Wild Bootstrap for Instrumental Variables Regression with Weak Instruments and Few Clusters

Wang, Wenjie

21 February 2021

Wild Bootstrap for Instrumental Variables Regression with Weak Instruments and Few Clusters

Wenjie Wang[*]

February 21, 2021

Abstract

Under a framework with a small number of clusters but large numbers of observations per cluster for instrumental variable (IV) regression, we show that an unstudentized wild bootstrap test based on IV estimators such as the two-stage least squares estimator is valid as long as the instruments are strong for at least one cluster. This is different from alternative methods proposed in the literature for inference with a small number of clusters, whose validity would require that the instruments be strong for all clusters. Moreover, for the leading case in empirical applications with a single instrument, the unstudentized wild bootstrap test generated by our procedure is fully robust to weak instrument in the sense that its limiting null rejection probability is no greater than the nominal level even if all clusters are "weak". However, such robustness is not shared by its studentized version; the wild bootstrap test that is based on the t-test statistic can have serious size distortion in this case. Furthermore, in the general case with multiple instruments, we show that an unstudentized version of bootstrap Anderson-Rubin (AR) test is fully robust to weak instruments, and is superior with regard to both size and power properties to alternative asymptotic and bootstrap AR tests that employ cluster-robust variance estimators. By contrast, we find that bootstrapping other weak-instrument-robust tests such as the Lagrange multiplier test and the conditional quasi-likelihood ratio test, no matter studentized or unstudentized, does not guarantee correct limiting null rejection probability when all clusters are "weak".

Keywords: Weak Instrument, Wild Bootstrap, Clustered Data, Randomization Test.

[*] Division of Economics, School of Social Sciences, Nanyang Technological University. HSS-04-65, 14 Nanyang Drive, 637332, Singapore. E-mail address: wang.wj@ntu.edu.sg.

1 Introduction

It is well known that in instrumental variables (IV) regressions, if the correlation between instruments and endogenous regressors is small, IV estimators such as two-stage least squares (TSLS) can be badly biased, and Wald-type t-tests can have serious size distortion and the coverage probability of conventional IV confidence intervals may be far lower than intended. Various recent surveys on papers published in leading economic journals suggest that these issues remain important concerns for empirical practice. For instance, Andrews, Stock, and Sun (2019) survey a sample of 230 IV regressions from 17 papers published in the American Economic Review (AER) from 2014 to 2018. They find that many of the first-stage F-statistics (and their nonhomoskedastic generalizations) in these papers are in a range that raise the concerns of weak instruments, and virtually all these papers reported at least one first-stage F with value smaller than 10. Brodeur, Cook, and Heyes (2020) investigate over 21,000 hypothesis tests published in 25 leading economic journals, and find that the extent of p-hacking and publication bias varies greatly by empirical methods such as randomized control trial, difference-in-differences, regression discontinuity design, and IV regressions. The authors highlight that IV regressions are particularly problematic and a sizable over-representation of first-stage F is documented just over the threshold of 10 (such pattern is also observed in Andrews et al. (2019)). They also find that the degree of p-hacking in the second stage is related to instrument strength in the first stage: IV regressions with relatively weak instruments have a much higher proportion of second-stage t-statistics being barely significant around 1.65 and 1.96. Furthermore, Young (2020) analyzes a sample of 1359 IV regressions in 31 papers published in the American Economic Association (AEA), and highlights that heteroskedastic errors and clustered data can significantly damage the quality of inference, so that normal approximations become rather unreliable. To address these issues, Young (2020) suggests applying (cluster-robust) bootstrap to IV estimates and Wald-type t statistics.

Although there are numerous evidences suggesting that appropriately designed bootstrap procedures can substantially improve the quality of inference for IV estimates and Wald-type t-tests (e.g., see also Davidson and MacKinnon (2008, 2010, 2014), Wang and Kaffo (2016), Finlay and Magnusson (2019)), it is well known that such bootstrap procedures are generally invalid under weak instruments; e.g., see the discussions in Section 3.1 and p.750 of Andrews et al. (2019). On the other hand, the econometric literature has developed various weak-instrument-robust tests and confidence sets, and bootstrap for such test statistics may remain valid regardless of instrument strength. Using the robust statistics may also help to alleviate the aforementioned problem of screening on first-stage F (by either researchers or journals), which can dramatically increase bias in published estimates and size distortion in published tests (e.g., see Andrews et al. (2019), Section 4.1).[1] In the case of homoskedastic errors, Moreira, Porter, and Suarez (2009) show validity of bootstrapped Lagrange multiplier (LM; Kleibergen, 2002) and Anderson-Rubin (1949, AR) tests under weak instruments. It is possible to extend their result of bootstrap validity to the case with heteroskedasticity and clustered data, under an asymptotic framework where the number of clusters goes to infinity. However, as emphasized in Ibragimov and Müeller (2010, 2016), Bester, Conley, and Hansen (2011), Cameron and Miller (2015), Canay, Romano, and Shaikh (2017), Canay, Santos, and Shaikh (2020) and Young (2020), many empirical studies motivate the consideration of an alternative framework in which the number of clusters is small, while the number of observations in each cluster is relatively large. In such case with few clusters, a fundamentally different framework is required to study the properties of bootstrap procedures for IV regressions. In particular, the bootstrap distribution can no longer consistently estimate the distribution of the statistics of interest, and it is thus not obvious what conditions are required to achieve bootstrap validity.

In this paper, we consider a linear IV model allowing for cluster heterogeneity in the strength of instruments; i.e., we allow for the case that the instruments may be strong for some clusters while weak for others. This setting is motivated by Young (2020)'s finding in his AEA samples that with the removal of just one cluster/observation, in the average paper 49% of reported 0.01 significant TSLS results can be rendered insignificant at that level and the first-stage F-statistics are also very sensitive to outlier clusters/observations. In terms of methodology, we exploit the connection between the wild cluster bootstrap with Rademacher weights and a

[1] See also Andrews (2018), who proposed a two-step procedure for GMM with controlled coverage distortions that is based on combining Wald-type and weak-identification-robust confidence sets. In addition, Andrews et al. (2019, Section 5.4) find that for the IV model with single endogenous regressor, a two-step procedure based on the effective F-statistic of Olea and Pflueger (2013), which uses a t-test if the effective F is larger than 10 and uses an Anderson-Rubin test otherwise, has at most mild size distortions in simulations calibrated to their AER data.

randomization test based on the group of sign changes in a framework in which the number of clusters is fixed, following the seminal study by Canay et al. (2020). First, under the condition that the available instruments are strong for at least one cluster, we establish the asymptotic validity results of the unstudentized and studentized wild bootstrap tests (i.e., percentile and percentile-t) for IV regressions similar to those obtained in Canay et al. (2020) for ordinary least squares. In particular, we notice that although having remarkable resemblance, the wild cluster bootstrap for IV regressions can have properties very different from the Fama-Macbeth type approach in Ibragimov and Müeller (2010, 2016) and the randomization test with sign changes in Canay, Romano, and Shaikh (2017), both of which are based on cluster-level estimates and would require strong instruments for all clusters to achieve validity in the current context. In this sense, the wild bootstrap tests are more robust to cluster-level heterogeneity/outlier in terms of instrument strength.

Second, we find that for the leading case in empirical applications of testing the value of the coefficient of single endogenous regressor with single instrument (e.g., 101 out of 230 specifications in Andrews and al. (2019) and 1087 out of 1359 in Young (2020)), the unstudentized wild bootstrap test generated by our particular procedure is fully robust to weak instrument in the sense that its null limiting rejection probability is no greater than the nominal level even when all clusters are "weak", while such robustness is not shared by its studentized version or bootstrap tests generated by alternative procedures such as the commonly employed pairs cluster bootstrap. Therefore, although in the standard strong-instrument case with a large number clusters, the studentized bootstrap test may achieve a higher order refinement as it is based on an asymptotically pivotal statistic, from the viewpoint of robustness, it could be more desirable to use the unstudentized bootstrap test with few clusters and single instrument.

Third, we find that in the general case with multiple instruments, an unstudentized version of the wild bootstrap AR test is valid irrespective of instrument strength, and its studentized version may only over-reject the null hypothesis by a small quantity that decreases exponentially with the number of clusters. In terms of size properties under a small number of clusters, we find that the wild bootstrap AR tests have substantial improvement, especially in the over-identified case, upon two alternative AR tests that are based on (null-imposed) cluster-robust variance estimators and conventional asymptotic critical values, one of which under-rejects or does not reject at all while the other can seriously over-reject. In addition, our simulation results suggest that in the over-identified case, the unstudentized bootstrap AR test typically has better power properties than its studentized version.

Furthermore, with regard to weak-instrument-robust tests other than the AR test, we are only able to establish the validity result for bootstrapping the LM and conditional quasi-likelihood ratio (CQLR) test when the instruments are strong for at least one cluster. This is because the validity of LM and CQLR tests (and various other robust statistics proposed in the literature) depends crucially on the asymptotic independence between sample moment and orthogonalized sample Jacobian. Such independence property holds under the standard framework where the number of observations/clusters is allowed to tend to infinity but no longer holds with a fixed number of clusters. In the presence of strong instruments for at least one cluster, we are still able to establish the connection between the wild bootstrap and randomization test even without such asymptotic independence, while their connection cannot be established if the instruments are weak for all clusters. Therefore, in the just-identified case bootstrapping these test statistics is valid, irrespective of instrument strength, as they are equivalent to the AR test in this case, while they could have large size distortions in the over-identified case, as illustrated in our simulation results.

A variety of weak-instrument-robust methods have been developed in the literature. For the case with homoskedastic errors, Kleibergen (2002) provides the LM test and Moreira (2003) proposes a conditional likelihood ratio (CLR) test. For subvector inference, Guggenberger, Kleibergen, Mavroeidis, and Chen (2012), Guggenberger, Kleibergen, and Mavroeidis (2019), and Wang and Doko Tchatoka (2018) propose AR-based methods. For the general case with non-homoskedastic errors, Kleibergen (2005) introduces LM and CQLR tests. Andrews (2016) introduces conditional linear combination tests, which are based on a data-dependent convex combination of the AR and LM statistics. Andrews and Mikusheva (2016) and Moreira and Moreira (2019) introduce a direct generalization of the CLR test. Andrews and Guggenberger (2019) introduce two alternative CQLR tests, which allow the variance matrix of the moments to be near singular or singular. However, the literature on the properties of the weak-instrument-robust tests with clustered data remains sparse, especially for the case with few clusters.

There is also a growing econometric literature studying the properties of wild bootstrap for clustered data, among them Cameron, Gelbach, and Miller (2008), MacKinnon and Webb (2017), Djogbenou, MacKinnon, and Nielsen (2019), MacKinnon, Nielsen, and Webb (2019), Roodman, Nielsen, MacKinnon, and Webb (2019), etc. Furthermore, the literature on bootstrap for the IV model includes Davidson and MacKinnon (2008, 2010, 2014), Moreira et al. (2004, 2009), Wang and Kaffo (2016), Kaffo and Wang (2017), Finlay and Magnusson (2019), among others. In particular, under the setting of homoskedastic errors, Moreira et al. (2004, 2009) show the bootstrap validity of AR, LM and CLR tests even under weak instruments. Wang and Kaffo (2016) show bootstrap inconsistency for estimating the distribution of IV estimators under the many/many weak instrument sequences of Bekker (1994) and Chao and Swanson (2005), and propose valid modified bootstrap procedure, which significantly improves upon asymptotic normal approximation. Davidson and MacKinnon (2010) and Finlay and Magnusson (2019) document through extensive simulations that a variety of wild bootstrap procedures have much better finite sample performance than asymptotic methods with heteroskedastic errors and clustered data, respectively.

The remainder of this paper is organized as follows. Section 2 presents the setting, test statistics and assumptions. Section 3 presents the main results for the bootstrap tests with few clusters. Section 4 investigates the finite sample size and power properties of the bootstrap tests and alternative methods using simulations. Conclusions are drawn in Section 5.

2 Setup and assumptions

We consider a setup with clustered data, where the clusters are indexed by $j \in J \equiv \{1, ..., q\}$ and units in the j-th cluster are indexed by $i \in I_{n,j} \equiv \{1, ..., n_j\}$. Our linear IV model can be written as

$$\begin{aligned} y_{i,j} &= X_{i,j}'\beta + W_{i,j}'\gamma + \epsilon_{i,j}, \\ X_{i,j} &= Z_{i,j}'\Pi_{z,j} + W_{i,j}'\Pi_w + v_{i,j}, \end{aligned} \tag{1}$$

where $y_{i,j} \in \mathbf{R}$ denotes an outcome of interest, while $X_{i,j} \in \mathbf{R}^{d_x}$, $W_{i,j} \in \mathbf{R}^{d_w}$, and $Z_{i,j} \in \mathbf{R}^{d_z}$ denote endogenous regressors, exogenous regressors, and instrumental variables, respectively. For example, $X_{i,j}$ may be certain treatment intervention or policy change that is endogenous in the sense that $X_{i,j}$ is correlated with the error $\epsilon_{i,j}$, and $W_{i,j}$ include exogenous control variables such as unit-level characteristics or cluster-level fixed effects. $\beta \in \mathbf{R}^{d_x}$ and $\gamma \in \mathbf{R}^{d_w}$ are unknown parameters of the structural form equation, while $\Pi_{z,j} \in \mathbf{R}^{d_z \times d_x}$ and $\Pi_w \in \mathbf{R}^{d_z \times d_w}$ are unknown parameters of the first-stage equation.

We also allow for the existence of cluster heterogeneity with regard to instrument strength in (1), by letting the first-stage coefficient $\Pi_{z,j}$ to vary across clusters. This setting is motivated by the fact that in empirical studies instruments often turn out to be strong for some subgroups and weak for some other subgroups, which can be determined by various factors such as ethnic groups and geographic regions (see Abadie, Gu, and Shen (2019) and the references therein). In experimental economics with clustered randomized trials, subjects' compliance with treatment assignment may also have substantial variations among clusters. For example, in Muralidharan, Niehaus, and Sukhtankar (2016)'s evaluation of a smartcard payment system, their random assignment was implemented at village level, and in some villages, 90% or more of the recipients complied with the treatment, while in many villages less than 10% complied. Furthermore, the setting is motivated by Young (2020)'s finding (e.g., see Figures I and II in his paper) that with the removal of just one cluster/observation in the average paper of his AEA samples, 49% of reported 0.01 significant TSLS results can be rendered insignificant at that level and the first-stage F-statistics are also very sensitive to outliers, e.g., the average paper F can be lowered to 72% of its original value with the removal of one cluster/observation.

Now we introduce the test statistics considered in the paper. The first set of test statistics are the ones based on the IV estimates and the standard Wald-type t-statistic with cluster-robust variance estimator. Specifically, for testing the null hypothesis

$$H_0^c : c'\beta = \lambda \quad \text{vs.} \quad H_1^c : c'\beta \neq \lambda, \tag{2}$$

where $c \in \mathbf{R}^{d_x}$ and $\lambda \in \mathbf{R}$, we consider the unstudentized test statistic

$$W_{U,n}(\lambda) \equiv |\sqrt{n}(c'\hat{\beta}_n - \lambda)|, \tag{3}$$

and the studentized test statistic

$$W_n(\lambda) \quad \equiv \quad \frac{|\sqrt{n}(c'\hat{\beta}_n - \lambda)|}{\sqrt{c'\widehat{V}_n(\hat{\beta}_n)c}}, \tag{4}$$

where

$$\widehat{V}_n(\hat{\beta}_n) \equiv \left(\widehat{Q}'_{\tilde{Z}X,n}\widehat{Q}^{-1}_{\tilde{Z}\tilde{Z},n}\widehat{Q}_{\tilde{Z}X,n}\right)^{-1}\widehat{Q}'_{\tilde{Z}X,n}\widehat{Q}^{-1}_{\tilde{Z}\tilde{Z},n}\widehat{\Omega}_n(\hat{\beta}_n)\widehat{Q}^{-1}_{\tilde{Z}\tilde{Z},n}\widehat{Q}_{\tilde{Z}X,n}\left(\widehat{Q}'_{\tilde{Z}X,n}\widehat{Q}^{-1}_{\tilde{Z}\tilde{Z},n}\widehat{Q}_{\tilde{Z}X,n}\right)^{-1},$$

$\widehat{\Omega}_n(\hat{\beta}_n) = n^{-1}\sum_{j\in J}\sum_{i\in I_{n,j}}\sum_{k\in I_{n,j}}\tilde{Z}_{i,j}\tilde{Z}'_{k,j}\hat{\epsilon}_{i,j}\hat{\epsilon}_{k,j}$, $\widehat{Q}_{\tilde{Z}\tilde{Z},n} = n^{-1}\sum_{j\in J}\sum_{i\in I_{n,j}}\tilde{Z}_{i,j}\tilde{Z}'_{i,j}$, $\widehat{Q}_{\tilde{Z}X,n} = n^{-1}\sum_{j\in J}\sum_{i\in I_{n,j}}\tilde{Z}_{i,j}X'_{i,j}$, $\hat{\epsilon}_{i,j} = y_{i,j} - X'_{i,j}\hat{\beta}_n - W'_{i,j}\hat{\gamma}_n$, $\hat{\beta}_n$ and $\hat{\gamma}_n$ are the TSLS estimators of β and γ in (1), and $\tilde{Z}_{i,j}$ is the residuals from regressing $Z_{i,j}$ on $W_{i,j}$ using full sample, i.e.,

$$\tilde{Z}_{i,j} \equiv Z_{i,j} - \widehat{\Gamma}'_n W_{i,j}, \tag{5}$$

where $\widehat{\Gamma}_n$, a $d_w \times d_z$-dimensional matrix, denotes the coefficients obtained from the regression of $Z_{i,j}$ on $W_{i,j}$ and satisfies the orthogonality conditions $\sum_{j\in J}\sum_{i\in I_{n,j}}\left(Z_{i,j} - \widehat{\Gamma}'_n W_{i,j}\right)W'_{i,j} = 0$.

It is well known that the conventional Wald-type t-test and confidence intervals can have serious distortion under weak instruments, thus we also consider the weak-instrument-robust test statistics. Following the econometric literature on weak instruments, for testing the joint null hypothesis

$$H_0 : \beta = \beta_0 \quad \text{vs.} \quad H_1 : \beta \neq \beta_0, \tag{6}$$

we define the AR statistic (with null-imposed cluster-robust variance estimator) as

$$AR_n(\beta_0) \equiv n\widehat{f}_n(\beta_0)'\widehat{\Omega}_n^{-1}(\beta_0)\widehat{f}_n(\beta_0), \tag{7}$$

with the sample moments and the estimator of their variance matrix denoted as

$$\begin{aligned} \widehat{f}_n(\beta) &\equiv n^{-1}\sum_{j\in J}\sum_{i\in I_{n,j}} f_{i,j}(\beta), \\ \widehat{\Omega}_n(\beta) &\equiv n^{-1}\sum_{j\in J}\sum_{i\in I_{n,j}}\sum_{k\in I_{n,j}} f_{i,j}(\beta)f_{k,j}(\beta)', \end{aligned} \tag{8}$$

where $f_{i,j}(\beta) = \tilde{Z}_{i,j}\left(y_{i,j} - X'_{i,j}\beta - W_{i,j}\bar{\gamma}^r_n\right)$, and $\bar{\gamma}^r_n$ is the null-restricted least squares estimator of γ, i.e., $\bar{\gamma}^r_n = \left(\sum_{j\in J}\sum_{i\in I_{n,j}} W_{i,j}W'_{i,j}\right)^{-1}\sum_{j\in J}\sum_{i\in I_{n,j}} W_{i,j}(y_{i,j} - X'_{i,j}\beta_0)$. The asymptotic critical value of the AR test rejects $H_0 : \beta = \beta_0$ if $AR_n(\beta_0) > \chi^2_{d_z,1-\alpha}$, where $\chi^2_{d_z,1-\alpha}$ is the $1-\alpha$ quantile of the chi-square distribution with d_z degree of freedom. We also consider an unstudentized version of the AR statistic, which take the form

$$AR_{U,n}(\beta_0) \equiv \left\|\sqrt{n}\widehat{f}_n(\beta_0)\right\|^2. \tag{9}$$

Another form of AR statistic widely applied in the literature (see, e.g., Chernozhukov and

Hansen (2008a, 2008b), Finlay and Magnusson (2009), Andrews et al. (2019), Roodman et al. (2019)) is based on the reduced form of the model in (1), which can be written as (under homogeneity in instrument strength, i.e., Π_z being the same for all clusters)

$$y_{i,j} - X_{i,j}'\beta_0 = Z_{i,j}'\delta + W_{i,j}'\theta + u_{i,j}. \tag{10}$$

where $\delta = \Pi_z(\beta - \beta_0)$, $\theta = \Pi_w(\beta - \beta_0) + \gamma$, and $u_{i,j} = v_{i,j}'(\beta - \beta_0) + \epsilon_{i,j}$. Notice that in this case testing $\beta = \beta_0$ is equivalent to testing $\delta = 0$, and this leads to a Wald-type AR statistic:

$$\begin{aligned} AR_{W,n}(\beta_0) &\equiv n\hat{\delta}_n'(\beta_0)\widehat{V}_{W,n}^{-1}(\beta_0)\hat{\delta}_n(\beta_0), \\ \widehat{V}_{W,n}(\beta_0) &\equiv \widehat{Q}_{\tilde{Z}\tilde{Z},n}^{-1}\widehat{\Omega}_{W,n}(\beta_0)\widehat{Q}_{\tilde{Z}\tilde{Z},n}^{-1}, \end{aligned} \tag{11}$$

where $\widehat{\Omega}_{W,n}(\beta_0) = n^{-1}\sum_{j\in J}\sum_{i\in I_{n,j}}\sum_{k\in I_{n,j}}\tilde{Z}_{i,j}\tilde{Z}_{k,j}'\hat{u}_{i,j}(\beta_0)\hat{u}_{k,j}(\beta_0)$, with $\hat{\delta}_n(\beta_0)$ and $\hat{u}_{i,j}(\beta_0)$ being the least squares estimator and residual of regressing $y_i - X_i'\beta_0$ on $Z_{i,j}$ and $W_{i,j}$, respectively. Different from (7), the procedure in (11) only requires conventional least squares-based estimation and cluster-robust inference, and uses the same critical values as $AR_n(\beta_0)$. We include the three forms of the AR statistics in the paper as they can have very different properties in the case with small number of clusters.

