Wild Bootstrap Inference for Instrumental Variables Regressions with Weak and Few Clusters*

Wenjie Wang† and Yichong Zhang‡

September 6, 2022

Abstract

We study the wild bootstrap inference for instrumental variable regressions with a small number of large clusters. We first show that the wild bootstrap Wald test controls size asymptotically up to a small error as long as the parameters of endogenous variables are strongly identified in at least one of the clusters. We further develop a wild bootstrap Anderson-Rubin test for the full-vector inference and show that it controls size asymptotically even under weak identification in all clusters. We illustrate their good performance using simulations and provide an empirical application to a well-known dataset about US local labor markets.

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

JEL codes: C12, C26, C31

1 Introduction

The instrument variable (IV) regression is one of the five most commonly used causal inference methods identified by Angrist and Pischke (2008), and it is often applied with clustered data. For example, Young (2021) analyzes 1,359 IV regressions in 31 papers published by the American Economic Association (AEA), out of which 24 papers account for clustering of observations. Three

*We are grateful to Aureo de Paula, Firmin Doko Tchatoka, Kirill Evdokimov, Qingfeng Liu, Morten Ørregaard Nielsen, Ryo Okui, Kyungchul (Kevin) Song, Naoya Sueishi, Yoshimasa Uematsu, and participants at the 2021 Annual Conference of the International Association for Applied Econometrics, the 2021 Asian Meeting of the Econometric Society, the 2021 China Meeting of the Econometric Society, the 2021 Australasian Meeting of the Econometric Society, the 2021 Econometric Society European Meeting and the 37th Canadian Econometric Study Group Meetings for their valuable comments. Special thanks to James MacKinnon for very insightful discussions. Wang acknowledges the financial support from NTU SUG Grant No.M4082262.SS0 and Singapore Ministry of Education Tier 1 grant RG53/20. Zhang acknowledges the financial support from Singapore Ministry of Education Tier 2 grant under grant MOE2018-T2-2-169 and the Lee Kong Chian fellowship. Any possible errors are our own.

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

‡School of Economics, Singapore Management University. E-mail address: yczhang@smu.edu.sg.

issues arise when running IV regressions with clustered data. First, the strength of IVs may be heterogeneous across clusters with one or two clusters providing the main identification power. Indeed, Young (2021) finds that in the average paper of his AEA samples, with the removal of just one cluster or observation, the first-stage F can decrease by 28%, and 38% of reported 0.05 significant two-stage least squares (TSLS) results can be rendered insignificant at that level. Second, the number of clusters is small in many IV applications. For instance, Acemoglu, Cantoni, Johnson, and Robinson (2011) cluster the standard errors at the country/polity level, resulting in 12-19 clusters, Glitz and Meyersson (2020) cluster at the sectoral level with 16 sectors, and Rogall (2021) clusters at the province or district level with 11 provinces and 30 districts. When the number of clusters is small, any inference procedures that require the number of clusters to diverge to infinity may not be reliable. Third, it is also possible that IVs are weak for all clusters, in which case researchers need to use weak-identification-robust inference methods (Andrews, Stock, and Sun, 2019).

Motivated by these issues, in this paper we study the wild bootstrap inference for IV regressions with a small, and thus, fixed number of clusters. First, we show that a wild bootstrap Wald test, with or without the cluster-robust covariance estimator (CCE), controls size asymptotically up to a small error, as long as there exists at least one strong cluster in which the parameters of endogenous variables are strongly identified. Second, we develop the full-vector inference based on a wild bootstrap Anderson and Rubin (1949, AR) test, which controls size asymptotically up to a small error regardless of instrument strength. Third, we establish conditions under which the wild bootstrap tests have power against local alternatives (e.g., there are at least 5 and 6 strong clusters for the nominal level α equal to 10% and 5%, respectively). Fourth, we show that in the special case with a single endogenous variable and single IV, a wild bootstrap test based on the unstudentized Wald statistic (i.e., the one without CCE) is asymptotically equivalent to a certain wild bootstrap AR test under both null and alternative, implying that in such a case it is fully robust to weak IV. Fifth, we establish the validity result for bootstrapping weak-IV-robust tests other than the AR test under at least one strong cluster.

