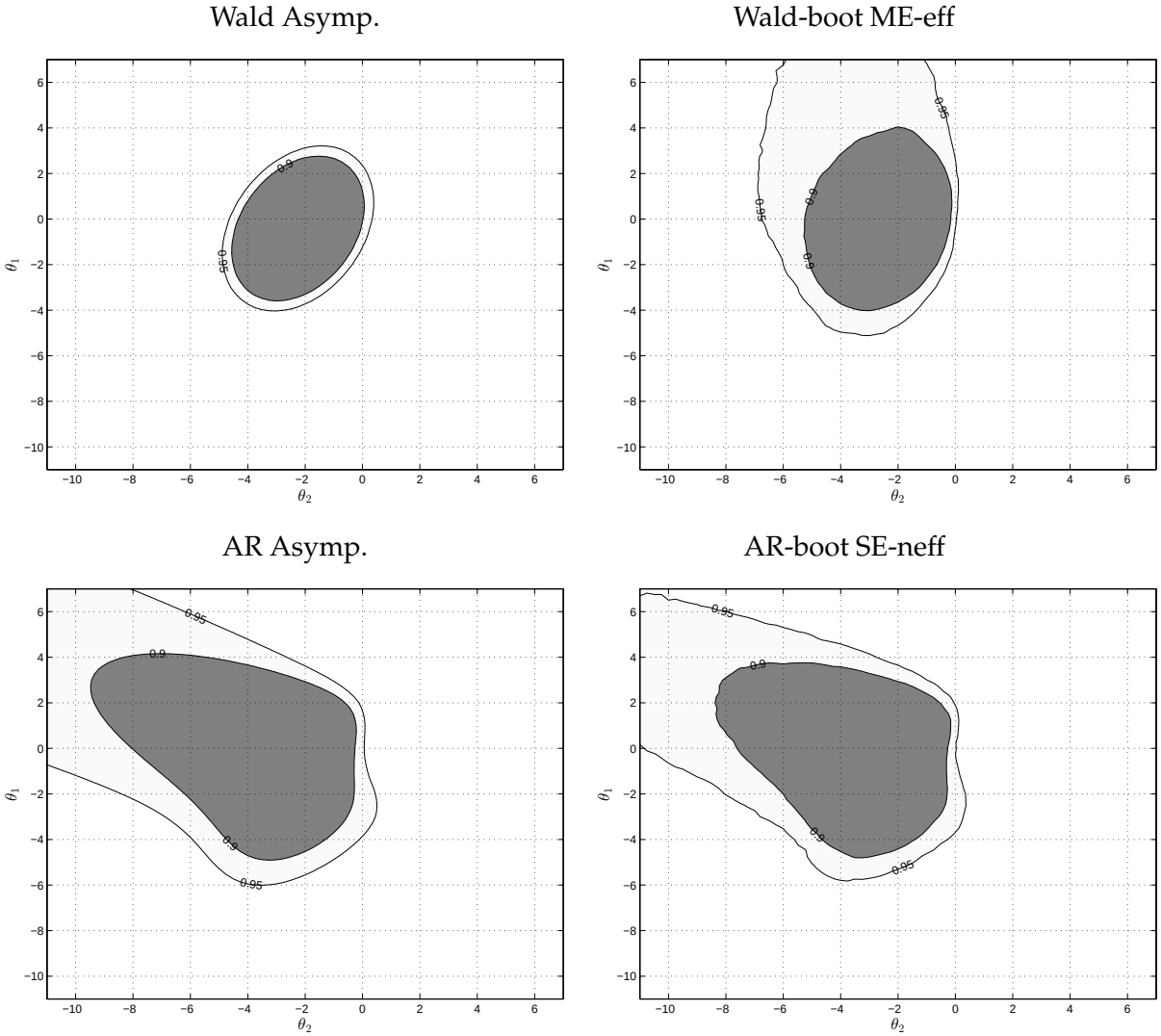
Notes: Authors' calculation from 10,000 Monte Carlo simulations. For bootstrap methods, each experiment consists of 499 bootstrap replications. The bootstrap tests uses Rademacher weights. Sample size is 410 observations, 20 clusters.