To introduce the other weak-instrument-robust statistics, we define the sample Jacobian as

$$\begin{aligned} \widehat{G}_n &\equiv \left(\widehat{G}_{1,n}, ..., \widehat{G}_{d_x,n}\right) \in \mathbf{R}^{d_z \times d_x}, \\ \widehat{G}_{l,n} &\equiv n^{-1}\sum_{j\in J}\sum_{i\in I_{n,j}}\tilde{Z}_{i,j}X_{i,j,l}, \text{ for } l = 1, ..., d_x, \end{aligned} \tag{12}$$

and define the orthogonalized sample Jacobian as

$$\begin{aligned} \widehat{D}_n(\beta) &\equiv \left(\widehat{D}_{1,n}(\beta), ..., \widehat{D}_{d_x,n}(\beta)\right) \in \mathbf{R}^{d_z \times d_x}, \text{ where} \\ \widehat{D}_{l,n}(\beta) &\equiv \widehat{G}_{l,n} - \widehat{\Gamma}_{l,n}(\beta)\widehat{\Omega}_n^{-1}(\beta)\widehat{f}_n(\beta) \in \mathbf{R}^{d_z} \text{ for } l = 1, ..., d_x, \\ \widehat{\Gamma}_{l,n}(\beta) &\equiv n^{-1}\sum_{j\in J}\sum_{i\in I_{n,j}}\sum_{k\in I_{n,j}}\left(\tilde{Z}_{i,j}\hat{v}_{i,j,l}\right)f_{k,j}(\beta)', \text{ for } l = 1, ..., d_x, \end{aligned} \tag{13}$$

where $\hat{v}_{i,j,l}$ is the residual of regressing $X_{i,j,l}$ on $Z_{i,j}$ and $W_{i,j}$. Therefore, under the null hypothesis in (6) and the standard asymptotic framework where the number of clusters tends to infinity, $\widehat{D}_n(\beta)$ equals the sample Jacobian matrix $\widehat{G}_n(\beta)$ adjusted to be asymptotically independent of the sample moments $\widehat{f}_n(\beta)$.

Then, the cluster-robust version of Kleibergen (2002, 2005)'s LM statistic is defined as

$$LM_n(\beta_0) \equiv n\widehat{f}_n(\beta_0)'\widehat{\Omega}_n^{-1/2}(\beta_0)P_{\widehat{\Omega}_n^{-1/2}(\beta_0)\widehat{D}_n(\beta_0)}\widehat{\Omega}_n^{-1/2}(\beta_0)\widehat{f}_n(\beta_0), \tag{14}$$

where $P_A = A(A'A)^{-}A'$ for any matrix A and $(\cdot)^{-}$ denotes any generalized inverse. The nominal size α asymptotic LM test rejects the null hypothesis when $LM_n(\beta_0) > \chi^2_{d_x,1-\alpha}$, where $\chi^2_{d_x,1-\alpha}$ is the $1-\alpha$ quantile of the chi-square distribution with d_x degree of freedom.

In addition, the CQLR statistic in Kleibergen (2005, 2007), Smith (2007), Newey and Windmeijer (2009), and Guggenberger, Ramalho, and Smith (2012) are adapted from Moreria (2003)'s CLR test, and its cluster-robust version takes the form

$$LR_n(\beta_0) \equiv \frac{1}{2}\left(AR_n(\beta_0) - rk_n(\beta_0) + \sqrt{(AR_n(\beta_0) - rk_n(\beta_0))^2 + 4LM_n(\beta_0)\cdot rk_n(\beta_0)}\right), \tag{15}$$

where $rk_n(\beta_0)$ is a conditioning statistic and the critical value of the CQLR test depends on $rk_n(\beta_0)$. Here, following Newey and Windmeijer (2009) and Guggenberger et al. (2012)[2], we let $rk_n(\beta) = n\widehat{D}_n'(\beta)\widehat{\Omega}_n^{-1}(\beta)\widehat{D}_n(\beta)$. The (conditional) asymptotic critical value of the CQLR test is $c(1-\alpha, rk_n(\beta))$, where $c(1-\alpha, r)$ is the $1-\alpha$ quantile of the distribution of $\frac{1}{2}\left(\chi^2_{d_x} + \chi^2_{d_z-d_x} - r + \sqrt{\left(\chi^2_{d_x} + \chi^2_{d_z-d_x} - r\right)^2 + 4\chi^2_{d_x}r}\right)$.

Similar to the bootstrap AR tests, we also study bootstrapping the unstudentized version of LM and CQRL statistics, i.e.,

$$\begin{aligned} LM_{U,n}(\beta_0) &\equiv \left\|\sqrt{n}\widehat{D}_n'(\beta_0)\widehat{\Omega}_n^{-1/2}(\beta_0)\widehat{f}_n(\beta_0)\right\|^2, \\ LR_{U,n}(\beta_0) &\equiv \frac{1}{2}\left(AR_{U,n}(\beta_0) - rk_n(\beta_0) + \sqrt{(AR_{U,n}(\beta_0) - rk_n(\beta_0))^2 + 4LM_{U,n}(\beta_0)\cdot rk_n(\beta_0)}\right). \end{aligned} \tag{16}$$

We next introduce the assumptions that will be used in our analysis of the asymptotic properties of the bootstrap tests under a small number of clusters.

Assumption 1 *The following statements hold:*

[2] Kleibergen (2005) uses alternative formula for $rk_n(\beta)$, and Andrews and Guggenberger (2019) introduce alternative CQLR test statistic. We can show similar result for these alternative CQLR tests under the framework with few clusters.

(i) The quantity

$$\frac{1}{\sqrt{n}}\sum_{j\in J}\sum_{i\in I_{n,j}}\begin{pmatrix}Z_{i,j}\epsilon_{i,j}\\ W_{i,j}\epsilon_{i,j}\end{pmatrix}$$

converges in distribution.

(ii) The quantities

$$\frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}}\begin{pmatrix}Z_{i,j}Z'_{i,j} & Z_{i,j}W'_{i,j}\\ W_{i,j}Z'_{i,j} & W_{i,j}W'_{i,j}\end{pmatrix}$$

and

$$\frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}}\begin{pmatrix}Z_{i,j}X'_{i,j}\\ W_{i,j}X'_{i,j}\end{pmatrix}$$

converges in probability to a positive-definite matrix and a full rank matrix, respectively.

Assumption 1 requires that the within-cluster dependence is weak enough to allow for the application of suitable law of large numbers and central limit theorems, and it ensures that the two-stage least squares estimators $\hat{\beta}_n$ and $\hat{\gamma}_n$ are well behaved. Assumption 1 also ensures that the restricted estimators $\hat{\beta}^r_n$ and $\hat{\gamma}^r_n$ are well behaved under H^c_0.

Assumption 2 *The following statements hold:*

(i) There exists a collection of independent random variables $\{\mathcal{Z}_j : j \in J\}$, where $\mathcal{Z}_j \equiv [\mathcal{Z}_{\epsilon,j} : \mathcal{Z}_{v,j}]$ with $\mathcal{Z}_{\epsilon,j} \in \boldsymbol{R}^{d_z}$ and $\mathcal{Z}_{v,j} \in \boldsymbol{R}^{d_z \times d_x}$, and $vec(\mathcal{Z}_j) \sim N(0, \Sigma_j)$ with Σ_j positive definite for all $j \in J$, such that

$$\left\{\left(\frac{1}{\sqrt{n_j}}\sum_{i\in I_{n,j}}\tilde{Z}_{i,j}\epsilon_{i,j},\ \frac{1}{\sqrt{n_j}}\sum_{i\in I_{n,j}}\tilde{Z}_{i,j}v'_{i,j}\right) : j\in J\right\} \xrightarrow{d} \{\mathcal{Z}_j : j \in J\}.$$

(ii) For each $j \in J$, $n_j/n \to \xi_j > 0$.

(iii) For each $j \in J$,

$$\frac{1}{n_j}\sum_{i\in I_{n,j}}\left\|W'_{i,j}\left(\widehat{\Gamma}_n - \widehat{\Gamma}^c_{n,j}\right)\right\|^2 \xrightarrow{P} 0,$$

where $\widehat{\Gamma}_n$ and $\widehat{\Gamma}^c_{n,j}$ denotes the coefficient from linearly regressing $Z_{i,j}$ on $W_{i,j}$ by using the entire sample and by only using the sample in the j-th cluster, respectively.

The assumptions are similar to those imposed in Canay et al. (2020). Assumption 2(i) is satisfied whenever the within-cluster dependence is sufficiently weak to permit applicaiton of a suitable central limit theorem and the data are independent across clusters. The assumption that $\mathcal{Z}_j$ have full rank covariance matrices requires that the instruments $Z_{i,j}$ can not be expressed as a linear combination of the exogenous regressrors $W_{i,j}$ within each cluster j. Assumption 2(ii) gives the restriction on relative sizes of the clusters. Assumption 2(iii) gives the condition on cluster homogeneity. As pointed out by Canay et al. (2020), this assumption is satisfied whenever the distributions of $(Z'_{i,j}, W'_{i,j})'$ are the same across clusters. Furthermore, in the case that $W_{i,j}$ includes only cluster-level fixed effects, then the assumption is immediately satisfied. It is also clear from the definition of $\widehat{\Pi}^c_{n,j}$ that it satisfies the cluster-level orthogonality condition; i.e., $\sum_{i\in I_{n,j}} \left(Z_{i,j} - \widehat{\Pi}^{c'}_{n,j} W_{i,j}\right) W'_{i,j} = 0$, for each $j \in J$.

The following assumption is with regard to the instrument strength, with Assumption 3(i) being stronger than Assumption 3(ii).

Assumption 3 *(i) There exists nonempty $J_s \subseteq J$ such that for each $j \in J_s$,*

$$\frac{1}{n_j} \sum_{i\in I_{n,j}} \tilde{Z}_{i,j} X_{i,j} \xrightarrow{P} Q_{\tilde{Z}X,j},$$

where $Q_{\tilde{Z}X,j}$ is a full rank matrix.
(ii) There exists nonempty $J_s \subseteq J$ such that for each $j \in J_s$,

$$\frac{1}{n_j} \sum_{i\in I_{n,j}} \tilde{Z}_{i,j} X_{i,j} \xrightarrow{P} a_j Q_{\tilde{Z}X},$$

where $a_j > 0$ and $Q_{\tilde{Z}X}$ is a full rank matrix.

Assumption 3(i) requires that the instruments are strong at least for one cluster, while Assumption 3(ii) further requires that the limits of the cluster-level sample Jacobian matrices $\sum_{i\in I_{n,j}} \tilde{Z}_{i,j} X_{i,j}/n_j$ are proportional to each other for these “strong” clusters. The bootstrap validity under few clusters requires different assumptions in terms of instrument strength, depending on the test statistics, hypothesis of interest, and specific application. In particular, Assumption 3(i) is needed for the bootstrap validity of testing $H_0 : \beta = \beta_0$ with the LM and CQLR tests. By contrast, the bootstrapped AR test does not require this assumption as it is fully robust to weak instruments even under few clusters. On the other hand, Assumption 3(ii) is needed for the bootstrap validity of testing the more general hypothesis $H_0^c : c'\beta = \lambda$ with the IV estimate and t-test in (3)-(4). However, we also notice that this assumption is not required for the bootstrapped IV estimate for testing $H_0 : \beta = \beta_0$ in the case with single instrument (i.e., testing the coefficient of single endogenous regressor with single instrument), as it is fully weak-instrument robust in this case (see Remark 3 in Section 3.1).

3 Main results

3.1 Wild bootstrap with IV estimate and t-statistic

In this section, we study the properties of the bootstrapped tests under the asymptotic framework where the number of clusters is kept fixed. The bootstrapped tests for $H_0^c : c'\beta = \lambda$ with the t-statistic and its unstudentized version are implemented through the following procedure:

1. Compute the null-restricted residual

$$\hat{\epsilon}^r_{i,j}(\lambda) = y_{i,j} - X'_{i,j}\hat{\beta}^r_n(\lambda) - W'_{i,j}\hat{\gamma}^r_n(\lambda), \tag{17}$$

   where $\hat{\beta}^r_n(\lambda)$ and $\hat{\gamma}^r_n(\lambda)$ are H_0^c-restricted two-stage least squares estimators of β and γ.

2. Let $\mathbf{G} = \{-1, 1\}^q$ and for any $g = (g_1, ..., g_q) \in \mathbf{G}$ generate

$$y^*_{i,j}(g) = X'_{i,j}\hat{\beta}^r_n(\lambda) + W'_{i,j}\hat{\gamma}^r_n(\lambda) + g_j\hat{\epsilon}^r_{i,j}(\lambda). \tag{18}$$

3. For each $g = (g_1, ..., g_q) \in \mathbf{G}$ compute $\hat{\beta}^*_n(g)$ and $\hat{\gamma}^*_n(g)$, the analogues of the two-stage least squares estimators $\hat{\beta}_n$ and $\hat{\gamma}_n$ using $y^*_{i,j}(g)$ in place of $y_{i,j}$ and the same $(Z'_{i,j}, X'_{i,j}, W'_{i,j})'$. For the bootstrapped t-statistic, also compute

$$\hat{\epsilon}^*_{i,j}(g) = y^*_{i,j}(g) - X'_{i,j}\hat{\beta}^*_n(g) - W'_{i,j}\hat{\gamma}^*_n(g). \tag{19}$$

4. Compute the bootstrap analogues of test statistics:

$$\begin{aligned} W^*_{U,n}(\lambda, g) &= |\sqrt{n}(c'\hat{\beta}^*_n(g) - \lambda)|, \\ W^*_n(\lambda, g) &= W^*_{U,n}(\lambda, g)/\sqrt{c'\widehat{V}^*_n(\hat{\beta}^*_n(g))c}, \end{aligned} \tag{20}$$

   where $\widehat{V}^*_n(\hat{\beta}^*_n(g)) = \left(\widehat{Q}'_{\tilde{Z}X,n}\widehat{Q}^{-1}_{\tilde{Z}\tilde{Z},n}\widehat{Q}_{\tilde{Z}X,n}\right)^{-1}\widehat{Q}'_{\tilde{Z}X,n}\widehat{Q}^{-1}_{\tilde{Z}\tilde{Z},n}\widehat{\Omega}^*_n(\hat{\beta}^*_n(g))\widehat{Q}^{-1}_{\tilde{Z}\tilde{Z},n}\widehat{Q}_{\tilde{Z}X,n}\left(\widehat{Q}'_{\tilde{Z}X,n}\widehat{Q}^{-1}_{\tilde{Z}\tilde{Z},n}\widehat{Q}_{\tilde{Z}X,n}\right)^{-1}$, and $\widehat{\Omega}^*_n(\hat{\beta}^*_n(g)) = n^{-1}\sum_{j\in J}\sum_{i\in I_{n,j}}\sum_{k\in I_{n,j}}\tilde{Z}_{i,j}\tilde{Z}'_{k,j}\hat{\epsilon}^*_{i,j}(g)\hat{\epsilon}^*_{k,j}(g)$.

5. To obtain the critical value for the bootstrapped t-test, we compute the $1 - \alpha$ quantile of

$\{W_n^*(\lambda, g) : g \in \mathbf{G}\}$:

$$\hat{c}_n^w(1-\alpha) \equiv \inf\left\{u \in \mathbf{R} : \frac{1}{|\mathbf{G}|}\sum_{g \in \mathbf{G}} I\{W_n^*(\lambda, g) \leq u\} \geq 1-\alpha\right\}, \tag{21}$$

where $I\{A\}$ equals one whenever the event A is true and equals zero otherwise. $\phi_n\left(W_n(\lambda)\right)$, the bootstrapped t-test for $H_0^c : c'\beta = \lambda$, rejects whenever $W_n(\lambda)$ exceeds its critical value:

$$\phi_n\left(W_n(\lambda)\right) \equiv I\{W_n(\lambda) > \hat{c}_n^w(1-\alpha)\}. \tag{22}$$

The bootstrapped test with $W_{U,n}(\lambda)$ is defined in the same fashion.

Notice that the above procedure takes the form of randomization inference with a group of sign change. Canay et al. (2020) point out the important connection between wild cluster bootstrap and randomization inference; e.g., the critical values defined in (21) may also be written as

$$\inf\left\{u \in \mathbf{R} : P\left\{W_n^*(\lambda, \omega) \leq u \middle| \left(y^{(n)}, X^{(n)}, Z^{(n)}, W^{(n)}\right)\right\} \geq 1-\alpha\right\}, \tag{23}$$

where $(y^{(n)}, X^{(n)}, Z^{(n)}, W^{(n)})$ denotes the full sample of observed data and ω is uniformly distributed on $\mathbf{G}$ independently of the observed data. As remarked by Canay et al. (2020), this way of writing the critical values coincides with the existing literature on the wild cluster bootstrap that sets $\omega = (\omega_1, ..., \omega_q)$ to be i.i.d. Rademacher random variables, which equals ±1 with equal probability.

The following theorem gives the properties of the bootstrapped test based on the IV estimates and t-statistic in the case with a small number of clusters.

Theorem 3.1 *If Assumptions 1-2, Assumption 3(ii), and $H_0^c : c'\beta = \lambda$ holds, then*

$$\alpha - \frac{1}{2^{q-1}} \leq \liminf_{n\to\infty} P\{W_{U,n}(\lambda) > \hat{c}_{u,n}^w(1-\alpha)\} \leq \limsup_{n\to\infty} P\{W_{U,n}(\lambda) > \hat{c}_{u,n}^w(1-\alpha)\} \leq \alpha,$$

and

$$\alpha - \frac{1}{2^{q-1}} \leq \liminf_{n\to\infty} P\{W_n(\lambda) > \hat{c}_n^w(1-\alpha)\} \leq \limsup_{n\to\infty} P\{W_n(\lambda) > \hat{c}_n^w(1-\alpha)\} \leq \alpha + \frac{1}{2^{q-1}},$$

where $\hat{c}_{u,n}^w(1-\alpha)$ and $\hat{c}_n^w(1-\alpha)$ denote the critical values of the $W_{U,n}(\lambda)$ and $W_n(\lambda)$-based bootstrap tests, respectively.

Theorem 3.1 states that as long as there exists at least one "strong" cluster, the bootstrap test with the unstudentized statistic $W_{U,n}$ is valid in the sense that its limiting null rejection probability is no greater than the nominal level α. Furthermore, the limiting null rejection probability of the bootstrap test with the studentized statistic W_n does not exceed the nominal level by $1/2^{q-1}$, which decreases exponentially with the total number of clusters (instead of the number of "strong" clusters). In addition, besides for the commonly used TSLS estimator, these validity results can also be shown for other estimators proposed in the IV literature.[3] We omit details for brevity but notice that these alternative estimators typically have smaller bias than TSLS in the overidentified case, and their corresponding bootstrap tests could therefore have better finite-sample size control since a randomization test with sign changes requires distributional symmetry around zero.