Our procedure is empirically relevant. First, it enriches practitioners' toolbox by providing a reliable inference for IV regressions with few clusters. Besides the aforementioned examples, the numbers of clusters may also be rather small in studies that estimate the region-wise effects of certain intervention if the partition of clusters is at the state level. We illustrate the usefulness of our bootstrap methods by applying them to the well-known dataset of Autor, Dorn, and Hanson (2013) in the estimation of the effects of Chinese imports on local labor markets in three US Census Bureau-designated regions with 11-16 clusters at the state level. Second, our bootstrap inference is flexible with respect to IV strength: the bootstrap Wald test allows for cluster-level heterogeneity in the first stage, while its AR counterpart is fully robust to weak IVs. Third, different from the analytical inference based on the widely used heteroskedasticity and autocorrelation consistent (HAC) estimators, our approach is agnostic about the within-cluster (weak) dependence structure and thus avoids the use of tuning parameters to estimate the covariance matrix for dependent data.

The contributions in the present paper relate to several strands of literature. First, it is related to the literature on the cluster-robust inference.[1] Djogbenou et al. (2019), MacKinnon et al. (2021), and Menzel (2021) show bootstrap validity under the asymptotic framework with a large number of clusters. However, as emphasized by Ibragimov and Müller (2010, 2016), Bester, Conley, and Hansen (2011), Cameron and Miller (2015), Canay, Romano, and Shaikh (2017), Hagemann (2019a,b, 2020), and Canay, Santos, and Shaikh (2021), many empirical studies motivate an alternative framework in which the number of clusters is small, while the number of observations in each cluster is relatively large. For the inference, we may consider applying the approaches developed by Bester et al. (2011), Hwang (2021), Ibragimov and Müller (2010, 2016), and Canay et al. (2017). However, Bester et al. (2011) and Hwang (2021) require an (asymptotically) equal cluster-level sample size, while Ibragimov and Müller (2010, 2016) and Canay et al. (2017) require strong identification in all clusters. In contrast, our bootstrap Wald tests are more flexible as it does not require an equal cluster size and only needs strong identification in one of the clusters. We also provide the bootstrap AR tests, which are fully robust to weak or partial identification.

Second, we follow Canay et al. (2021) to show the asymptotic equivalence between the wild bootstrap test and a randomization test with sign changes, but complement their results in the following aspects. First, Canay et al. (2021) focus on the linear regression with exogenous regressors and then extend the analysis to a score bootstrap for the GMM estimator. Instead, we focus on extending the wild restricted efficient cluster (WREC) bootstrap advocated by Finlay and Magnusson (2014, 2019), Davidson and MacKinnon (2010), Roodman, Nielsen, MacKinnon, and Webb (2019), and

[1] See Cameron, Gelbach, and Miller (2008), Conley and Taber (2011), Imbens and Kolesar (2016), Abadie, Athey, Imbens, and Wooldridge (2022), Hagemann (2017, 2019a,b, 2020), MacKinnon and Webb (2017), Djogbenou, MacKinnon, and Nielsen (2019), MacKinnon, Nielsen, and Webb (2021), Ferman and Pinto (2019), Hansen and Lee (2019), Menzel (2021), MacKinnon (2021), among others, and MacKinnon, Nielsen, and Webb (2022) for a recent survey.

MacKinnon (2021).[2] Therefore, our procedure cannot be formulated as a score bootstrap in the GMM setting. In fact, in order to follow the WREC procedure, the first stage of our bootstrap has to be carefully designed to ensure its validity with few clusters. Second, we consider Wald statistics based on general k-class IV estimators, including TSLS, bias-adjusted TSLS, limited information maximum likelihood (LIML), and modified LIML estimators as special cases. Third, we establish the local power for the Wald test both with and without CCE under strong identification. The former result is derived based on the Sherman–Morrison–Woodbury formula and new to the literature. We also find that the two types of bootstrap critical values behave rather differently (as summarized in Table 1), and further establish the power superiority of the Wald test with CCE for the empirically prevalent case of testing a single restriction. Fourth, we consider the wild bootstrap AR test, with or without CCE, for the full-vector weak-IV-robust inference, and further establish the local power for the one without CCE.