Figure 2: Power Curve for Testing $H_0 : \theta = \theta_0$ at 5% significance level, using $\rho = 0.95$



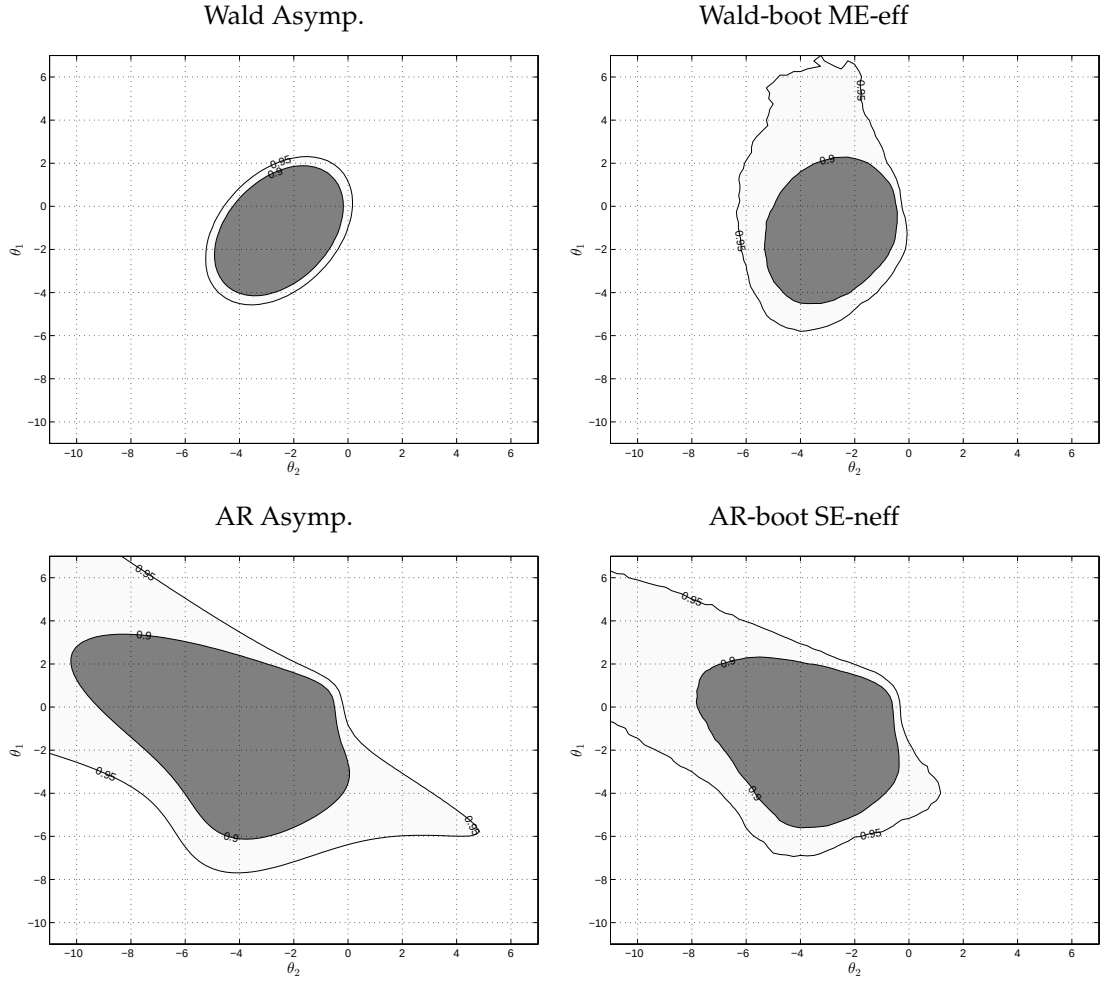
Notes: Authors' calculation from 10,000 Monte Carlo simulations. For bootstrap methods, each experiment consists of 499 bootstrap replications. The bootstrap tests uses Rademacher weights. Sample size is 410 observations, 20 clusters.

Figure 3A: Asymptotic and wild bootstrap tests, 90% and 95% confidence regions.



Notes: Dependent variable is Civil Conflict ≥ 25 Deaths. The right hand side endogenous variables are the economic growth rate and its lagged value, instrumented by growth in rainfall at t and growth in rainfall at $t-1$. The included exogenous variables are country specific time trends, log of the GDP per capita in 1979, log of the proportion that a country is mountainous, log of the nation population at $t-1$, an indicator for the countries which are oil-exporters, ethnolinguistic fractionization, religious fractionization, and measures of democracy. Original Miguel et al. (2004) data set is used. For bootstrap methods, each experiment consists of 1999 bootstrap replications, using Rademacher weights.

Figure 3B: Asymptotic and wild bootstrap tests, 90% and 95% confidence regions.



Notes: Dependent variable is Civil Conflict $\geq$ 25 Deaths. The right hand side endogenous variables are the economic growth rate and its lagged value, instrumented by growth in rainfall at t and growth in rainfall at $t-1$. The included exogenous variables are country specific dummies, country specific trends. Original Miguel et al. (2004) data set is used. For bootstrap methods, each experiment consists of 1999 bootstrap replications, using Rademacher weights.