We also note that instead of applying the procedure described in (17)-(19), one might consider to employ an alternative double-equation bootstrap procedure (e.g., see Moreira et al. (2009), Davidson and MacKinnon (2010), Finlay and Magnusson (2019), Roodman et al. (2019) and Young (2020)):

$$\begin{aligned} X^*_{i,j}(g) &= Z'_{i,j}\widehat{\Pi}_z + W'_{i,j}\widehat{\Pi}_w + g_j\hat{v}_{i,j}, \\ y^*_{i,j}(g) &= X^{*\prime}_{i,j}(g)\hat{\beta}^r_n(\lambda) + W'_{i,j}\hat{\gamma}^r_n(\lambda) + g_j\hat{\epsilon}_{i,j}(\lambda), \end{aligned} \tag{24}$$

where $\widehat{\Pi}_z$ and $\widehat{\Pi}_w$ are the first-stage least squares estimators computed using the full sample, $\hat{v}_{i,j}$ is the corresponding residual[4], and the bootstrap analogues of the TSLS estimator use $\big(y^*_{i,j}(g), X^{*\prime}_{i,j}(g)\big)$ generated by (24) in place of $(y_{i,j}, X'_{i,j})$ with the same $(Z'_{i,j}, W'_{i,j})$. The results in Theorem 3.1 also holds for this procedure as it is asymptotically equivalent to the procedure in (17)-(19) in the case with at least one "strong" cluster.

Remark 1. The bootstrap tests with $W_{U,n}$ and W_n have remarkable resemblance to the Fama-Macbeth type approach in Ibragimov and Müeller (2010, IM) and the randomization test with sign changes in Canay et al. (2017, CRS), which are based on the asymptotic independence

[3] For example, the limited information maximum likelihood (LIML) estimator, Fuller (1977)'s modified LIML estimator, the bias-adjusted TSLS estimator (e.g., Nagar (1959), Rothenberg (1984)), and various jackknife IV estimators (JIVEs; e.g., Phillips and Hale (1977), Angrist, Imbens, and Krueger (1999), Chao, Swanson, Hausman, Newey, and Woutersen (2012), Hausman, Newey, Woutersen, Chao, and Swanson (2012))

[4] Besides $\widehat{\Pi}_z$ and $\widehat{\Pi}_w$, one might consider to generate the bootstrap samples by using more efficient estimators proposed by Davidson and MacKinnon (2010, 2012, 2014).

of cluster-level estimators (say, $\hat{\beta}_{n,1}, ..., \hat{\beta}_{n,q}$) when applied to the setting of clustered data. In addition, IM's approach requires the asymptotic normality of the q cluster-level estimators and CRS's approach requires that these estimators have limiting distributions that are symmetric about zero (after an appropriate recentering). We notice that in the context of IV regressions, the bootstrap tests can be very different from these two approaches with regard to the required instrument strength. In particular, to achieve asymptotic validity, IM and CRS would require the instruments being strong for all clusters; e.g., for all clusters one needs to rule out the presence of weak instruments in the sense of Staiger and Stock (1997) (i.e., $\Pi_{z,j} = n_j^{-1/2} C_j$, where C_j has a fixed full rank value), as the cluster-level IV estimators of the "weak" clusters would become inconsistent and have highly nonstandard limiting distributions, violating the assumptions underlying IM and CRS's approaches. By contrast, the results in Theorem 3.1 hold even with only one "strong" cluster, since the randomization with sign changes for the bootstrap procedure in (17)-(20) is implemented on the score component of the full-sample estimator rather than directly on the cluster-level estimators. In this sense, the bootstrap tests are more robust to cluster heterogeneity/outlier in terms of instrument strength.

Moreover, when the IV estimator applied in the regression has substantial finite sample bias (e.g., TSLS in the over-identified case), the bootstrap tests may perform better as they are based on a full-sample estimator, rather than an average of cluster-level estimators whose finite sample bias may not average out. By contrast, in the case that all clusters are "strong" and/or the cluster-level IV estimators have minimal bias, the approaches of IM and CRS have advantage over the bootstrap as they require neither the condition on cluster homogeneity in Assumption 2(iii) nor the condition that the limits of cluster-level Jacobian being proportional to each other as in Assumption 3(ii)[5]. Therefore, the wild bootstrap and the cluster-level estimator-based approaches can be considered as complements as there are scenarios where one would be preferred to the other.

Remark 2. In general, the results in Theorem 3.1 do not hold for the two bootstrap tests when all clusters are "weak". Intuitively, further complication arises because $\sum_{i \in I_{n,j}} \tilde{Z}_{i,j} v_{i,j} / \sqrt{n}$, the noise part in the first-stage of the model in (1), enters the distributions of interest. Indeed,

[5] In the case of testing $H_0 : \beta = \beta_0$, Assumption 3(ii) is not required to establish Theorem 3.1 for the two wild bootstrap tests, but Assumption 2(iii) would still be required.

under the weak-instrument parameter sequence such that $\Pi_{z,j} = n_j^{-1/2} C_j$ with some fixed full rank C_j for all $j \in J$, the sample Jacobian

$$\frac{1}{\sqrt{n}} \sum_{j \in J} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} X_{i,j}' \xrightarrow{d} \sum_{j \in J} \sqrt{\xi_j} Q_{\tilde{Z}\tilde{Z},j} C_j + \sum_{j \in J} \sqrt{\xi_j} \mathcal{Z}_{v,j}, \tag{25}$$

where $\sum_{j \in J} \sqrt{\xi_j} Q_{\tilde{Z}\tilde{Z},j} C_j$, the signal part of the first-stage equation, is of the same order of magnitude as the noise part $\sum_{j \in J} \sqrt{\xi_j} \mathcal{Z}_{v,j}$. A randomization test with sign changes would not work in this case because for each $j \in J$, (i) the distribution of $\sqrt{\xi_j} \left(Q_{\tilde{Z}\tilde{Z},j} C_j + \mathcal{Z}_{v,j} \right)$ is not symmetric around zero, and (ii) C_j cannot be consistently estimated so that one could not demean either. In particular, the double-equation procedure in (24) would result in the following limiting distribution:

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{j \in J} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} X_{i,j}^{*\prime}(g) \quad \xrightarrow{d} \quad & \sum_{j \in J} \sqrt{\xi_j} Q_{\tilde{Z}\tilde{Z},j} C_j + \sum_{j \in J} \sqrt{\xi_j} g_j \mathcal{Z}_{v,j} \\ & + \sum_{j \in J} \xi_j (1 - g_j) Q_{\tilde{Z}\tilde{Z},j} Q_{\tilde{Z}\tilde{Z}}^{-1} \left(\sum_{\tilde{j} \in J} \sqrt{\xi_{\tilde{j}}} \mathcal{Z}_{v,j} \right). \end{aligned} \tag{26}$$

The first term in (26) equals the signal part in (25), the second term equals the **G**-transformed version of the noise part in (25), while the third is an extra term: the procedure mimics the noise correctly when $g_j = 1$ but over-states the noise when $g_j = -1$.

Remark 3. However, for the empirically prevalent case of testing the coefficient of single endogenous regressor with single instrument (e.g., 101 out of 230 specifications in Andrews et al. (2019)'s sample and 1087 out of 1359 in Young (2020)'s sample), the $W_{U,n}$-based unstudentized bootstrap test is fully robust to weak instrument. Indeed, in this particular case the unstudentized bootstrap test is equivalent to certain version of bootstrap AR test (the $AR_{U,n}$-based unstudentized bootstrap test in Section 3.2), and its asymptotic null rejection probability is no larger than the nominal level irrespective of instrument strength. We notice that such equivalence also holds for the standard framework in which the number of observations/clusters tends to infinity, and the unstudentized wild bootstrap test is thus fully robust to weak instrument under such framework as well. By contrast, the studentized wild bootstrap test, which is more widely used in practice (e.g., see Cameron et al. (2008), Cameron and Miller (2015), MacKinnon and Webb (2017), and Roodman et al. (2019)), is not weak-instrument robust no matter under the standard framework or the framework with few clusters, and thus may produce substantial size distortions even in the case with single instrument, as illustrated by simulations in Section 4.

Therefore, although we expect that in the strong-instrument case with a large number of observations/clusters (so that the bootstrap consistently estimates distributions of interest), bootstrapping an asymptotically pivotal statistic such as W_n can achieve a higher order refinement (e.g., see Beran (1988), Hall (1992), Horowitz (2001), Djogbenou et al. (2019)), here it could be more desirable to use the unstudentized wild bootstrap test from the viewpoint of robustness, especially when the number of clusters is small. Furthermore, notice that its validity under both weak instrument and few clusters depends crucially on the Rademacher weight and the specific procedure in (17)-(20), and thus could not be extended to alternative procedures such as the double-equation procedure in (24) or the commonly employed pairs cluster bootstrap (including percentile, percentile-t, and bootstrap standard error).

3.2 Wild bootstrap with weak-instrument-robust statistics

Similarly, we may define the procedure of the bootstrapped tests for $H_0 : \beta = \beta_0$ with the AR, LM, and CQLR statistics and their unstudentized versions under the form of randomization inference with sign changes:

1. Compute the null-restricted residual

$$\hat{\epsilon}^r_{i,j}(\beta_0) = y_{i,j} - X'_{i,j}\beta_0 - W'_{i,j}\bar{\gamma}^r_n(\beta_0), \tag{27}$$

   where $\bar{\gamma}^r_n(\beta_0)$ is the H_0-restricted least squares estimator of γ.

2. Let $\mathbf{G} = \{-1, 1\}^q$ and for any $g = (g_1, ..., g_q) \in \mathbf{G}$ define

$$\begin{aligned} \widehat{f}^*_n(\beta_0, g) &= n^{-1} \sum_{j \in J} \sum_{i \in I_{n,j}} f^*_{i,j}(\beta_0, g_j), \\ \widehat{\Omega}^*_n(\beta_0, g) &= n^{-1} \sum_{j \in J} \sum_{i \in I_{n,j}} \sum_{k \in I_{n,j}} f^*_{i,j}(\beta_0, g_j) f^*_{k,j}(\beta_0, g_j)', \\ \widehat{\Omega}^*_{W,n}(\beta_0, g) &= n^{-1} \sum_{j \in J} \sum_{i \in I_{n,j}} \sum_{k \in I_{n,j}} \tilde{Z}_{i,j} \tilde{Z}_{k,j} \hat{u}^*_{i,j}(\beta_0, g) \hat{u}^*_{k,j}(\beta_0, g), \end{aligned} \tag{28}$$

   where $f^*_{i,j}(\beta_0, g_j) = \tilde{Z}_{i,j}\epsilon^*_{i,j}(\beta_0, g_j)$, $\epsilon^*_{i,j}(\beta_0, g_j) = g_j\hat{\epsilon}^r_{i,j}(\beta_0)$ and $\hat{u}^*_{i,j}(\beta_0, g_j)$ equals the residual of regressing $\epsilon^*_{i,j}(\beta_0, g_j)$ on $Z_{i,j}$ and $W_{i,j}$.

For the bootstrapped LM and CQLR tests, also compute

$$\begin{aligned}
\widehat{D}_n^*(\beta_0,g) &= \left(\widehat{D}_{1,n}^*(\beta_0,g),...,\widehat{D}_{d_x,n}^*(\beta_0,g)\right), \\
\widehat{D}_{l,n}^*(\beta_0,g) &= \widehat{G}_{l,n}-\widehat{\Gamma}_{l,n}^*(\beta_0,g)\widehat{\Omega}_n^{*-1}(\beta_0,g)\widehat{f}_n^*(\beta_0,g), \\
\widehat{\Gamma}_{l,n}^*(\beta_0,g) &= n^{-1}\sum_{j\in J}\sum_{i\in I_{n,j}}\sum_{k\in I_{n,j}}\left(\tilde{Z}_{i,j}\hat{v}_{i,j,l}^*(g_j)\right)f_{k,j}^*(\beta_0,g_j)',\ \text{for } l=1,...,d_x, \qquad (29)
\end{aligned}$$

where $\hat{v}_{i,j,l}^*(g_j)$ equals the residual of regressing $v_{i,j,l}^*(g_j)=g_j\hat{v}_{i,j,l}$ on $Z_{i,j}$ and $W_{i,j}$.

3. Compute the bootstrap analogues of the test statistics:

$$\begin{aligned}
AR_n^*(\beta_0,g) &= n\widehat{f}_n^*(\beta_0,g)'\widehat{\Omega}_n^{*-1}(\beta_0,g)\widehat{f}_n^*(\beta_0,g), \\
AR_{W,n}^*(\beta_0,g) &= n\widehat{f}_n^*(\beta_0,g)'\widehat{\Omega}_{W,n}^{*-1}(\beta_0,g)\widehat{f}_n^*(\beta_0,g), \\
AR_{U,n}^*(\beta_0,g) &= \left\|\sqrt{n}\widehat{f}_n^*(\beta_0,g)\right\|^2, \\
LM_n^*(\beta_0,g) &= n\widehat{f}_n^*(\beta_0,g)'\widehat{\Omega}_n^{*-1/2}(\beta_0,g)P_{\widehat{\Omega}_n^{*-1/2}(\beta_0,g)\widehat{D}_n^*(\beta_0,g)}\widehat{\Omega}_n^{*-1/2}(\beta_0,g)\widehat{f}_n^*(\beta_0,g), \\
LM_{U,n}^*(\beta_0,g) &= \left\|\sqrt{n}\widehat{D}_n^{*'}(\beta_0,g)\widehat{\Omega}_n^{*-1/2}(\beta_0,g)\widehat{f}_n^*(\beta_0,g)\right\|^2, \\
LR_n^*(\beta_0,g) &= \frac{1}{2}\left(AR_n^*(\beta_0,g)-rk_n(\beta_0)+\sqrt{\left(AR_n^*(\beta_0,g)-rk_n(\beta_0)\right)^2+4LM_n^*(\beta_0,g)\cdot rk_n(\beta_0)}\right), \\
LR_{U,n}^*(\beta_0,g) &= \frac{1}{2}\Bigg(AR_{U,n}^*(\beta_0,g)-rk_n(\beta_0) \\
&\quad +\sqrt{\left(AR_{U,n}^*(\beta_0,g)-rk_n(\beta_0)\right)^2+4LM_{U,n}^*(\beta_0,g)\cdot rk_n(\beta_0)}\Bigg). \qquad (30)
\end{aligned}$$

4. The bootstrapped tests and the corresponding critical values are defined in the same fashion as in Step 5 of the bootstrapped t-test.

The following theorem shows that in the general case with multiple instruments, the $AR_{U,n}(\beta_0)$-based unstudentized wild bootstrap test is fully robust to weak instruments and few clusters in the sense that its limiting null rejection probability is no greater than the nominal level α, irrespective of instrument strength. In addition, its limiting null rejection probability is bounded from below by $\alpha-1/2^{q-1}$. On the other hand, the theorem also shows that when the number of instruments is smaller than the total number of clusters, the limiting null rejection probabilities of the two studentized bootstrap AR tests are bounded by $\alpha-1/2^{q-1}$ from below and by $\alpha+1/2^{q-1}$ from above, respectively.

Theorem 3.2 *If Assumption 2(i)-(iii) and $H_0 : \beta = \beta_0$ holds, then*

$$\alpha - \frac{1}{2^{q-1}} \leq \liminf_{n\to\infty} P\{AR_{U,n}(\beta_0) > \hat{c}^{ar}_{u,n}(1-\alpha)\} \leq \limsup_{n\to\infty} P\{AR_{U,n}(\beta_0) > \hat{c}^{ar}_{u,n}(1-\alpha)\} \leq \alpha,$$

and if further $d_z < q$, then

$$\alpha - \frac{1}{2^{q-1}} \leq \liminf_{n\to\infty} P\{AR_n(\beta_0) > \hat{c}^{ar}_{n}(1-\alpha)\} \leq \limsup_{n\to\infty} P\{AR_n(\beta_0) > \hat{c}^{ar}_{n}(1-\alpha)\} \leq \alpha + \frac{1}{2^{q-1}};$$

$$\alpha - \frac{1}{2^{q-1}} \leq \liminf_{n\to\infty} P\{AR_{W,n}(\beta_0) > \hat{c}^{ar}_{r,n}(1-\alpha)\} \leq \limsup_{n\to\infty} P\{AR_{W,n}(\beta_0) > \hat{c}^{ar}_{r,n}(1-\alpha)\} \leq \alpha + \frac{1}{2^{q-1}},$$

where $\hat{c}^{ar}_{u,n}(1-\alpha)$, $\hat{c}^{ar}_{n}(1-\alpha)$ and $\hat{c}^{ar}_{r,n}(1-\alpha)$ denote the critical values of the $AR_{U,n}(\beta_0)$, $AR_n(\beta_0)$ and $AR_{W,n}(\beta_0)$-based bootstrap tests, respectively.

Remark 4. In terms of size properties under a small number of clusters, the bootstrap AR tests have substantial improvement over the AR tests with conventional asymptotic critical values. We notice that the $AR_n(\beta_0)$-based asymptotic test typically under-rejects or does not reject at all in the over-identified case (in the simulations of Section 4, its null rejection frequencies equal zero for the cases with 10 clusters and 3 or 5 instruments). In particular, the null rejection probabilities of this AR test decreases toward zero when the number of instruments d_z approaches the number of clusters q; in fact, when d_z is equal to q, the statistic $AR_n(\beta_0)$ will be exactly equal to d_z (or q) since for $\widehat{F}_n(\beta_0) = \left(\widehat{f}_{1,n}(\beta_0), ..., \widehat{f}_{q,n}(\beta_0)\right)'$ and $\widehat{f}_{j,n}(\beta_0) = n^{-1}\sum_{i\in I_{n,j}} f_{i,j}(\beta_0)$ with $j = 1, ..., q$,

$$AR_n(\beta_0) = \ell'\widehat{F}_n(\beta_0)\left(\widehat{F}_n(\beta_0)'\widehat{F}_n(\beta_0)\right)^{-1}\widehat{F}_n(\beta_0)'\ell = \ell'\ell = d_z \tag{31}$$

as long as $\widehat{F}_n(\beta_0)$ is invertible, where ℓ denotes a q-dimensional vector of ones. The $AR_{W,n}(\beta_0)$ statistic also cannot be employed in this case since its variance-covariance matrix estimator $\widehat{V}_{W,n}$ would become singular. Moreover, the asymptotic test that is based on $AR_{W,n}(\beta_0)$ tends to have substantial over-rejections in the case with few clusters, as illustrated by the simulations.

Compared with the asymptotic tests, all the three bootstrap tests typically have much better size controls. With regard to power properties, in the over-identified case the $AR_n(\beta_0)$ and $AR_{W,n}(\beta_0)$-based studentized bootstrap tests may suffer from the issue of low power due to similar problems as those for the asymptotic tests (e.g., the value of the bootstrap analogue of $AR_n(\beta_0)$ will also be exactly equal to d_z when d_z is equal to q). On the other hand, the $AR_{U,n}(\beta_0)$-based unstudentized bootstrap test does not have such issue and also works well even in the case with d_z larger than q. Overall, we recommend to use the unstudentized bootstrap AR test instead of the others when the number of clusters is small.

Remark 5. It is also possible to modify the bootstrap AR tests so that they can be applied to the cases where Assumption 2(iii) of cluster homogeneity may not hold. For instance, the modified $AR_{U,n}(\beta_0)$ statistic can be defined as

$$AR^c_{U,n}(\beta_0) \quad \equiv \quad \left\| \sqrt{n} \widehat{f}^c_n(\beta_0) \right\|^2, \tag{32}$$

where $\widehat{f}^c_n(\beta_0) = n^{-1} \sum_{j \in J} \sum_{i \in I_{n,j}} f^c_{i,j}(\beta_0)$, $f^c_{i,j}(\beta_0) = \tilde{Z}^c_{i,j} \hat{\epsilon}^c_{i,j}(\beta_0)$, $\hat{\epsilon}^c_{i,j}(\beta_0) = y_{i,j} - X'_{i,j}\beta_0 - W_{i,j}\hat{\gamma}^r_{n,j}(\beta_0)$, with $\tilde{Z}^c_{i,j}$ and $\hat{\gamma}^r_{n,j}(\beta_0)$ being the cluster-level residuals from regressing $Z_{i,j}$ on $W_{i,j}$ (i.e., $\tilde{Z}^c_{i,j} = Z_{i,j} - \widehat{\Gamma}^{c\prime}_{n,j} W_{i,j}$) and the null-restricted least squares estimator of γ only using the sample in the j-th cluster, respectively. Assuming that for all $j \in J$,

$$\left\{ \frac{1}{\sqrt{n_j}} \sum_{i \in I_{n,j}} \tilde{Z}^c_{i,j} \epsilon_{i,j} : j \in J \right\} \xrightarrow{d} \left\{ \mathcal{Z}^c_{\epsilon,j} : j \in J \right\},$$

where $\mathcal{Z}^c_{\epsilon,j} \sim N\left(0, \Sigma^c_j\right)$ with some positive definite Σ^c_j, then the result for $AR_{U,n}(\beta_0)$ in Theorem 3.2 can be established under arbitrary cluster heterogeneity for $AR^c_{U,n}(\beta_0)$-based bootstrap test with $\widehat{f}^{c*}_n(\beta_0, g) = n^{-1} \sum_{j \in J} \sum_{i \in I_{n,j}} f^{c*}_{i,j}(\beta_0, g_j)$, $f^{c*}_{i,j}(\beta_0, g_j) = \tilde{Z}^c_{i,j} \epsilon^{c*}_{i,j}(\beta_0, g_j)$, and $\epsilon^{c*}_{i,j}(\beta_0, g_j) = g_j \hat{\epsilon}^c_{i,j}(\beta_0)$. We may do similar modifications to $AR_n(\beta_0)$ and $AR_{W,n}(\beta_0)$ as well.