Third, our paper is related to the literature on weak identification, in which various normal approximation-based inference approaches are available for nonhomoskedastic cases, among them Stock and Wright (2000), Kleibergen (2005), Andrews and Cheng (2012), Andrews (2016), Andrews and Mikusheva (2016), Andrews (2018), Moreira and Moreira (2019), and Andrews and Guggenberger (2019). As Andrews et al. (2019, p.750) remark, an important question concerns the quality of the normal approximations with influential observations or clusters. On the other hand, when implemented appropriately, bootstrap may substantially improve the inference for IV regressions.[3] We complement this literature by establishing bootstrap validity for the weak-IV-robust statistics with few clusters.

Last, we note that although empirical applications often involve settings with substantial first-stage heterogeneity, related econometric literature remains rather sparse. Abadie, Gu, and Shen (2019) exploit such heterogeneity to improve the asymptotic mean squared error of IV estimators with independent and conditionally homoskedastic observations. Instead, we focus on developing bootstrap inference methods that are robust to the first-stage heterogeneity for data with a small number of clusters, while allowing for (weak) within-cluster dependence and heteroskedasticity.

The remainder of this paper is organized as follows. Section **??** presents the main results for

[2] The WREC bootstrap has superior finite sample performance for IV regressions with nonhomoskedastic errors and is very popular among empirical researchers. In our paper, we extend this bootstrap procedure and give conditions under which it is also valid with few clusters, so that empirical researchers can use the WRE-type procedures in a wide range of scenarios.

[3] See, for example, Davidson and MacKinnon (2008, 2010), Moreira, Porter, and Suarez (2009), Wang and Kaffo (2016), Finlay and Magnusson (2019), and Young (2021), among others.

wild bootstrap IV regression. Section 7 provides simulation results. The empirical application is presented in Section 8. We conclude and provide practical recommendations in Section 9.

2 Setup, Estimation, and Inference Procedure

2.1 Setup

Throughout the paper, we consider the setup of a linear IV regression with clustered data,

$$y_{i,j} = X_{i,j}^\top \beta + W_{i,j}^\top \gamma + \varepsilon_{i,j}, \tag{1}$$

where the clusters are indexed by $j \in J = \{1, ..., q\}$ and units in the j-th cluster are indexed by $i \in I_{n,j} = \{1, ..., n_j\}$. In (1), we denote $y_{i,j} \in \mathbf{R}$, $X_{i,j} \in \mathbf{R}^{d_x}$, $W_{i,j} \in \mathbf{R}^{d_w}$, and $Z_{i,j} \in \mathbf{R}^{d_z}$ as an outcome of interest, endogenous regressors, exogenous regressors, and IVs, respectively. $\beta \in \mathbf{R}^{d_x}$ and $\gamma \in \mathbf{R}^{d_w}$ are unknown structural parameters.

We let the parameter of interest β to shift with respect to (w.r.t.) the sample size to incorporate the analyses of size and local power in a concise manner: $\beta_n = \beta_0 + \mu_\beta/\sqrt{n}$, where $\mu_\beta \in \mathbf{R}^{d_x}$ is the local parameter. We let $\lambda_\beta^\top \beta_0 = \lambda_0$, where $\lambda_\beta \in \mathbf{R}^{d_x \times d_r}$, $\lambda_0 \in \mathbf{R}^{d_r}$ and d_r denotes the number of restrictions under the null hypothesis. Define $\mu = \lambda_\beta^\top \mu_\beta$. Then, the null and local alternative hypotheses studied in this paper can be written as

$$\mathcal{H}_0 : \mu = 0 \quad v.s. \quad \mathcal{H}_{1,n} : \mu \neq 0. \tag{2}$$