We also notice that different from the original bootstrap test, the modified bootstrap test requires the parameter of interest to be identified within each cluster (this is similar to IM and CRS's approaches; e.g., see the discussions in p.1025 of CRS). For example, consider a clustered regression model with endogenous treatment effect,

$$y_{i,j} \quad = \quad \theta + \beta X_{i,j} + W'_{i,j}\gamma + \epsilon_{i,j}, \tag{33}$$

where $y_{i,j}$ denotes the outcome of unit i in group or area j, $X_{i,j}$ a single endogenous regressor (e.g., the treatment status or dose), $W_{i,j}$ a vector of covariates that vary within each cluster, Z_j the cluster-level random assignment status of treatment, and the quantity of interest is the treatment effect β. Let J_1 the set of clusters such that $Z_j = 1$ and J_0 the set of clusters such that $Z_j = 0$. To implement the test, we need to define the clusters by forming pairs of groups or areas, that is, by matching each group in J_1 with a group in J_0 (e.g., in experimental settings, such pairs can be determined by the treatment assignment status of each group). Such pairing would reduce the number of clusters available for inference by half, while there is no need for pairing when implementing the original bootstrap AR tests. Therefore, the original and modified bootstrap AR tests are also complement to each other.

Remark 6. For empirical applications involving treatment effect such as the one in (33), we may consider an alternative AR-type procedure by imposing the null hypothesis $H_0 : \beta = \beta_0$ into the structural form equation so that

$$y_{i,j} - \beta_0 X_{i,j} = \theta + W_{i,j}' \gamma + \epsilon_{i,j}. \tag{34}$$

Since θ can be identified in each cluster $j \in J_0 \cup J_1$, we may therefore run the least squares regressions for the q clusters separately, and obtain their estimates as $(\hat{\theta}_{n,1}, ..., \hat{\theta}_{n,q})$. Then, we may define a two-sample test statistic based on the cluster-level estimates:

$$\frac{1}{|J_1|} \sum_{j \in J_1} \hat{\theta}_{n,j} - \frac{1}{|J_0|} \sum_{j \in J_0} \hat{\theta}_{n,j}. \tag{35}$$

Notice that under our current framework and the null hypothesis, $\sqrt{n}\left(\hat{\theta}_{n,j} - \theta\right) \xrightarrow{d} N\left(0, \sigma_j^2\right)$ for $j \in J_0 \cup J_1$, so the two-sample t-test in Ibragimov and Müller (2016) and the adjusted permutation test in Hagemann (2019), which is based on permuting $(\hat{\theta}_1, ..., \hat{\theta}_q)$ and adjusted critical values, will be asymptotically valid for the test statistic in (35) with arbitrary cluster heterogeneity. The number of clusters available for inference under these procedures is equal to q (if one use IV estimator-based statistics instead, than again one has to pair the treatment and control groups for identification). We also notice that (35) is closely related to the permutation test proposed by Rosenbaum (1996) and Imbens and Rosenbaum (2005), which is exact for testing sharp null hypothesis under a finite-population perspective.

The behaviour of the LM and CQLR tests is more complicated than the AR test as they depend on the adjusted sample Jacobian $\widehat{D}_n(\beta_0)$. Similar to the bootstrap IV estimate and t-test, further complication arises for the bootstrap LM and CQLR tests in the case that all the clusters are "weak", as the noise part in the first-stage enters the distributions of interest. For instance, let us consider the LM statistic and also suppose that $k_x = 1$ for notational simplicity.

For the adjusted sample Jacobian we notice that

$$\begin{aligned}&\sqrt{n}\widehat{D}_n(\beta_0)\\ =\ &\sqrt{n}\widehat{G}_n-\left(\sum_{j\in J}\frac{n_j}{n}\left(\frac{1}{\sqrt{n_j}}\sum_{i\in I_{n,j}}\tilde{Z}_{i,j}\hat{v}_{i,j}\right)\left(\frac{1}{\sqrt{n_j}}\sum_{k\in I_{n,k}}f_{k,l}(\beta_0)\right)'\right)\widehat{\Omega}_n^{-1}(\beta_0)\left(\sqrt{n}\widehat{f}_n(\beta_0)\right),\end{aligned} \tag{36}$$

where the distributions of $\sqrt{n}\widehat{G}_n$ and $\sum_{i\in I_{n,j}}\tilde{Z}_{i,j}\hat{v}_{i,j}/\sqrt{n_j}$, the two terms related to the first-stage equation, cannot be well mimicked by the **G**-transformation with sign changes when all the clusters are "weak" for similar reason as that noted in Remark 3. Furthermore, it is clear from (36) that $\sqrt{n}\widehat{D}_n(\beta_0)$ is no longer asymptotically independent from $\sqrt{n}\widehat{f}_n(\beta_0)$ under the framework with fixed number of clusters as the orthogonalization adjustment is no longer valid in this case, thus resulting in a highly nonstandard null limiting distribution for the LM statistic. We therefore cannot give the lower and upper bounds of the limiting null rejection probabilities of the two bootstrap tests in the case that all the clusters are "weak". This is different from Moreira et al. (2004, 2009), who show the bootstrap validity for the LM and CLR tests under the weak-instrument asymptotics and homoskedastic errors.

However, when there is at least one "strong" cluster, we are still able to establish the connection between a randomization test with sign changes and the bootstrap LM and CQLR tests, as shown in Theorem 3.3. In particular, $\widehat{G}_n$ in (36) becomes dominant in this case and

$$LM_n(\beta_0)\xrightarrow{d}\left\|\left(Q'_{\tilde{Z}X}\left(\sum_{j\in J}\xi_j\mathcal{Z}_{\epsilon,j}\mathcal{Z}'_{\epsilon,j}\right)^{-1}Q_{\tilde{Z}X}\right)^{-1/2}Q'_{\tilde{Z}X}\left(\sum_{j\in J}\xi_j\mathcal{Z}_{\epsilon,j}\mathcal{Z}'_{\epsilon,j}\right)^{-1}\sum_{j\in J}\sqrt{\xi_j}\mathcal{Z}_{\epsilon,j}\right\|^2. \tag{37}$$

Although the distribution on the right-hand side of (37) is still nonstandard, we can establish the connection by showing that

$$(LM_n(\beta_0),\{LM_n^*(\beta_0,g):g\in\mathbf{G}\})=(T_{lm}(S_n),\{T_{lm}(gS_n):g\in\mathbf{G}\})+o_P(1), \tag{38}$$

for some statistic S_n and function $T_{lm}(\cdot)$ defined in the proofs of Theorem 3.3. Then, we can show the asymptotic equivalence of the bootstrap LM and bootstrap CQLR tests in this case. Similar arguments are used for their unstudentized version.

Theorem 3.3 *If Assumption 2(i)-(iii), Assumption 3(i), $H_0 : \beta = \beta_0$ holds and $d_z < q$, then*

$$\alpha - \frac{1}{2^{q-1}} \leq \liminf_{n\to\infty} P\{LM_n(\beta_0) > \hat{c}_n^{lm}(1-\alpha)\} \leq \limsup_{n\to\infty} P\{LM_n(\beta_0) > \hat{c}_n^{lm}(1-\alpha)\} \leq \alpha + \frac{1}{2^{q-1}};$$

$$\alpha - \frac{1}{2^{q-1}} \leq \liminf_{n\to\infty} P\{LM_{U,n}(\beta_0) > \hat{c}_{u,n}^{lm}(1-\alpha)\} \leq \limsup_{n\to\infty} P\{LM_{U,n}(\beta_0) > \hat{c}_{u,n}^{lm}(1-\alpha)\} \leq \alpha + \frac{1}{2^{q-1}};$$

$$\alpha - \frac{1}{2^{q-1}} \leq \liminf_{n\to\infty} P\{LR_n(\beta_0) > \hat{c}_n^{lr}(1-\alpha)\} \leq \limsup_{n\to\infty} P\{LR_n(\beta_0) > \hat{c}_n^{lr}(1-\alpha)\} \leq \alpha + \frac{1}{2^{q-1}};$$

$$\alpha - \frac{1}{2^{q-1}} \leq \liminf_{n\to\infty} P\{LR_{U,n}(\beta_0) > \hat{c}_{u,n}^{lr}(1-\alpha)\} \leq \limsup_{n\to\infty} P\{LR_{U,n}(\beta_0) > \hat{c}_{u,n}^{lr}(1-\alpha)\} \leq \alpha + \frac{1}{2^{q-1}},$$

where $\hat{c}_n^{lm}(1-\alpha)$, $\hat{c}_{u,n}^{lm}(1-\alpha)$, $\hat{c}_{u,n}^{lr}(1-\alpha)$ and $\hat{c}_n^{lr}(1-\alpha)$ denote the critical values of the $LM_n(\beta_0)$, $LM_{U,n}(\beta_0)$, $LR_{U,n}(\beta_0)$ and $LR_n(\beta_0)$-based bootstrap tests, respectively.

Remark 7. We emphasize that in our framework, it is assumed for all $j \in J$ that $n_j \to \infty$ as $n \to \infty$, but the number of clusters, q, is fixed, thus very different from the asymptotic framework considered in Djogbenou et al. (2019), MacKinnon et al. (2019), and Hansen and Lee (2019), where the number of clusters tends to infinity with the sample size. Under such framework with $q \to \infty$, one can show that the wild bootstrap procedure for the AR, LM, and CQLR tests are all asymptotically valid for testing the joint null hypothesis $H_0 : \beta = \beta_0$ (by extending the results in Moreira et al. (2009)), no matter the instruments are strong or weak. In particular, $\widehat{f}_n(\beta_0)$ and the orthogonalized Jacobian $\widehat{D}_n(\beta_0)$ are asymptotically independent in such case, and the LM and CQLR statistics will thus follow the limiting distributions given in the weak-instrument literature, which can be consistently estimated by the wild bootstrap. However, weak-instrument-robust inference with regard to a subvector of β would be substantially more complicated since unrestricted structural parameters enter the problem as additional nuisance parameter. Indeed, Wang and Doko Tchatoka (2018) and Wang (2020) show that both residual-based and nonparametric bootstrap procedures are inconsistent even for the subvector AR test under conditional homoskedasticity.

4 Monte Carlo simulation

In this section, we investigate the finite-sample performance of the bootstrap tests with a simulation study. The data is generated as

$$\begin{aligned} y_{i,j} &= \gamma + X_{i,j}\beta + \sigma(Z_{i,j})\left(\eta_{\epsilon,j} + \epsilon_{i,j}\right), \\ X_{i,j} &= \gamma + Z'_{i,j}\Pi_z + \sigma(Z_{i,j})\left(\eta_{v,j} + v_{i,j}\right), \end{aligned} \tag{39}$$

for $i = 1, ..., n$ and $j = 1, ..., q$. The total sample size n is equal to 500, the number of clusters q is equal to 10, and the cluster size is set to be the same. The disturbances $(\epsilon_{i,j}, v_{i,j})$ and cluster effects $(\eta_{\epsilon,j}, \eta_{v,j})$ are specified as follows: $(\epsilon_{i,j}, u_{i,j}) \sim N(0, I_2)$, $v_{i,j} = \rho\epsilon_{i,j} + (1 - \rho^2)^{1/2}u_{i,j}$, $(\eta_{\epsilon,j}, \eta_{u,j}) \sim N(0, I_2)$, $\eta_{v,j} = \rho\eta_{\epsilon,j} + (1 - \rho^2)^{1/2}\eta_{u,j}$. $\rho \in \{0, 0.2, 0.4, 0.6, 0.8, 0.99\}$ corresponds to the degree of endogeneity. The instruments are generated by $Z_{i,j} \sim N(0, I_{d_z})$ and $\sigma(Z_{i,j}) = |\sum_{k=1}^{d_z} Z_{i,j,k}|$. The instrument strength is characterized by the concentration parameter $\Pi'_z \left(\sum_{j\in J}\sum_{i\in I_{n,j}} Z_{i,j}Z'_{i,j}\right)\Pi_z$ equal to 10, 100, and 200. The number of Monte Carlo replications is equal to 5,000.

Figure 1 reports the null empirical rejection frequencies of the cluster-robust tests that are based on the TSLS estimates, including the studentized and unstudentized (single-equation) wild bootstrap tests in Section 3.1, the group-based t-test of Ibragimov and Müeller (2010, 2016), the randomization test of Canay et al. (2017), and the cluster-robust t-test with the conventional asymptotic normal critical values and the critical values proposed by Bester et al. (2011), We notice that size distortions increase for all the tests when the instruments become weak and/or the degree of endogeneity becomes high. The studentized wild bootstrap test has size properties similar to the t-tests with the asymptotic normal or Bester et al. (2011)'s critical value when the instruments are weak, while it typically has smaller size distortions when the instrument becomes strong. Furthermore, the unstudentized wild bootstrap test is found to have the smallest size distortions among these test procedures. In particular, we notice that in line with the discussions in Remark 3, it does not have size distortions in the case with one instrument, irrespective of the instrument strength. Figure 2 reports the results for the studentized and unstudentized double-equation wild bootstrap tests in (24), the studentized and unstudentized pairs bootstrap tests, and the t-test with bootstrap standard error. We notice that the pairs bootstrap tests typically have larger size distortions than their wild bootstrap counterparts, and the t-test with bootstrap standard error has performance very similar to that of the unstudentized pairs bootstrap test.

Figure 3 reports the null empirical rejection frequencies of the same set of tests as those in Figure 1 but with LIML estimates instead of TSLS. We notice that in this case, the approaches of Ibragimov and Müeller (2010, 2016) and Canay et al. (2017) have large improvement upon the case with TSLS estimates. This is due to the fact that these tests are based on cluster-level estimates, which could produce serious finite-sample bias when TSLS is employed, especially in the over-identified case. All the other procedures, including the studentized wild bootstrap test, also have improvement upon their TSLS-based counterparts in Figure 1. Again, the unstudentized wild bootstrap test turns out to have the best size control across different settings of instrument strength, degree of endogeneity, and number of instruments, with null rejection frequencies no larger than 10% in these simulations. Figure 4 reports the results for the double-equation wild bootstrap tests, the pairs bootstrap tests, and the t-test with bootstrap standard error, all based on the LIML estimates. The two studentized bootstrap tests seem to have relatively large size distortions across different settings. The t-test with bootstrap standard error has smaller size distortions than the unstudentized pairs bootstrap test when the degree of endogeneity is high, but may be more conservative in other cases.

Figure 5 reports the rejection frequencies of the AR-based tests, including the AR test and Wald-AR test that are based on the asymptotic critical values, the studentized and unstudentized bootstrap AR tests, and the bootstrap Wald-AR test. Figure 6 reports the rejection frequencies of the asymptotic LM and CQLR tests, and the bootstrap LM and CQLR tests for both studentized and unstudentized versions. We highlight some findings below. First, it turns out that the asymptotic AR test can be very conservative and even does not reject at all when, e.g., the number of instruments equals 3 or 5, while the asymptotic Wald-AR test has serious over-rejections across various settings, with over-rejections increasing with the number of instruments. Second, we notice that in line with our analysis in Section 3.2, the bootstrap LM and CQLR tests also have size distortions when the instruments are weak and/or the degree of endogeneity is high. Moreover, the unstudentized versions tend to under-reject while the studentized versions tend to over-reject when the instruments are weak and/or the degree of endogeneity is high. The bootstrap CQLR tests have slightly smaller distortions than their LM counterparts. By contrast, the three bootstrap AR tests always have rejection frequencies very close to the nominal size. In particular, the studentized bootstrap AR test is able to correct the conservativeness of the asymptotic AR test, and the studentized bootstrap Wald-AR test also largely erases the size distortions of the asymptotic Wald-AR test.

Figures 7 reports the power properties of the AR-based tests with $d_z = 3$, and the results are in line with those found in Figure 5. In particular, among the tests that are able to have good size control (namely, the three bootstrap AR tests), the unstudentized bootstrap AR test has remarkably superior power performance compared with the alternative methods, as discussed in Remark 4. Figure 8 reports the power curves with $d_z = 5$. We observe that the two asymptotic tests become even more distorted in this case with the AR test not rejecting at all while the Wald-AR test having very large size distortions.

5 Conclusion

In this paper, we study the properties of wild bootstrap tests under a framework with a small number of clusters but large numbers of observations per cluster for IV regressions. Our setting allows for cluster heterogeneity in terms of instrument strength, and we show that an unstudentized wild bootstrap test based on IV estimators is valid as long as the instruments are strong for at least one cluster. This is different from alternative methods proposed in the literature for inference with a small number of clusters (e.g., IM and CRS's approaches that are based on cluster-level estimates), whose validity would require that the instruments be strong for all clusters. Moreover, for the leading case in empirical applications with a single instrument, the unstudentized wild bootstrap test generated by our procedure is fully robust to weak instrument in the sense that its limiting null rejection probability is no greater than the nominal level even if all clusters are "weak". However, such robustness is not shared by the studentized wild bootstrap test or the commonly used pairs cluster bootstrap, which may result in serious size distortion in this case. Furthermore, in the general case with multiple instruments, we show that an unstudentized version of bootstrap AR test is fully robust to weak instruments, and is superior with regard to both size and power properties to alternative asymptotic and bootstrap AR tests that employ cluster-robust variance estimators. By contrast, we find that bootstrapping other weak-instrument-robust tests such as the LM and CQLR tests, no matter studentized or unstudentized, does not guarantee correct limiting null rejection probability when all clusters are “weak”. Overall, when the weak instrument issue is a concern and the number of available clusters is small, we recommend to use the unstudentized bootstrap test with TSLS in the case with single instrument, and to use the unstudentized bootstrap AR test in the case with multiple instruments.

Figure 1: Null empirical rejection frequencies of TSLS-based tests (1)

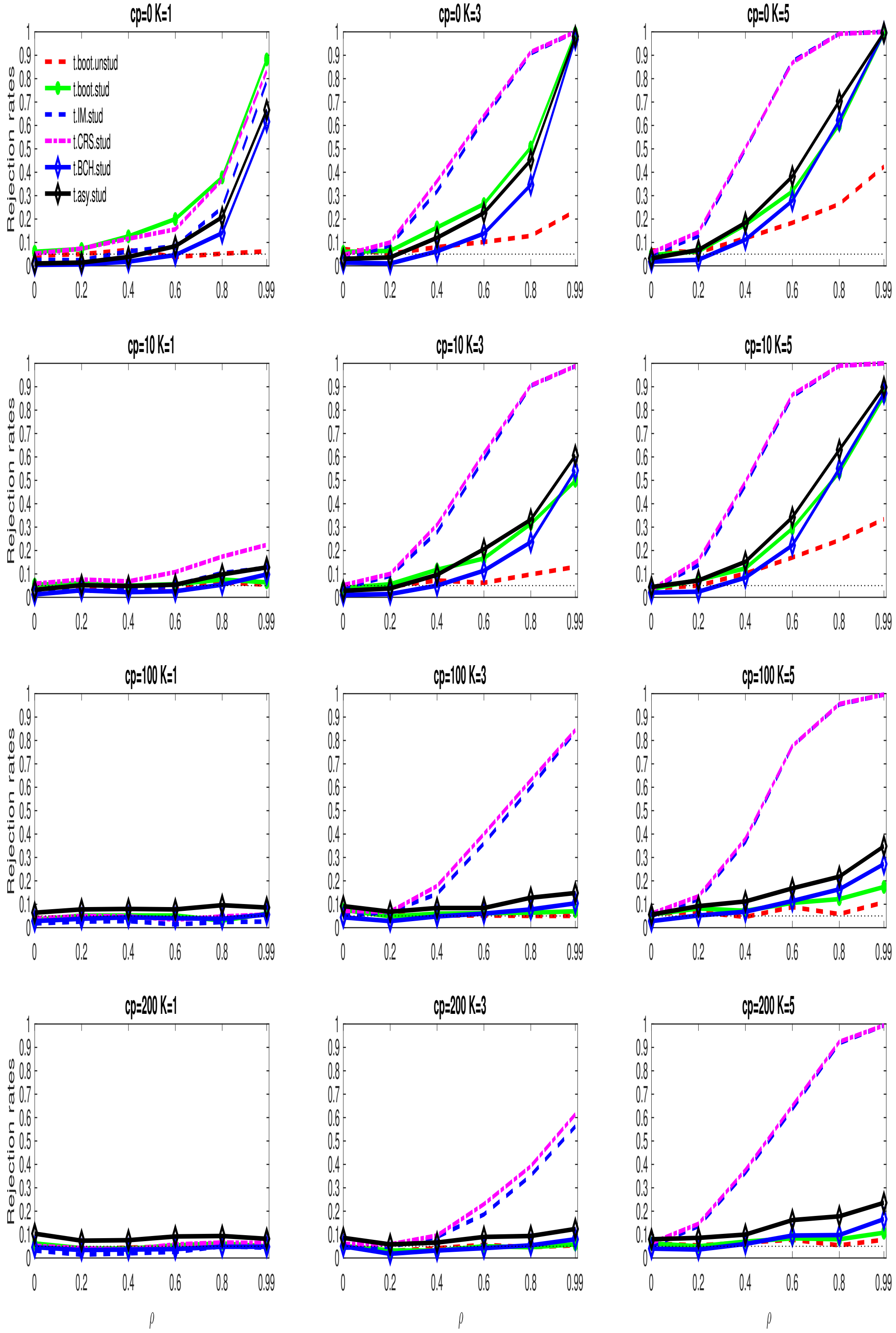


Figure 2: Null empirical rejection frequencies of TSLS-based tests (2)

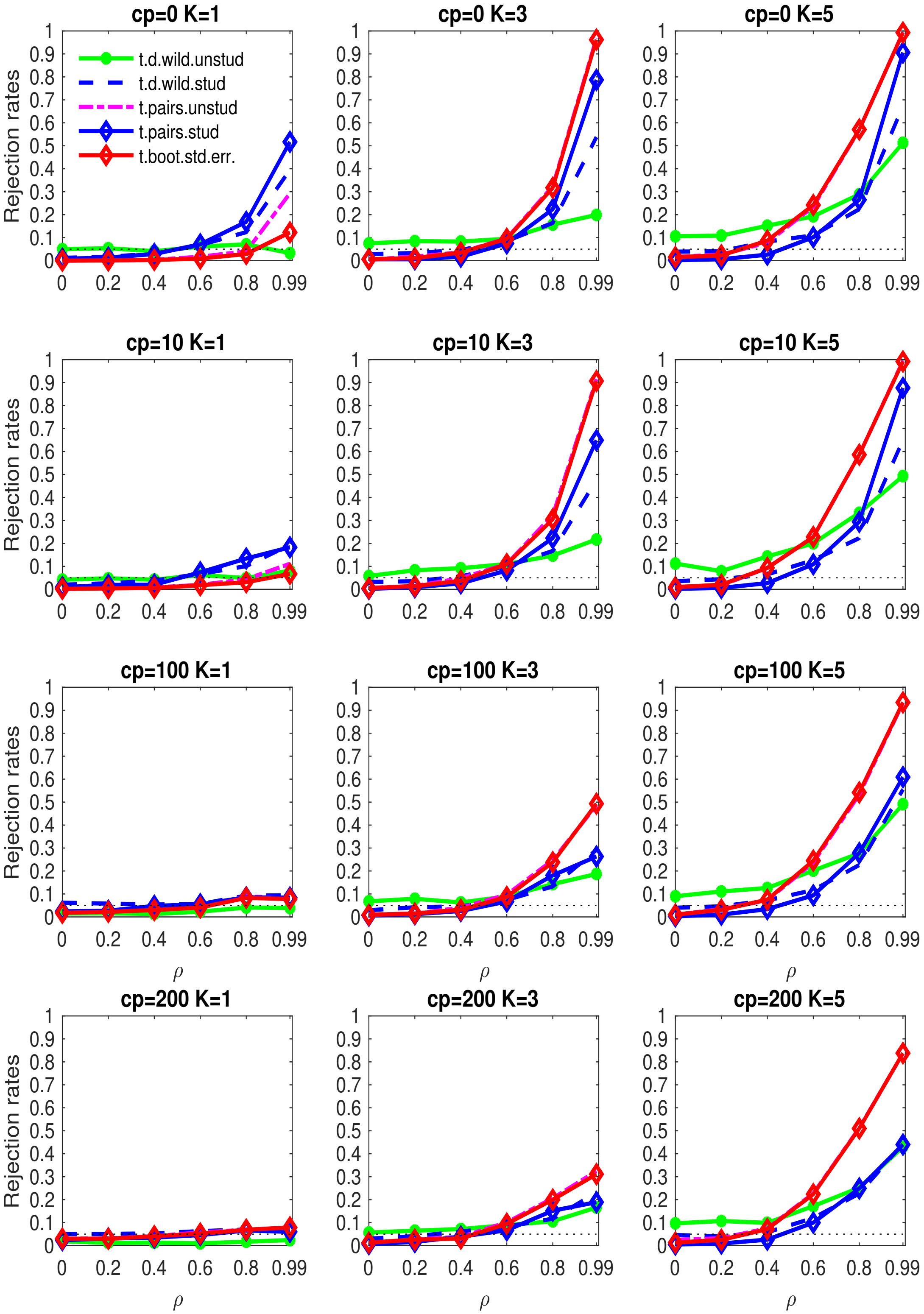


Figure 3: Null empirical rejection frequencies of LIML-based tests (1)

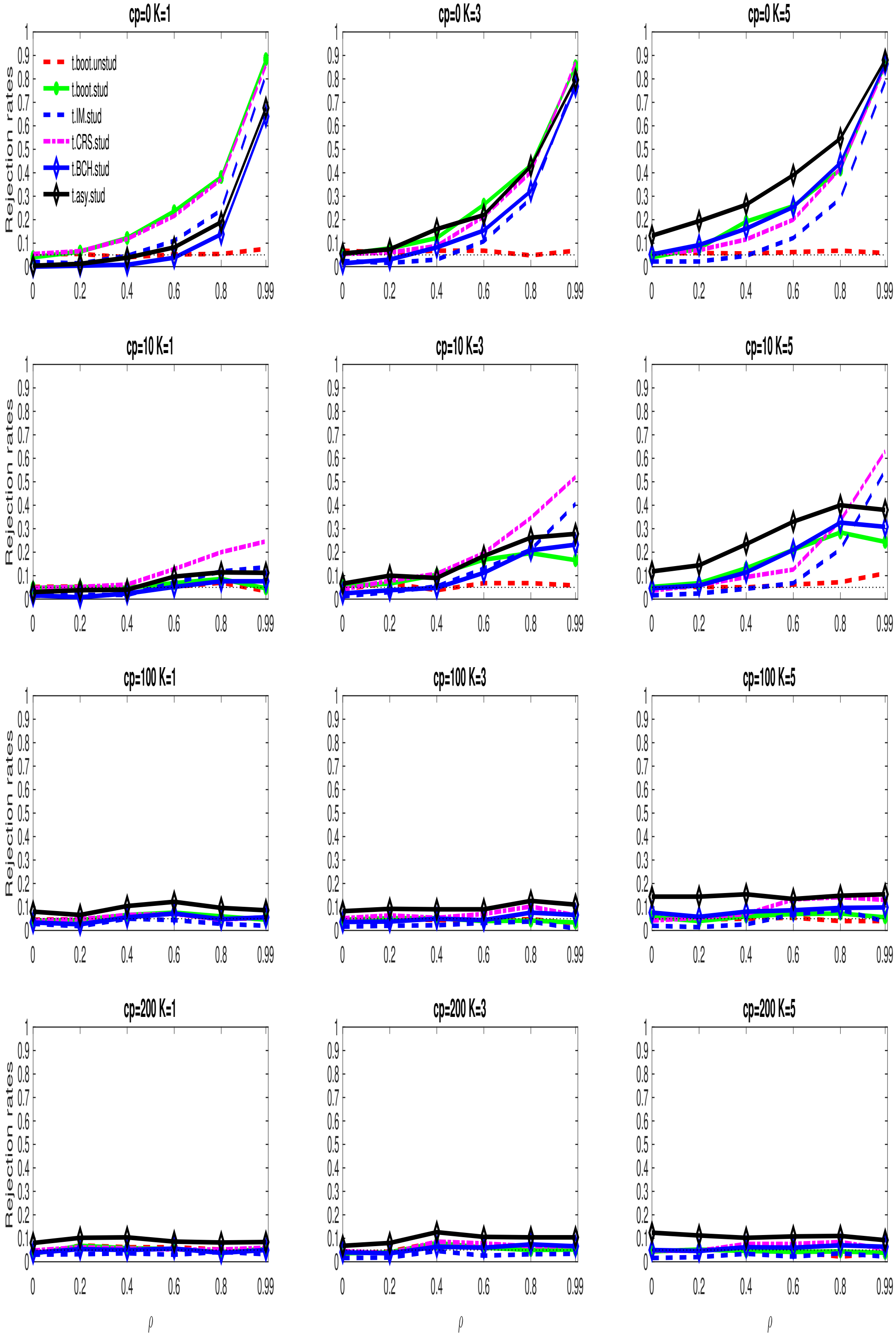


Figure 4: Null empirical rejection frequencies of LIML-based tests (2)

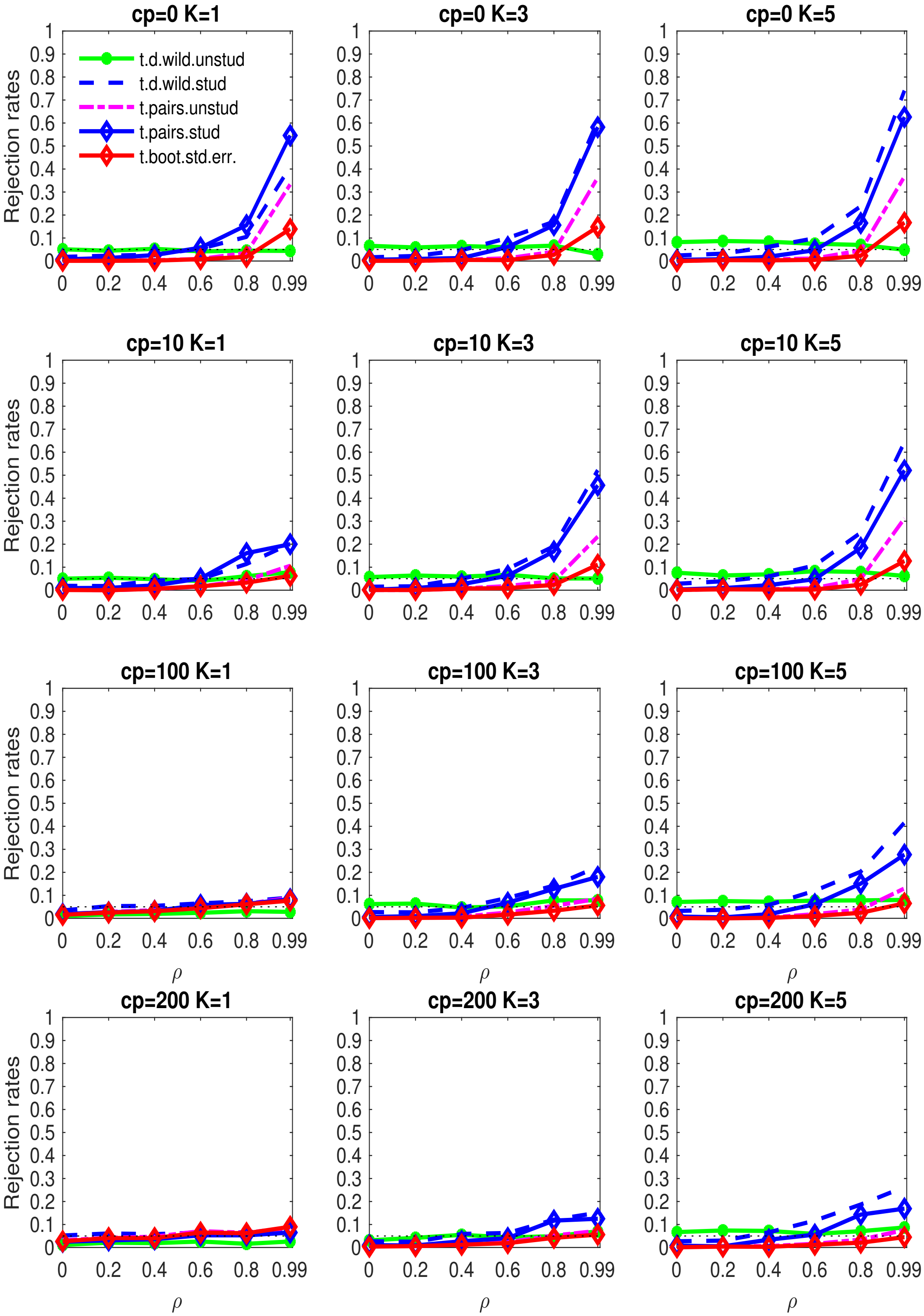


Figure 5: Null empirical rejection frequencies of AR-based tests

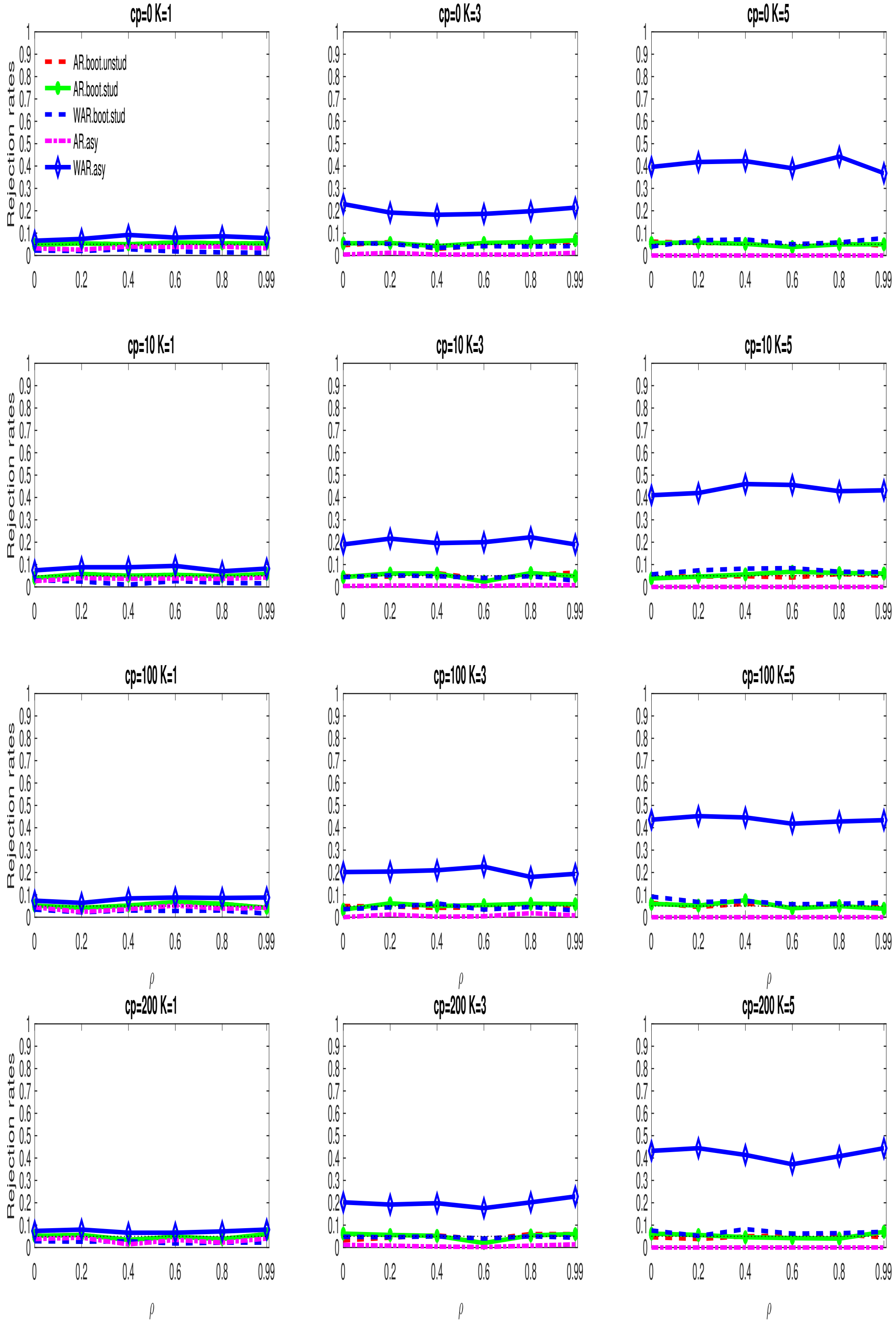


Figure 6: Null empirical rejection frequencies of LM and CQLR-based tests

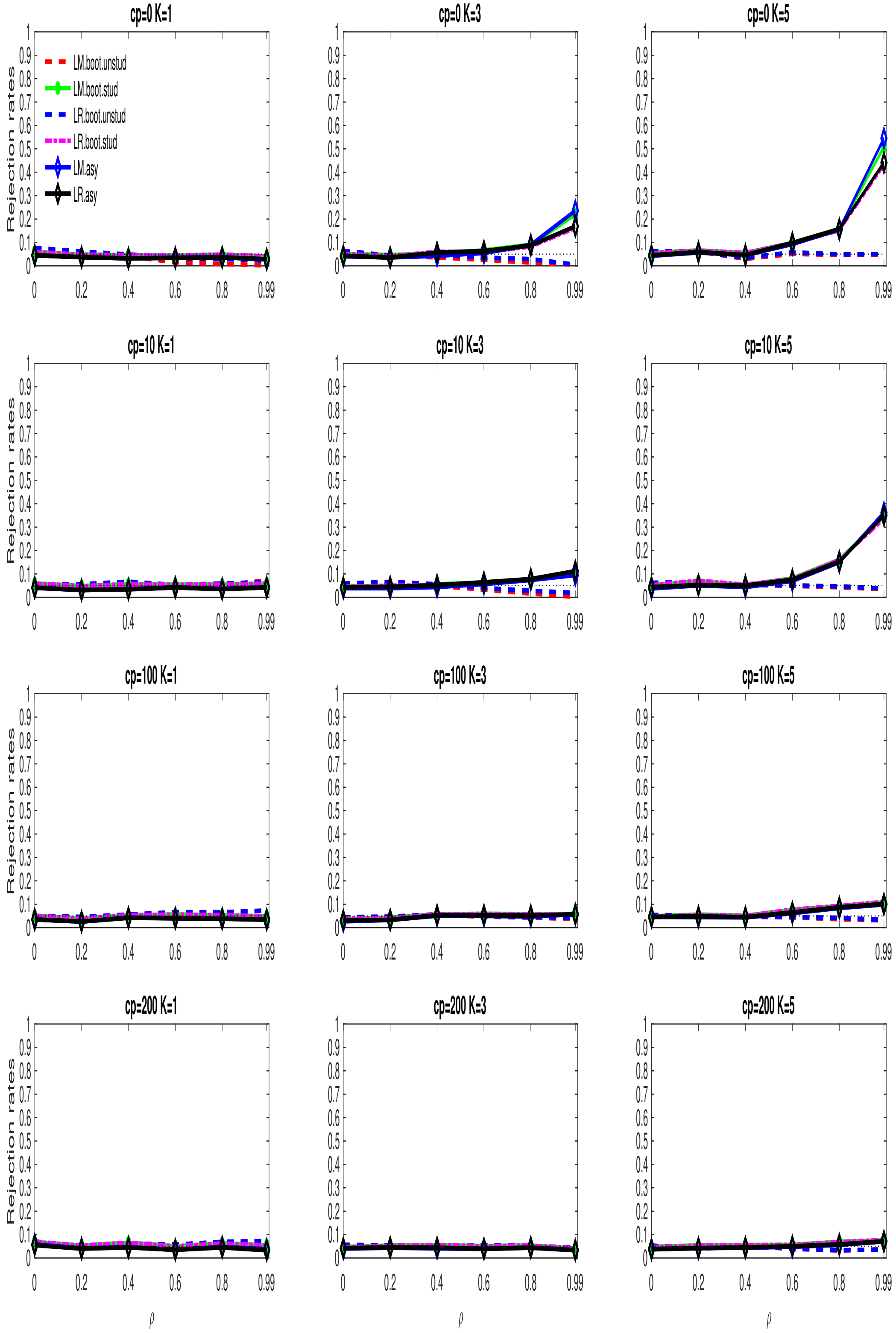


Figure 7: Power of AR-based tests with $d_z = 3$

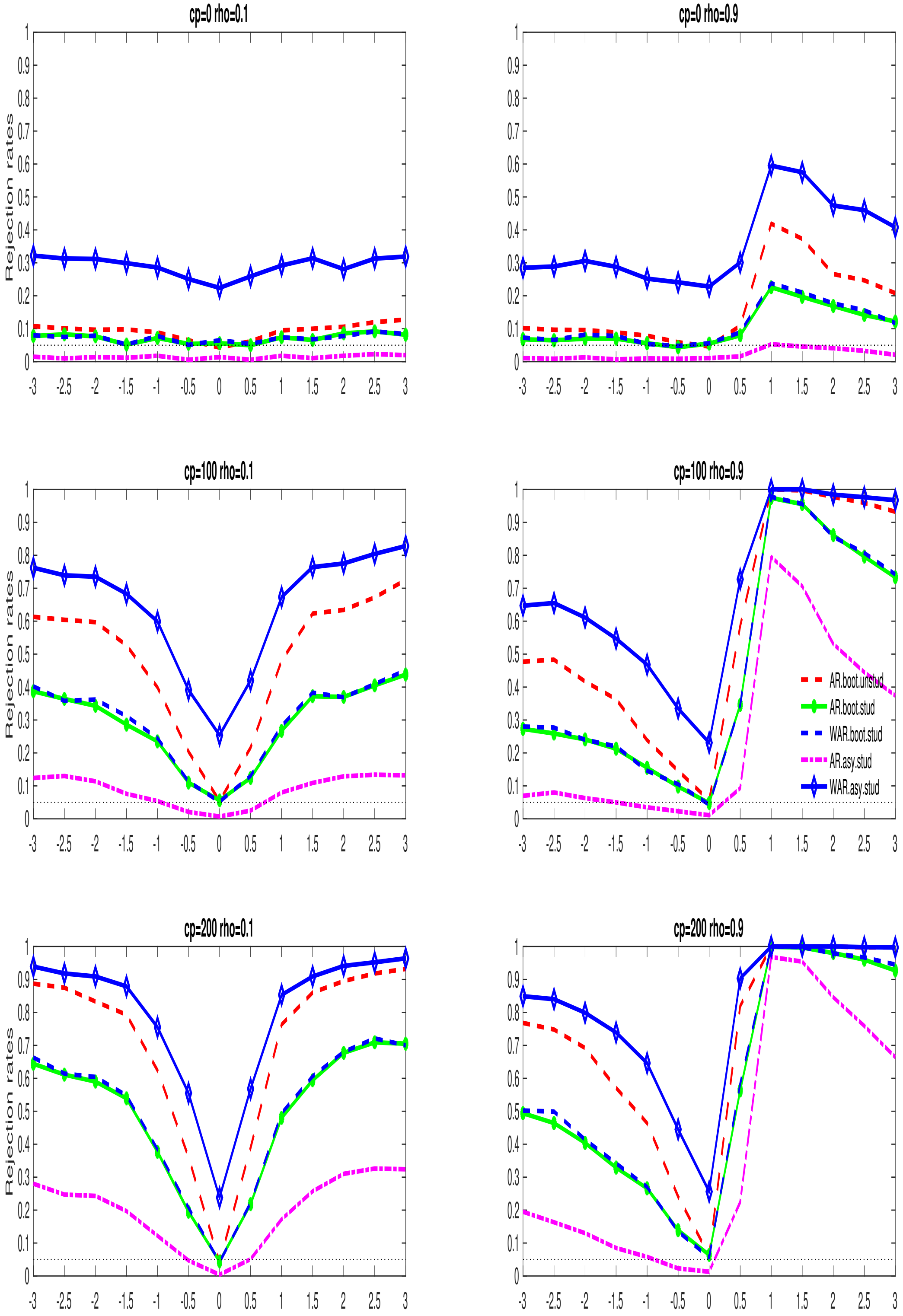
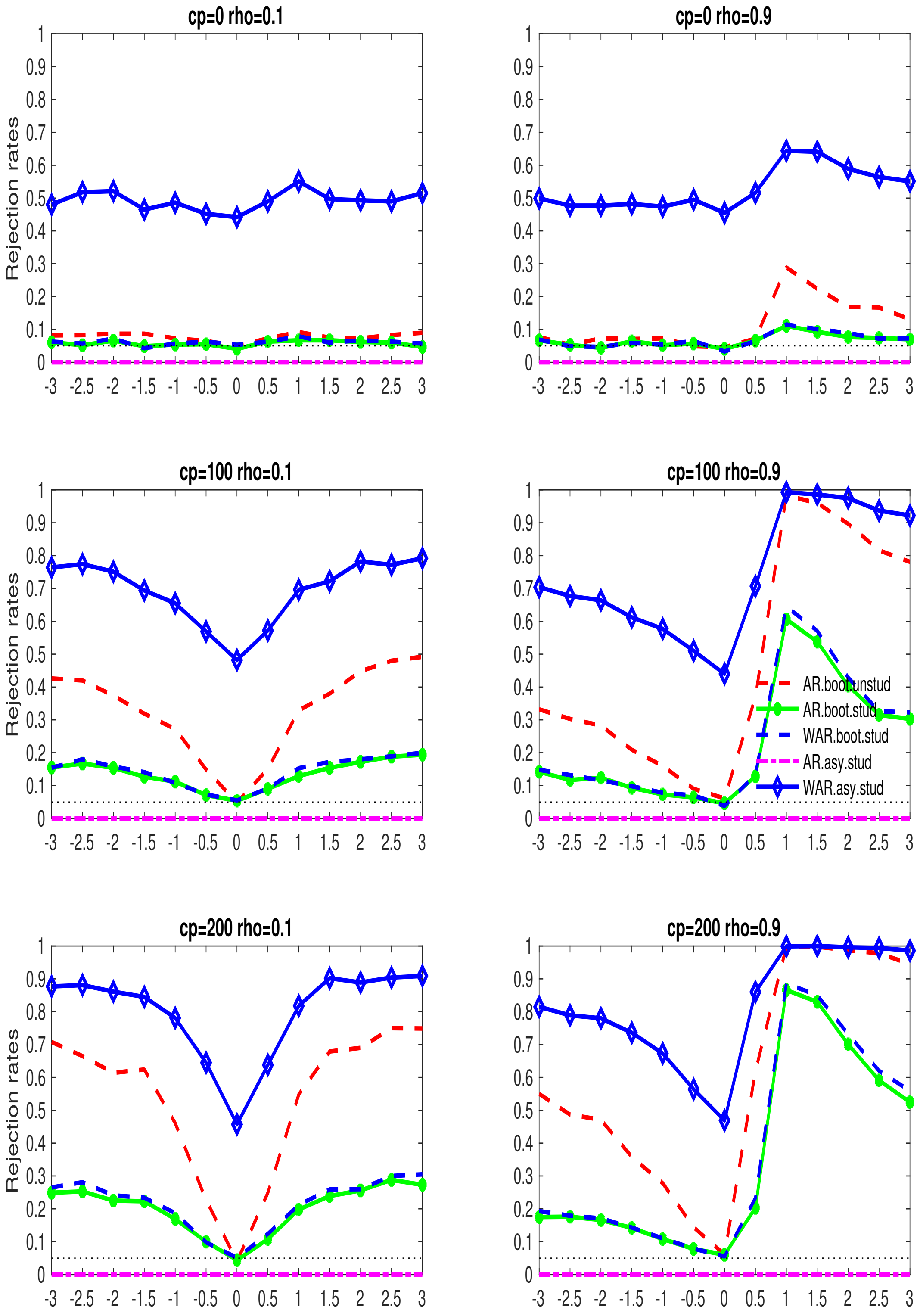


Figure 8: Power of AR-based tests with $d_z = 5$



References

Abadie, A., J. Gu, and S. Shen (2019): "Instrumental Variable Estimation with First Stage Heterogeneity," Discussion paper, Working paper, MIT.

Anderson, T. W., and H. Rubin (1949): "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *Annals of Mathematical Statistics*, 20(1), 46–63.

Andrews, D. W., and P. Guggenberger (2019): "Identification-and singularity-robust inference for moment condition models," *Quantitative Economics*, 10(4), 1703–1746.

Andrews, I. (2016): "Conditional linear combination tests for weakly identified models," *Econometrica*, 84(6), 2155–2182.

——— (2018): "Valid two-step identification-robust confidence sets for GMM," *Review of Economics and Statistics*, 100(2), 337–348.

Andrews, I., and A. Mikusheva (2016): "Conditional inference with a functional nuisance parameter," *Econometrica*, 84(4), 1571–1612.

Andrews, I., J. H. Stock, and L. Sun (2019): "Weak instruments in instrumental variables regression: Theory and practice," *Annual Review of Economics*, 11, 727–753.

Angrist, J., G. Imbens, and A. Krueger (1999): "Jackknife Instrumental Variables Estimates," *Journal of Applied Econometrics*, 14(1), 57–67.

Bekker, P. (1994): "Alternative Approximations to the Distributions of Instrumental Variable Estimators," *Econometrica*, 62(3), 657–681.

Beran, R. (1988): "Prepivoting test statistics: a bootstrap view of asymptotic refinements," *Journal of the American Statistical Association*, 83(403), 687–697.

Bester, C. A., T. G. Conley, and C. B. Hansen (2011): "Inference with dependent data using cluster covariance estimators," *Journal of Econometrics*, 165(2), 137–151.

Brodeur, A., N. Cook, and A. Heyes (2020): "Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics," *American Economic Review*.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008): "Bootstrap-based improvements for inference with clustered errors," *The Review of Economics and Statistics*, 90(3), 414–427.

Cameron, A. C., and D. L. Miller (2015): "A practitioner's guide to cluster-robust inference," *Journal of human resources*, 50(2), 317–372.

Canay, I. A., J. P. Romano, and A. M. Shaikh (2017): "Randomization tests under an approximate symmetry assumption," *Econometrica*, 85(3), 1013–1030.

Canay, I. A., A. Santos, and A. M. Shaikh (2020): "The wild bootstrap with a" small" number of" large" clusters," *Review of Economics and Statistics*, p. Forthcoming.

Chao, J. C., and N. R. Swanson (2005): "Consistent Estimation with a Large Number of Weak Instruments," *Econometrica*, 73(5), 1673–1692.

Chao, J. C., N. R. Swanson, J. A. Hausman, W. K. Newey, and T. Woutersen (2012): "Asymptotic Distribution Of JIVE In A Heteroskedastic IV Regression With Many Instruments," *Econometric Theory*, 28(1), 42–86.

Chernozhukov, V., and C. Hansen (2008a): "Instrumental variable quantile regression: A robust inference approach," *Journal of Econometrics*, 142(1), 379–398.

——— (2008b): "The reduced form: A simple approach to inference with weak instruments," *Economics Letters*, 100(1), 68–71.

Davidson, R., and J. G. MacKinnon (2008): "Bootstrap inference in a linear equation estimated by instrumental variables," *The Econometrics Journal*, 11(3), 443–477.

——— (2010): "Wild bootstrap tests for IV regression," *Journal of Business & Economic Statistics*, 28(1), 128–144.

——— (2014): "Bootstrap confidence sets with weak instruments," *Econometric Reviews*, 33(5-6), 651–675.

Djogbenou, A. A., J. G. MacKinnon, and M. Ø. Nielsen (2019): "Asymptotic theory and wild bootstrap inference with clustered errors," *Journal of Econometrics*, 212(2), 393–412.

FINLAY, K., AND L. M. MAGNUSSON (2009): "Implementing weak-instrument robust tests for a general class of instrumental-variables models," *The Stata Journal*, 9(3), 398–421.

——— (2019): "Two applications of wild bootstrap methods to improve inference in cluster-IV models," *Journal of Applied Econometrics*, 34(6), 911–933.

GUGGENBERGER, P., F. KLEIBERGEN, AND S. MAVROEIDIS (2019): "A more powerful subvector Anderson Rubin test in linear instrumental variables regression," *Quantitative Economics*, 10(2), 487–526.

GUGGENBERGER, P., F. KLEIBERGEN, S. MAVROEIDIS, AND L. CHEN (2012): "On the asymptotic sizes of subset Anderson–Rubin and Lagrange multiplier tests in linear instrumental variables regression," *Econometrica*, 80(6), 2649–2666.

HAGEMANN, A. (2019): "Permutation inference with a finite number of heterogeneous clusters," *arXiv preprint arXiv:1907.01049*.

HALL, P. (1992): "The bootstrap and Edgeworth expansion," in *The bootstrap and Edgeworth expansion*, ed. by P. Hall. Springer-Verlag New York, Inc, New York.

HANSEN, B. E., AND S. LEE (2019): "Asymptotic theory for clustered samples," *Journal of econometrics*, 210(2), 268–290.

HAUSMAN, J. A., W. K. NEWEY, T. WOUTERSEN, J. C. CHAO, AND N. R. SWANSON (2012): "Instrumental variable estimation with heteroskedasticity and many instruments," *Quantitative Economics*, 3(2), 211–255.

HOROWITZ, J. L. (2001): "The bootstrap," in *Handbook of Econometrics*, ed. by J. Heckman, and E. E. Leamer. Elsvier Science, Amsterdam, The Netherlands.

IBRAGIMOV, R., AND U. K. MÜLLER (2010): "t-Statistic based correlation and heterogeneity robust inference," *Journal of Business & Economic Statistics*, 28(4), 453–468.

——— (2016): "Inference with few heterogeneous clusters," *Review of Economics and Statistics*, 98(1), 83–96.

Imbens, G. W., and P. R. Rosenbaum (2005): "Robust, accurate confidence intervals with a weak instrument: quarter of birth and education," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(1), 109–126.

Kaffo, M., and W. Wang (2017): "On bootstrap validity for specification testing with many weak instruments," *Economics Letters*, 157, 107–111.

Kleibergen, F. (2002): "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression," *Econometrica*, 70(5), 1781–1803.

——— (2005): "Testing Parameters in GMM Without Assuming That They are Identified," *Econometrica*, 73, 1103–1124.

MacKinnon, J. G., M. Ø. Nielsen, and M. D. Webb (2019): "Wild bootstrap and asymptotic inference with multiway clustering," *Journal of Business & Economic Statistics*, pp. 1–15.

MacKinnon, J. G., and M. D. Webb (2017): "Wild bootstrap inference for wildly different cluster sizes," *Journal of Applied Econometrics*, 32(2), 233–254.

Moreira, H., and M. J. Moreira (2019): "Optimal two-sided tests for instrumental variables regression with heteroskedastic and autocorrelated errors," *Journal of Econometrics*, 213(2), 398–433.

Moreira, M. J. (2003): "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica*, 71(4), 1027–1048.

Moreira, M. J., J. Porter, and G. Suarez (2009): "Bootstrap Validity for the Score Test when Instruments may be Weak," *Journal of Econometrics*, 149(1), 52–64.

Muralidharan, K., P. Niehaus, and S. Sukhtankar (2016): "Building state capacity: Evidence from biometric smartcards in India," *American Economic Review*, 106(10), 2895–2929.

Nagar, A. L. (1959): "The Bias and Moment Matrix of the Genaralizea k-Class Estimators of the Parameters in Simultaneous Equations," *Econometrica*, 27, 575–595.

Newey, W. K., and F. Windmeijer (2009): “Generalized method of moments with many weak moment conditions,” *Econometrica*, 77(3), 687–719.

Olea, J. L. M., and C. Pflueger (2013): “A robust test for weak instruments,” *Journal of Business & Economic Statistics*, 31(3), 358–369.

Phillips, G., and C. Hale (1977): “The Bias of Instrumental Variable Estimators of Simultaneous Equation Systems,” *International Economic Review*, 18(1), 219–228.

Roodman, D., M. Ø. Nielsen, J. G. MacKinnon, and M. D. Webb (2019): “Fast and wild: Bootstrap inference in Stata using boottest,” *The Stata Journal*, 19(1), 4–60.

Rosenbaum, P. R. (1996): “Identification of causal effects using instrumental variables: Comment,” *Journal of the American Statistical Association*, 91(434), 465–468.

Rothenberg, T. (1984): “Approximating the Distributions of Econometric Estimators and Test Statistics. Ch. 14 in: Handbook of Econometrics, vol 2, ed. Z. Griliches and M. Intriligator,” .

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

Wang, W. (2020): “On the inconsistency of nonparametric bootstraps for the subvector Anderson-Rubin test,” *Economics Letters*, p. 109157.

Wang, W., and F. Doko Tchatoka (2018): “On Bootstrap inconsistency and Bonferroni-based size-correction for the subset Anderson–Rubin test under conditional homoskedasticity,” *Journal of Econometrics*, 207(1), 188–211.

Wang, W., and M. Kaffo (2016): “Bootstrap inference for instrumental variable models with many weak instruments,” *Journal of Econometrics*, 192(1), 231–268.

Young, A. (2020): “Consistency without inference: Instrumental variables in practical application,” Discussion paper, London School of Economics.

Appendices

A Proofs

Proof of Theorem 3.1 Let $\mathbb{S} \equiv \mathbf{R}^{d_z \times d_x} \times \mathbf{R}^{d_z \times d_z} \times \otimes_{j \in J} \mathbf{R}^{d_z}$ and write an element $s \in \mathbb{S}$ by $s = (s_1, s_2, \{s_{3,j} : j \in J\})$ where $s_{3,j} \in \mathbf{R}^{d_z}$ for any $j \in J$. Define the function T^u_w: $\mathbb{S} \to \mathbf{R}$ to be given by

$$T^u_w(s) \equiv \left| c' \left(s_1' s_2^{-1} s_1 \right)^{-1} s_1' s_2^{-1} \left(\sum_{j \in J} s_{3,j} \right) \right| \tag{A.1}$$

for any $s \in \mathbb{S}$ such that s_2 and $s_1' s_2^{-1} s_1$ are invertible, and let $T^u_w(s) = 0$ otherwise. We also identify any $(g_1, ..., g_q) = g \in \mathbf{G} = \{-1, 1\}^q$ with an action on $s \in \mathbb{S}$ given by $gs = (s_1, s_2, \{g_j s_{3,j} : j \in J\})$. For any $s \in \mathbb{S}$ and $\mathbf{G'} \subseteq \mathbf{G}$, denote the ordered values of $\{T^u_w(gs) : g \in \mathbf{G'}\}$ by

$$T_w^{u(1)}(s|\mathbf{G'}) \leq \ldots \leq T_w^{u(|\mathbf{G'}|)}(s|\mathbf{G'}). \tag{A.2}$$

Given this notation we can define the statistics $S_n, \widehat{S}_n \in \mathbb{S}$ as

$$\begin{aligned} S_n &\equiv \left(\widehat{Q}_{\tilde{Z}X,n}, \widehat{Q}_{\tilde{Z}\tilde{Z},n}, \left\{ \frac{\sqrt{n_j}}{\sqrt{n}} \frac{1}{\sqrt{n_j}} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} \epsilon_{i,j} : j \in J \right\} \right), \\ \widehat{S}_n &\equiv \left(\widehat{Q}_{\tilde{Z}X,n}, \widehat{Q}_{\tilde{Z}\tilde{Z},n}, \left\{ \frac{\sqrt{n_j}}{\sqrt{n}} \frac{1}{\sqrt{n_j}} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} \hat{\epsilon}^r_{i,j}(\lambda) : j \in J \right\} \right). \end{aligned} \tag{A.3}$$

Let A_n denote the event

$$A_n \equiv I \left\{ \widehat{Q}_{\tilde{Z}X,n} \text{ is of full rank value and } \widehat{Q}_{\tilde{Z}\tilde{Z},n} \text{ is invertible} \right\}, \tag{A.4}$$

and note that whenever $A_n = 1$ and $H_0^c : c'\beta = \lambda$ is true, the Frisch-Waugh-Lovell theorem implies that

$$\begin{aligned} W_{U,n}(\lambda) &= \left| \sqrt{n} \left(c' \hat{\beta}_n - \lambda \right) \right| = \left| \sqrt{n} c' \left(\hat{\beta}_n - \beta \right) \right| \\ &= \left| c' \left(\widehat{Q}'_{\tilde{Z}X,n} \widehat{Q}^{-1}_{\tilde{Z}\tilde{Z},n} \widehat{Q}_{\tilde{Z}X,n} \right)^{-1} \widehat{Q}'_{\tilde{Z}X,n} \widehat{Q}^{-1}_{\tilde{Z}\tilde{Z},n} \sum_{j \in J} \frac{1}{\sqrt{n}} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} \epsilon_{i,j} \right| \\ &= T^u_w(S_n). \end{aligned} \tag{A.5}$$

Similarly, we have for any action $g \in \mathbf{G}$ that

$$\begin{aligned} W^*_{U,n}(\lambda, g) &= \left| \sqrt{n} c' \left(\hat{\beta}^*_n(g) - \hat{\beta}^r_n \right) \right| \\ &= \left| c' \left(\widehat{Q}'_{\tilde{Z}X,n} \widehat{Q}^{-1}_{\tilde{Z}\tilde{Z},n} \widehat{Q}_{\tilde{Z}X,n} \right)^{-1} \widehat{Q}'_{\tilde{Z}X,n} \widehat{Q}^{-1}_{\tilde{Z}\tilde{Z},n} \sum_{j \in J} \frac{1}{\sqrt{n}} \sum_{i \in I_{n,j}} g_j \tilde{Z}_{i,j} \hat{\epsilon}^r_{i,j}(\lambda) \right| \\ &= T^u_w(g\widehat{S}_n). \end{aligned} \tag{A.6}$$

Therefore, for any $x \in \mathbf{R}$ letting $\lceil x \rceil$ denote the smallest integer larger than x and $k^* \equiv \lceil |\mathbf{G}|(1-\alpha) \rceil$, we obtain from (A.5)-(A.6) that

$$I\left\{ W_{U,n}(\lambda) > \hat{c}^w_{u,n}(1-\alpha) \right\} = I\left\{ T^u_w(S_n) > T^{u(k^*)}_w(\widehat{S}_n|\mathbf{G}) \right\}. \tag{A.7}$$

In addition, Assumption 1 implies that

$$\liminf_{n \to \infty} P\{A_n = 1\} = 1. \tag{A.8}$$

Furthermore, let $\ell \in \mathbf{G}$ correspond to the identity action, i.e., $\ell \equiv (1, ..., 1) \in \mathbf{R}^q$, and similarly define $-\ell \equiv (-1, ..., -1) \in \mathbf{R}^q$. Then note that since $T^u_w(-\ell\widehat{S}_n) = T^u_w(\ell\widehat{S}_n)$, we obtain from (A.6) that

$$\begin{aligned} T^u_w\left(-\ell\widehat{S}_n\right) &= T^u_w\left(\ell\widehat{S}_n\right) \\ &= \left| c' \left(\widehat{Q}'_{\tilde{Z}X,n} \widehat{Q}^{-1}_{\tilde{Z}\tilde{Z},n} \widehat{Q}_{\tilde{Z}X,n} \right)^{-1} \widehat{Q}'_{\tilde{Z}X,n} \widehat{Q}^{-1}_{\tilde{Z}\tilde{Z},n} \sum_{j \in J} \frac{1}{\sqrt{n}} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} \left(y_{i,j} - X'_{i,j}\hat{\beta}^r_n - W'_{i,j}\hat{\gamma}^r_n \right) \right|, \\ &= \left| c' \left(\widehat{Q}'_{\tilde{Z}X,n} \widehat{Q}^{-1}_{\tilde{Z}\tilde{Z},n} \widehat{Q}_{\tilde{Z}X,n} \right)^{-1} \widehat{Q}'_{\tilde{Z}X,n} \widehat{Q}^{-1}_{\tilde{Z}\tilde{Z},n} \sum_{j \in J} \frac{1}{\sqrt{n}} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} \left(y_{i,j} - X'_{i,j}\hat{\beta}^r_n \right) \right|, \\ &= \left| \sqrt{n} c' (\hat{\beta}_n - \hat{\beta}^r_n) \right| = T^u_w\left(S_n\right), \end{aligned} \tag{A.9}$$

where the third equality follows from $\sum_{j \in J} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} W'_{i,j} = 0$. (A.9) implies that if $k^* \equiv \lceil |\mathbf{G}|(1-\alpha) \rceil > |\mathbf{G}| - 2$, then $I\{T^u_w(S_n) > T^{u(k^*)}_w(\widehat{S}_n|\mathbf{G})\} = 0$, and this gives the upper bound in Theorem 3.1. We therefore assume that $k^* \equiv \lceil |\mathbf{G}|(1-\alpha) \rceil \leq |\mathbf{G}| - 2$, in which case

$$\begin{aligned} \limsup_{n \to \infty} E\left[\phi_n\left(W_{U,n}(\lambda)\right)\right] &= \limsup_{n \to \infty} P\{T^u_w(S_n) > T^{u(k^*)}_w(\widehat{S}_n|\mathbf{G}); A_n = 1\} \\ &= \limsup_{n \to \infty} P\{T^u_w(S_n) > T^{u(k^*)}_w(\widehat{S}_n|\mathbf{G} \setminus \{\pm\ell\}); A_n = 1\} \\ &\leq \limsup_{n \to \infty} P\{T^u_w(S_n) \geq T^{u(k^*)}_w(\widehat{S}_n|\mathbf{G} \setminus \{\pm\ell\}); A_n = 1\}, \end{aligned} \tag{A.10}$$

where the first equality follows from (A.7), the second equality from (A.9) and $k^* \leq |\mathbf{G}| - 2$, and the final inequality follows by set inclusion.

Then, to examine the right hand side of (A.10), first note that by Assumptions 1-2, Assumption 3(ii) and the continuous mapping theorem we have

$$\left(\widehat{Q}_{\tilde{Z}X,n},\widehat{Q}_{\tilde{Z}\tilde{Z},n},\left\{\frac{\sqrt{n_j}}{\sqrt{n}}\frac{1}{\sqrt{n_j}}\sum_{i\in I_{n,j}}\tilde{Z}_{i,j}\epsilon_{i,j}:j\in J\right\}\right)\xrightarrow{d}\left(\bar{a}Q_{\tilde{Z}X},Q_{\tilde{Z}\tilde{Z}},\left\{\sqrt{\xi_j}\mathcal{Z}_j:j\in J\right\}\right)\equiv S, \tag{A.11}$$

where $\xi_j>0$ for all $j\in J$ by Assumption 2(i), and $Q_{\tilde{Z}\tilde{Z}}$ denotes the limit of $\widehat{Q}_{\tilde{Z}\tilde{Z},n}$. We further note that whenever $A_n=1$, for every $g\in\mathbf{G}$,

$$\begin{aligned}
&\left|T_w^u(gS_n)-T_w^u(g\widehat{S}_n)\right|\\
\leq\ &\left|c'\left(\widehat{Q}'_{\tilde{Z}X,n}\widehat{Q}^{-1}_{\tilde{Z}\tilde{Z},n}\widehat{Q}_{\tilde{Z}X,n}\right)^{-1}\widehat{Q}'_{\tilde{Z}X,n}\widehat{Q}^{-1}_{\tilde{Z}\tilde{Z},n}\sum_{j\in J}\frac{n_j}{n}\frac{1}{n_j}\sum_{i\in I_{n,j}}g_j\tilde{Z}_{i,j}X'_{i,j}\sqrt{n}(\beta-\hat{\beta}^r_n)\right|\\
&+\left|c'\left(\widehat{Q}'_{\tilde{Z}X,n}\widehat{Q}^{-1}_{\tilde{Z}\tilde{Z},n}\widehat{Q}_{\tilde{Z}X,n}\right)^{-1}\widehat{Q}'_{\tilde{Z}X,n}\widehat{Q}^{-1}_{\tilde{Z}\tilde{Z},n}\sum_{j\in J}\frac{n_j}{n}\frac{1}{n_j}\sum_{i\in I_{n,j}}g_j\tilde{Z}_{i,j}W'_{i,j}\sqrt{n}(\gamma-\hat{\gamma}^r_n)\right|.
\end{aligned}\tag{A.12}$$

Note that whenever $c'\beta=\lambda$ it follows from Assumption 1 and Amemiya (1985, Eq.(1.4.5)) that $\sqrt{n}(\hat{\beta}^r_n-\beta)$ and $\sqrt{n}(\hat{\gamma}^r_n-\gamma)$ are bounded in probability. In addition, we have $\sum_{i\in I_{n,j}}\tilde{Z}_{i,j}W'_{i,j}/n_j=o_P(1)$ by using the same argument as in Lemma A.2 of Canay et al. (2020). Therefore,

$$\begin{aligned}
&\limsup_{n\to\infty}P\left\{\left|c'\left(\widehat{Q}'_{\tilde{Z}X,n}\widehat{Q}^{-1}_{\tilde{Z}\tilde{Z},n}\widehat{Q}_{\tilde{Z}X,n}\right)^{-1}\widehat{Q}'_{\tilde{Z}X,n}\widehat{Q}^{-1}_{\tilde{Z}\tilde{Z},n}\sum_{j\in J}\frac{n_j}{n}\frac{1}{n_j}\sum_{i\in I_{n,j}}g_j\tilde{Z}_{i,j}W'_{i,j}\sqrt{n}(\gamma-\hat{\gamma}^r_n)\right|>\epsilon;A_n=1\right\}\\
&=0.
\end{aligned}\tag{A.13}$$

Moreover, Assumption 3(ii) yields for any $\epsilon>0$ that

$$\begin{aligned}
&\limsup_{n\to\infty}P\left\{\left|c'\left(\widehat{Q}'_{\tilde{Z}X,n}\widehat{Q}^{-1}_{\tilde{Z}\tilde{Z},n}\widehat{Q}_{\tilde{Z}X,n}\right)^{-1}\widehat{Q}'_{\tilde{Z}X,n}\widehat{Q}^{-1}_{\tilde{Z}\tilde{Z},n}\sum_{j\in J}\frac{n_j}{n}\frac{1}{n_j}\sum_{i\in I_{n,j}}g_j\tilde{Z}_{i,j}X'_{i,j}\sqrt{n}(\beta-\hat{\beta}^r_n)\right|>\epsilon;A_n=1\right\}\\
=\ &\limsup_{n\to\infty}P\left\{\left|c'\left(\bar{a}^2Q'_{\tilde{Z}X}Q^{-1}_{\tilde{Z}\tilde{Z}}Q_{\tilde{Z}X}\right)^{-1}\bar{a}Q'_{\tilde{Z}X}Q^{-1}_{\tilde{Z}\tilde{Z}}\sum_{j\in J}\xi_jg_ja_jQ_{\tilde{Z}X}\sqrt{n}(\beta-\hat{\beta}^r_n)\right|>\epsilon;A_n=1\right\}\\
=\ &\limsup_{n\to\infty}P\left\{\left|c'\sum_{j\in J}\frac{\xi_jg_ja_j}{\bar{a}}\sqrt{n}(\beta-\hat{\beta}^r_n)\right|>\epsilon;A_n=1\right\}\\
=\ &\limsup_{n\to\infty}P\left\{\left|\sum_{j\in J}\frac{\xi_jg_ja_j}{\bar{a}}\sqrt{n}(c'\beta-c'\hat{\beta}^r_n)\right|>\epsilon;A_n=1\right\}=0,
\end{aligned}\tag{A.14}$$

where $\bar{a} = \sum_{j \in J} \xi_j a_j$, and the last equality holds because $c'\hat{\beta}_n^r = \lambda$ under $H_0^c : c'\beta = \lambda$.

Notice that $T_w^u(g\widehat{S}_n) = T_w^u(gS_n)$ whenever $A_n = 0$ as we have defined $T_w^u(s) = 0$ for any $s = (s_1, s_2, \{s_{3,j} : j \in J\})$ whenever s_2 or $s_1' s_2^{-1} s_1$ is not invertible. Therefore, results in (A.12), (A.13), and (A.14) imply that $T_w^u(g\widehat{S}_n) = T_w^u(gS_n) + o_P(1)$ for any $g \in \mathbf{G}$. We thus obtain from (A.11) that

$$\left(T_w^u(S_n), \left\{T_w^u(g\widehat{S}_n) : g \in \mathbf{G}\right\}\right) \xrightarrow{d} \left(T_w^u(S), \{T_w^u(gS) : g \in \mathbf{G}\}\right). \tag{A.15}$$

Moreover, since $\liminf_{n\to\infty} P\{A_n = 1\} = 1$, it follows that $\left(T_w^u(S_n), A_n, \{T_w^u(g\widehat{S}_n) : g \in \mathbf{G}\}\right)$ converge jointly as well. Hence, Portmanteau's theorem implies that

$$\begin{aligned} &\limsup_{n\to\infty} P\{T_w^u(S_n) \geq T_w^{u(k^*)}(\widehat{S}_n | \mathbf{G} \setminus \{\pm\ell\}); A_n = 1\} \\ \leq\ & P\{T_w^u(S) \geq T_w^{u(k^*)}(S | \mathbf{G} \setminus \{\pm\ell\})\} = P\{T_w^u(S) > T_w^{u(k^*)}(S | \mathbf{G} \setminus \{\pm\ell\})\}, \end{aligned} \tag{A.16}$$

where the equality follows from $P\{T_w^u(S) = T_w^u(gS)\} = 0$ for all $g \in \mathbf{G} \setminus \{\pm\ell\}$ since the covariance matrix of $\mathcal{Z}_j$ is full rank for all $j \in J$, and the limit of $\left(\widehat{Q}_{\tilde{Z}X,n}' \widehat{Q}_{\tilde{Z}\tilde{Z},n}^{-1} \widehat{Q}_{\tilde{Z}X,n}\right)^{-1} \widehat{Q}_{\tilde{Z}X,n}' \widehat{Q}_{\tilde{Z}\tilde{Z},n}^{-1}$ is of full rank by Assumption 1.

Finally, since $T_w^u(\ell S) = T_w^u(-\ell S)$, we obtain that $T_w^u(S) > T_w^{u(k^*)}(S | \mathbf{G} \setminus \{\pm\ell\})$ if and only if $T_w^u(S) > T_w^{u(k^*)}(S | \mathbf{G})$, which together with (A.10) and (A.16) yields

$$\limsup_{n\to\infty} E\left[\phi_n(W_{U,n}(\lambda))\right] \leq P\{T_w^u(S) > T_w^{u(k^*)}(S | \mathbf{G} \setminus \{\pm\ell\})\} \;=\; P\{T_w^u(S) > T_w^{u(k^*)}(S | \mathbf{G})\} \leq \alpha, \tag{A.17}$$

where the final inequality follows by $gS \overset{d}{=} S$ for all $g \in \mathbf{G}$ and the properties of randomization tests. This completes the proof of the upper bound in the statement of the Theorem.

For the lower bound, first note that $k^* \equiv \lceil |\mathbf{G}|(1-\alpha) \rceil > |\mathbf{G}| - 2$ implies that $\alpha - \frac{1}{2^{q-1}} \leq 0$, in which case the result trivially follows. Now assume $k^* \equiv \lceil |\mathbf{G}|(1-\alpha) \rceil \leq |\mathbf{G}| - 2$ and note that

$$\begin{aligned} \limsup_{n\to\infty} E\left[\phi_n\left(W_{U,n}(\lambda)\right)\right] &\geq \liminf_{n\to\infty} P\{T_w^u(S_n) > T_w^{u(k^*)}(S_n | \mathbf{G})\} \\ &\geq P\{T_w^u(S) > T_w^{u(k^*)}(S | \mathbf{G})\} \\ &\geq P\{T_w^u(S) > T_w^{u(k^*+2)}(S | \mathbf{G})\} + P\{T_w^u(S) = T_w^{u(k^*+2)}(S | \mathbf{G})\} \\ &\geq \alpha - \frac{1}{2^{q-1}}, \end{aligned} \tag{A.18}$$

where the first inequality follows from (A.7), the second inequality follows from Portmanteau's theorem, the third inequality holds because $P\{T_w^{u(\mathbf{z}+2)}(S|\mathbf{G}) > T_w^{u(\mathbf{z})}(S|\mathbf{G})\} = 1$ for any integer $\mathbf{z} \leq |\mathbf{G}| - 2$ by (A.1) and Assumption 2(i)-(ii), and the last equality follows from noticing that $k^* + 2 = \lceil |\mathbf{G}|((1-\alpha) + 2/|\mathbf{G}|)\rceil = \lceil |\mathbf{G}|(1-\alpha')\rceil$ with $\alpha' = \alpha - \frac{1}{2^{q-1}}$ and the properties of randomization tests. This completes the proof of the lower bound.

The proof for the studentized wild bootstrap test follows similar arguments as those for the unstudentized wild bootstrap test and the arguments in the proof of Theorem 3.3 in Canay et al. (2020), and is thus omitted. ■

Proof of Theorem 3.2 The proof follows similar arguments as those in Theorem 3.1, and thus we keep exposition more concise. Let $\mathbb{S} \equiv \otimes_{j\in J}\mathbf{R}^{d_z}$ and write an element $s \in \mathbb{S}$ by $s = \{s_j : j \in J\}$ where $s_j \in \mathbf{R}^{d_z}$ for any $j \in J$. Define the function T_{ar}^u: $\mathbb{S} \to \mathbf{R}$ to be given by

$$T_{ar}^u(s) \equiv \left\| \sum_{j\in J} s_j \right\|^2 . \tag{A.19}$$

Given this notation we can define the statistics $S_n, \widehat{S}_n \in \mathbb{S}$ as

$$S_n \equiv \left\{ \frac{\sqrt{n_j}}{\sqrt{n}} \frac{1}{\sqrt{n_j}} \sum_{i\in I_{n,j}} \tilde{Z}_{i,j}\epsilon_{i,j} : j \in J \right\}, \widehat{S}_n \equiv \left\{ \frac{\sqrt{n_j}}{\sqrt{n}} \frac{1}{\sqrt{n_j}} \sum_{i\in I_{n,j}} \tilde{Z}_{i,j}\hat{\epsilon}_{i,j}(\beta_0) : j \in J \right\}. \tag{A.20}$$

Note that by the Frisch-Waugh-Lovell theorem,

$$AR_{U,n}(\beta_0) \quad = \quad \left\| \sum_{j\in J} \frac{1}{\sqrt{n}} \sum_{i\in I_{n,j}} \tilde{Z}_{i,j}\epsilon_{i,j} \right\|^2 = T_{ar}^u(S_n). \tag{A.21}$$

Similarly, we have for any action $g \in \mathbf{G}$ that

$$AR_{U,n}^*(\beta_0, g) \quad = \quad \left\| \sum_{j\in J} \frac{1}{\sqrt{n}} \sum_{i\in I_{n,j}} g_j \tilde{Z}_{i,j}\hat{\epsilon}_{i,j}(\beta_0) \right\|^2 = T_{ar}^u(g\widehat{S}_n). \tag{A.22}$$

Therefore, for any $x \in \mathbf{R}$ letting $\lceil x \rceil$ denote the smallest integer larger than x and $k^* \equiv \lceil |\mathbf{G}|(1-\alpha)\rceil$, we obtain from (A.21)-(A.22) that

$$I\left\{AR_{U,n}(\beta_0) > \hat{c}_{u,n}^{ar}(1-\alpha)\right\} = I\left\{T_{ar}^u(S_n) > T_{ar}^{u(k^*)}(\widehat{S}_n|\mathbf{G})\right\}. \tag{A.23}$$

Furthermore, similar to the arguments in the proof of Theorem 3.1, we have

$$\begin{aligned} T\left(-\ell\widehat{S}_n\right) &= T\left(\ell\widehat{S}_n\right) = \left\| \sum_{j\in J} \frac{1}{\sqrt{n}} \sum_{i\in I_{n,j}} \tilde{Z}_{i,j}\left(y_{i,j} - X'_{i,j}\beta_0 - W'_{i,j}\hat{\gamma}^r_n(\beta_0)\right) \right\|^2, \\ &= \left\| \sum_{j\in J} \frac{1}{\sqrt{n}} \sum_{i\in I_{n,j}} \tilde{Z}_{i,j}\left(\epsilon_{i,j} + W'_{i,j}(\gamma - \hat{\gamma}^r_n(\beta_0))\right) \right\|^2 = T\left(S_n\right). \qquad \text{(A.24)} \end{aligned}$$

(A.24) implies that if $k^* \equiv \lceil |\mathbf{G}|(1-\alpha)\rceil > |\mathbf{G}| - 2$, then $I\{T(S_n) > T^{(k^*)}(\widehat{S}_n|\mathbf{G})\} = 0$, and this gives the upper bound in Theorem 3.2. We therefore assume that $k^* \equiv \lceil |\mathbf{G}|(1-\alpha)\rceil \leq |\mathbf{G}| - 2$, in which case

$$\begin{aligned} \limsup_{n\to\infty} E\left[\phi_n\left(AR_{U,n}(\beta_0)\right)\right] &= \limsup_{n\to\infty} P\{T^u_{ar}(S_n) > T^{u(k^*)}_{ar}(\widehat{S}_n|\mathbf{G})\} \\ &= \limsup_{n\to\infty} P\{T^u_{ar}(S_n) > T^{u(k^*)}_{ar}(\widehat{S}_n|\mathbf{G}\setminus\{\pm\ell\})\} \\ &\leq \limsup_{n\to\infty} P\{T^u_{ar}(S_n) \geq T^{u(k^*)}_{ar}(\widehat{S}_n|\mathbf{G}\setminus\{\pm\ell\})\}. \qquad \text{(A.25)} \end{aligned}$$

Then, to examine the right hand side of (A.25), first note that by Assumption 2(i) and the continuous mapping theorem we have

$$\left\{\frac{\sqrt{n_j}}{\sqrt{n}}\frac{1}{\sqrt{n_j}}\sum_{i\in I_{n,j}} \tilde{Z}_{i,j}\epsilon_{i,j} : j \in J\right\} \xrightarrow{d} \left\{\sqrt{\xi_j}\mathcal{Z}_j : j\in J\right\} \equiv S, \qquad \text{(A.26)}$$

where $\xi_j > 0$ for all $j \in J$ by Assumption 2(ii). We further note that for every $g \in \mathbf{G}$,

$$\begin{aligned} g\widehat{S}_n &= \left\{ g_j \frac{1}{\sqrt{n}} \sum_{i\in I_{n,j}} \tilde{Z}_{i,j}\epsilon_{i,j} - \frac{g_j}{n}\sum_{i\in I_{n,j}} \tilde{Z}_{i,j}W'_{i,j}\sqrt{n}(\hat{\gamma}^r_n(\beta_0) - \gamma) : j\in J \right\} \\ &= \left\{ g_j \frac{1}{\sqrt{n}} \sum_{i\in I_{n,j}} \tilde{Z}_{i,j}\epsilon_{i,j} + o_P(1) : j \in J \right\}, \qquad \text{(A.27)} \end{aligned}$$

by Assumption 2(iii), which implies that

$$T^u_{ar}(g\widehat{S}_n) = \left\| \sum_{j\in J} g_j \frac{1}{\sqrt{n}} \sum_{i\in I_{n,j}} \tilde{Z}_{i,j}\epsilon_{i,j} + o_P(1) \right\|^2 = T^u_{ar}(gS_n) + o_P(1). \qquad \text{(A.28)}$$

We thus obtain from results in (A.26)-(A.28) and the continuous mapping theorem that

$$\left(T^u_{ar}(S_n), \left\{T^u_{ar}(g\widehat{S}_n) : g\in\mathbf{G}\right\}\right) \xrightarrow{d} \left(T^u_{ar}(S), \{T^u_{ar}(gS) : g\in\mathbf{G}\}\right). \qquad \text{(A.29)}$$

Hence, Portmanteau's theorem implies that

$$\begin{aligned}&\limsup_{n\to\infty} P\{T^u_{ar}(S_n) \geq T^{u(k^*)}_{ar}(\widehat{S}_n|\mathbf{G}\setminus\{\pm\ell\})\}\\ \leq\ & P\{T^u_{ar}(S) \geq T^{u(k^*)}_{ar}(S|\mathbf{G}\setminus\{\pm\ell\})\} = P\{T^u_{ar}(S) > T^{u(k^*)}_{ar}(S|\mathbf{G}\setminus\{\pm\ell\})\}, \qquad \text{(A.30)}\end{aligned}$$

where the equality follows from $P\{T(S) = T(gS)\} = 0$ for all $g \in \mathbf{G}\setminus\{\pm\ell\}$ since the covariance matrix of $\mathcal{Z}_j$ is full rank for all $j \in J$. Finally, using arguments similar to those in the proof of Theorem 3.1, we obtain

$$\limsup_{n\to\infty} E\left[\phi_n(AR_{U,n}(\beta_0))\right] \leq P\{T^u_{ar}(S) > T^{u(k^*)}_{ar}(S|\mathbf{G}\setminus\{\pm\ell\})\} \;=\; P\{T^u_{ar}(S) > T^{u(k^*)}_{ar}(S|\mathbf{G})\} \leq \alpha, \qquad \text{(A.31)}$$

where the final inequality follows by $gS \overset{d}{=} S$ for all $g \in \mathbf{G}$ and the properties of randomization tests. This completes the proof of the upper bound in the statement of the Theorem. The proof of the lower bound follows the same arguments as those for Theorem 3.1.

The proof for the studentized AR test follows similar arguments as those for the unstudentized version, and we keep exposition concise. Define the function $T_{ar} : \mathbb{S} \to \mathbf{R}$ to be given by

$$T_{ar}(s) \equiv \left\| \left(\sum_{j\in J} s_j s_j'\right)^{-1/2} \sum_{j\in J} s_j \right\|^2, \qquad \text{(A.32)}$$

for any $s \in \mathbb{S}$ such that $\sum_{j\in J} s_j s_j'$ is invertible, and set $T_{ar}(s) = 0$ whenever $\sum_{j\in J} s_j s_j'$ is not invertible. We set $A_n \in \mathbf{R}$ to equal

$$A_n \equiv I\left\{\sum_{j\in J} \widehat{S}_{n,j}\widehat{S}_{n,j}' \text{ is invertible}\right\}, \qquad \text{(A.33)}$$

where $\widehat{S}_{n,j} = \frac{1}{\sqrt{n}}\sum_{i\in I_{n,j}} \tilde{Z}_{i,j}\hat{\epsilon}_{i,j}(\beta_0)$, and we have

$$\liminf_{n\to\infty} P\{A_n = 1\} = 1, \qquad \text{(A.34)}$$

which follows from $\left\{\sqrt{\xi_j}\mathcal{Z}_{\epsilon,j} : j \in J\right\}$ being independent and continuously distributed with covariance matrices that are full rank. It follows that whenever $A_n = 1$,

$$(AR_n(\beta_0), \{AR^*_n(\beta_0, g) : g \in \mathbf{G}\}) = (T_{ar}(S_n), \{T_{ar}(gS_n) : g \in \mathbf{G}\}) + o_P(1). \qquad \text{(A.35)}$$

Next, we have

$$\begin{aligned}
& \limsup_{n\to\infty} P\left\{AR_n(\beta_0) > \hat{c}_n^{ar}(1-\alpha)\right\} \\
\leq\ & \limsup_{n\to\infty} P\left\{AR_n(\beta_0) \geq \hat{c}_n^{ar}(1-\alpha); A_n = 1\right\} \\
\leq\ & P\left\{T_{ar}(S) \geq \inf\left\{u \in \mathbf{R} : \frac{1}{|\mathbf{G}|}\sum_{g\in\mathbf{G}} I\{T_{ar}(gS) \leq u\} \geq 1-\alpha\right\}\right\}, \qquad \text{(A.36)}
\end{aligned}$$

where the final inequality follows from (A.34), (A.35), the continuous mapping theorem and Portmanteau's theorem.

Therefore, setting $k^* \equiv \lceil |\mathbf{G}|(1-\alpha)\rceil$, we can then obtain from (A.36) that

$$\begin{aligned}
& \limsup_{n\to\infty} P\left\{AR_n(\beta_0) > \hat{c}_n^{ar}(1-\alpha)\right\} \\
\leq\ & P\left\{T_{ar}(S) > T_{ar}^{(k^*)}(S|\mathbf{G})\right\} + P\left\{T_{ar}(S) = T_{ar}^{(k^*)}(S|\mathbf{G})\right\} \\
\leq\ & \alpha + P\left\{T_{ar}(S) = T_{ar}^{(k^*)}(S|\mathbf{G})\right\}, \qquad \text{(A.37)}
\end{aligned}$$

where the final inequality follows by $gS \overset{d}{=} S$ for all $g \in \mathbf{G}$ and the properties of randomization tests. Furthermore, by applying Lehmann and Romano (2005, Theorem 15.2.2), we obtain

$$P\left\{T_{ar}(S) = T_{ar}^{(k^*)}(S|\mathbf{G})\right\} = E\left[\frac{1}{|\mathbf{G}|}\sum_{g\in\mathbf{G}} I\left\{T_{ar}(gS) = T_{ar}^{(k^*)}(S|\mathbf{G})\right\}\right]. \qquad \text{(A.38)}$$

For any $g = (g_1, ..., g_q) \in \mathbf{G}$ then let $-g = (-g_1, ..., -g_q) \in \mathbf{G}$, and note that $T_{ar}(gS) = T_{ar}(-gS)$ with probability one. However, if $\tilde{g}, g \in \mathbf{G}$ are such that $\tilde{g} \notin \{g, -g\}$, then

$$P\left\{T_{ar}(gS) = T_{ar}(\tilde{g}S)\right\} = 0 \qquad \text{(A.39)}$$

since $\xi_j > 0$ for all $j \in J$ and $\{\mathcal{Z}_j : j \in J\}$ are independent with full rank covariance matrices by Assumption 2(i)-(ii). Hence,

$$\frac{1}{|\mathbf{G}|}\sum_{g\in\mathbf{G}} I\left\{T_{ar}(gS) = T_{ar}^{(k^*)}(S|\mathbf{G})\right\} = \frac{1}{|\mathbf{G}|} \times 2 = \frac{1}{2^{q-1}} \qquad \text{(A.40)}$$

with probability one. The claim of the upper bound in the theorem then follows from (A.37) and (A.40). The proof for the lower bound follows similar arguments as those for the unstudentized AR test and thus are omitted. ■

Proof of Theorem 3.3

The proof for the studentized LM test follows similar arguments as those for the studentized version of the AR rest. Let $\mathbb{S} \equiv \mathbf{R}^{d_z \times d_x} \times \otimes_{j\in J}\mathbf{R}^{d_z}$, and write an element $s \in \mathbb{S}$ by $s = (\{s_{1,j} : j \in J_s\}, \{s_{2,j} : j \in J\})$. We identify any $(g_1, ..., g_q) = g \in \mathbf{G} = \{-1, 1\}^q$ with an action on $s \in \mathbb{S}$ given by $gs = (\{s_{1,j} : j \in J_s\}, \{g_j s_{2,j} : j \in J\})$. We define the function $T_{lm} : \mathbb{S} \to \mathbf{R}$ to be given by

$$T_{lm}(s) \equiv \left\| \left(D(s)' \left(\sum_{j\in J} s_{2,j}s_{2,j}' \right)^{-1} D(s) \right)^{-1/2} D(s)' \left(\sum_{j\in J} s_{2,j}s_{2,j}' \right)^{-1} \sum_{j\in J} s_{2,j} \right\|^2, \quad \text{(A.41)}$$

for any $s \in \mathbb{S}$ such that $\sum_{j\in J} s_{2,j}s_{2,j}'$ and $D(s)' \left(\sum_{j\in J} s_{2,j}s_{2,j}'\right)^{-1} D(s)$ are invertible and set $T_{lm}(s) = 0$ whenever one of the two is not invertible, where

$$\begin{aligned} D(s) &\equiv (D_1(s), ..., D_{d_x}(s)), \\ D_l(s) &\equiv \sum_{j\in J_s} s_{1,j,l} - \left(\sum_{j\in J_s} s_{1,j,l}s_{2,j}' \right) \left(\sum_{j\in J} s_{2,j}s_{2,j}' \right)^{-1} \sum_{j\in J} s_{2,j}, \qquad \text{(A.42)} \end{aligned}$$

for $s_{1,j} = (s_{1,j,1}, ..., s_{1,j,d_x})$ and $l = 1, ..., d_x$.

Furthermore, define the statistic S_n as

$$S_n \equiv \left(\left\{ \frac{1}{n} \sum_{i\in I_{n,j}} \tilde{Z}_{i,j}X_{i,j}' : j \in J_s \right\}, \left\{ \frac{1}{\sqrt{n}} \sum_{i\in I_{n,j}} \tilde{Z}_{i,j}\epsilon_{i,j} : j \in J \right\} \right), \quad \text{(A.43)}$$

Note that for $l = 1, ..., d_x$, we have

$$\begin{aligned} \frac{1}{n} \sum_{i\in I_{n,j}} \tilde{Z}_{i,j}X_{i,j,l} &= \frac{n_j}{n} \left(\frac{1}{n_j} \sum_{i\in I_{n,j}} \tilde{Z}_{i,j}\tilde{Z}_{i,j}'\Pi_{z,j,l} + \frac{1}{n_j} \sum_{i\in I_{n,j}} \tilde{Z}_{i,j}v_{i,j,l} \right) \\ &= \frac{n_j}{n} \left(\frac{1}{n_j} \sum_{i\in I_{n,j}} \tilde{Z}_{i,j}\tilde{Z}_{i,j}'\Pi_{z,j,l} \right) + o_P(1) \\ &\xrightarrow{P} Q_{\tilde{Z}X,j,l}, \qquad \text{(A.44)} \end{aligned}$$

where $Q_{\tilde{Z}X,j,l}$ denotes the l-th column of the $d_z \times d_x$-dimensional matrix $Q_{\tilde{Z}X,j}$, the second equality follows from Assumption 2(i), and the convergence in probability follows from Assumption 3(i). Then, by Assumptions 2 and the continuous mapping theorem we have

$$S_n \xrightarrow{d} \left(\{\xi_j a_j Q_{\tilde{Z}X} : j \in J_s\}, \left\{ \sqrt{\xi_j}\mathcal{Z}_j : j \in J \right\} \right) \equiv S, \quad \text{(A.45)}$$

where $\xi_j > 0$ for all $j \in J$. Also notice that for $l = 1, ..., d_x$,

$$\begin{aligned} \widehat{D}_{l,n}(\beta_0) &= \sum_{j \in J} \left(\frac{1}{n} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} X_{i,j,l} \right) - \left(\sum_{j \in J} \left(\frac{1}{n} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} X_{i,j,l} \right) \left(\frac{1}{\sqrt{n}} \sum_{k \in I_{k,j}} \tilde{Z}_{k,j} \hat{\epsilon}_{k,j}(\beta_0) \right)' \right) \\ &\quad \cdot \left(\sum_{j \in J} \left(\frac{1}{\sqrt{n}} \sum_{i \in I_{i,j}} \tilde{Z}_{i,j} \hat{\epsilon}_{i,j}(\beta_0) \right) \left(\frac{1}{\sqrt{n}} \sum_{k \in I_{k,j}} \tilde{Z}_{k,j} \hat{\epsilon}_{k,j}(\beta_0) \right)' \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i \in I_{i,j}} \tilde{Z}_{i,j} \hat{\epsilon}_{i,j}(\beta_0). \end{aligned} \tag{A.46}$$

Here, we set $A_n \in \mathbf{R}$ to equal

$$A_n \equiv I \left\{ \widehat{D}_n(\beta_0) \text{ is of full rank value and } \widehat{\Omega}_n(\beta_0) \text{ is invertible} \right\}, \tag{A.47}$$

and we have

$$\liminf_{n \to \infty} P\{A_n = 1\} = 1, \tag{A.48}$$

which holds because $\left\{ \sqrt{\xi_j} \mathcal{Z}_j : j \in J \right\}$ are independent and continuously distributed with covariance matrices that are of full rank, and $Q_{\tilde{Z}X,j}$ are of full rank for all $j \in J$, by Assumption 2 and Assumption 3(i).

It follows that whenever $A_n = 1$,

$$(LM_n(\beta_0), \{LM_n^*(\beta_0, g) : g \in \mathbf{G}\}) = (T_{lm}(S_n), \{T_{lm}(gS_n) : g \in \mathbf{G}\}) + o_P(1). \tag{A.49}$$

In what follows, we denote the ordered values of $\{T_{lm}(gs) : g \in \mathbf{G}\}$ by

$$T_{lm}^{(1)}(s|\mathbf{G}) \leq ... \leq T_{lm}^{|\mathbf{G}|}(s|\mathbf{G}). \tag{A.50}$$

Next, we have

$$\begin{aligned} &\limsup_{n \to \infty} P\left\{ LM_n(\beta_0) > \hat{c}_n^{lm}(1-\alpha) \right\} \\ \leq\ &\limsup_{n \to \infty} P\left\{ LM_n(\beta_0) \geq \hat{c}_n^{lm}(1-\alpha); A_n = 1 \right\} \\ \leq\ &P\left\{ T_{lm}(S) \geq \inf \left\{ u \in \mathbf{R} : \frac{1}{|\mathbf{G}|} \sum_{g \in \mathbf{G}} I\{T_{lm}(gS) \leq u\} \geq 1 - \alpha \right\} \right\}, \end{aligned} \tag{A.51}$$

where the final inequality follows from (A.43), (A.45), (A.48), (A.49), the continuous mapping theorem and Portmanteau's theorem. Therefore, setting $k^* \equiv \lceil |\mathbf{G}|(1-\alpha) \rceil$, we can then obtain from (A.51) that

$$\begin{aligned}
&\limsup_{n\to\infty} P\left\{LM_n(\beta_0) > \hat{c}_n^{lm}(1-\alpha)\right\} \\
\leq\ & P\left\{T_{lm}(S) > T_{lm}^{(k^*)}(S|\mathbf{G})\right\} + P\left\{T_{lm}(S) = T_{lm}^{(k^*)}(S|\mathbf{G})\right\} \\
\leq\ & \alpha + P\left\{T_{lm}(S) = T_{lm}^{(k^*)}(S|\mathbf{G})\right\}, \qquad \text{(A.52)}
\end{aligned}$$

where the final inequality follows by $gS \overset{d}{=} S$ for all $g \in \mathbf{G}$ and the properties of randomization tests. Then, using similar arguments as those for the studentized AR test, we obtain

$$P\left\{T_{lm}(S) = T_{lm}^{(k^*)}(S|\mathbf{G})\right\} = \frac{1}{2^{q-1}}. \qquad \text{(A.53)}$$

The claim of the upper bound in the theorem then follows from (A.52) and (A.53). The proof for the lower bound is similar to that for the bootstrap AR test, and thus is omitted.

To prove the result for the CQLR test, we note that

$$\begin{aligned}
& LR_n(\beta_0) \\
=\ & \frac{1}{2}\left\{AR_n(\beta_0) - rk_n(\beta_0) + \sqrt{(AR_n(\beta_0) - rk_n(\beta_0))^2 + 4 \cdot LM_n(\beta_0) \cdot rk_n(\beta_0)}\right\} \\
=\ & \frac{1}{2}\left\{AR_n(\beta_0) - rk_n(\beta_0) + |AR_n(\beta_0) - rk_n(\beta_0)| \sqrt{1 + \frac{4 \cdot LM_n(\beta_0) \cdot rk_n(\beta_0)}{(AR_n(\beta_0) - rk_n(\beta_0))^2}}\right\} \\
=\ & \frac{1}{2}\left\{AR_n(\beta_0) - rk_n(\beta_0) + |AR_n(\beta_0) - rk_n(\beta_0)| \left(1 + 2 \cdot LM_n(\beta_0) \frac{rk_n(\beta_0)}{(AR_n(\beta_0) - rk_n(\beta_0))^2}(1 + o_P(1))\right)\right\} \\
=\ & LM_n(\beta_0) \frac{rk_n(\beta_0)}{rk_n(\beta_0) - AR_n(\beta_0)}(1 + o_P(1)) = LM_n(\beta_0) + o_P(1),
\end{aligned} \qquad \text{(A.54)}$$

where the third equality follows from the mean value expansion $\sqrt{1+x} = 1 + (1/2)(x + o(1))$, the fourth and last equalities follow from $AR_n(\beta_0) - rk_n(\beta_0) < 0$ w.p.a.1 since $AR_n(\beta_0) = O_P(1)$ while $rk_n(\beta_0) \to \infty$ w.p.a.1 under Assumption 3(i). Using arguments similar to those in (A.54), we obtain that for each $g \in \mathbf{G}$,

$$LR_n^*(\beta_0, g) = LM_n^*(\beta_0, g) \frac{rk_n(\beta_0)}{rk_n(\beta_0) - AR_n^*(\beta_0, g)}(1 + o_P(1)) = LM_n^*(\beta_0, g) + o_P(1), \qquad \text{(A.55)}$$

by $AR_n^*(\beta_0, g) - rk_n(\beta_0) < 0$ w.p.a.1 since $AR_n^*(\beta_0, g) = O_P(1)$ for each $g \in \mathbf{G}$. Then, it follows that whenever $A_n = 1$,

$$(LR_n(\beta_0), \{LR_n^*(\beta_0, g) : g \in \mathbf{G}\}) = (T_{lm}(S_n), \{T_{lm}(gS_n) : g \in \mathbf{G}\}) + o_P(1). \qquad \text{(A.56)}$$

Then, we obtain that

$$\begin{aligned}
& \limsup_{n\to\infty} P\left\{LR_n(\beta_0) > \hat{c}_n^{lr}(1-\alpha)\right\} \\
\leq\ & \limsup_{n\to\infty} P\left\{LR_n(\beta_0) \geq \hat{c}_n^{lr}(1-\alpha); A_n = 1\right\} \\
\leq\ & P\left\{T_{lm}(S) \geq \inf\left\{u \in \mathbf{R} : \frac{1}{|\mathbf{G}|}\sum_{g\in\mathbf{G}} I\{T_{lm}(gS) \leq u\} \geq 1-\alpha\right\}\right\}, \qquad \text{(A.57)}
\end{aligned}$$

where the second inequality follows from (A.43), (A.45), (A.48), (A.56), the continuous mapping theorem and Portmanteau's theorem. Finally, the upper and lower bounds for the studentized bootstrap CQLR test follows from the previous arguments for the bootstrap LM test. The proofs for the unstudentized bootstrap LM and CQLR test follow from similar arguments, and thus are omitted. ■