2.2 K-Class IV Estimators

Throughout the paper, we consider estimators of the form:

$$\left(\hat{\beta}^\top, \hat{\gamma}^\top\right)^\top = \left(\vec{X}^\top \vec{X} - \hat{\kappa} \vec{X}^\top M_{\vec{Z}} \vec{X}\right)^{-1} \left(\vec{X}^\top Y - \hat{\kappa} \vec{X}^\top M_{\vec{Z}} Y\right), \tag{3}$$

where $\vec{Z} = [Z : W]$, $\vec{X} = [X : W]$, Y, X, Z, and W are $n \times 1$, $n \times d_x$, $n \times d_z$, and $n \times d_w$-dimensional vectors and matrices formed by $y_{i,j}$, $X_{i,j}^\top$, $Z_{i,j}^\top$, and $W_{i,j}^\top$, respectively, and $P_A = A(A^\top A)^{-1} A^\top$, $M_A = I_n - P_A$, where A is an n-dimensional matrix and I_n is an n-dimensional identity matrix. This class includes all of the familiar k-class IV estimators. Specifically, we focus on four cases: (1) the two-stage least squares (TSLS) estimator, where $\hat{\kappa} = \hat{\kappa}_{tsls} = 1$, (2) the limited information maximum likelihood (LIML) estimator, where

$$\hat{\kappa} = \hat{\kappa}_{liml} = \min_r r^\top \vec{Y}^\top M_W \vec{Y} r / (r^\top \vec{Y}^\top M_{\vec{Z}} \vec{Y} r), \quad \vec{Y} = [Y : X], \quad \text{and} \quad r = (1, -\beta^\top)^\top,$$

(3) the modified LIML estimator proposed by Fuller (1977, hereafter FULL estimator), where $\hat{\kappa} = \hat{\kappa}_{full} = \hat{\kappa}_{liml} - C/(n - d_z - d_w)$ with some constant C, and (4) the bias-adjusted TSLS (BA) estimator proposed by Nagar (1959) and Rothenberg (1984), where $\hat{\kappa} = \hat{\kappa}_{ba} = n/(n - d_z + 2)$.

Theoretically, we show that all k-class IV estimators are asymptotically the same. However, in our simulation, we find that Fuller's modified LIML estimator has the best finite sample performance.

2.3 Wild Bootstrap Inference

2.3.1 Inference Procedure by Wald Statistics

For inference, we construct Wald statistics based on the the k-class estimator $\hat{\beta}_L$ defined in (3) with $\hat{k} = \hat{k}_L$ for $L \in \{\text{tsls}, \text{liml}, \text{full}, \text{ba}\}$. When the $d_r \times d_r$ weighting matrix $\hat{A}_r$ is asymptotically deterministic in the sense of Assumption 4 below (such as $\hat{A}_r = I_{d_r}$, the $d_r \times d_r$ identity matrix), we denote T_n as the Wald statistic without CCE and define it as

$$T_n = ||\sqrt{n}(\lambda_\beta^\top \hat{\beta}_L - \lambda_0)||_{\hat{A}_r}, \tag{4}$$

where $||u||_A = \sqrt{u^\top A u}$ for a generic vector u and a weighting matrix A. When we use $\hat{A}_{r,CR}$, the inverse of the CCE as defined in (14) in Section 5 as the weighting matrix, we denote $T_{CR,n}$ as the Wald statistic with CCE and define it as

$$T_n = ||\sqrt{n}(\lambda_\beta^\top \hat{\beta}_L - \lambda_0)||_{\hat{A}_{r,CR}}. \tag{5}$$

We reject the null hypothesis if T_n and $T_{CR,n}$ are greater than their corresponding critical values $\hat{c}_n$ and $\hat{c}_{CR,n}$, respectively. We compute the critical values by a wild bootstrap procedure described below.

Step 1: For $L \in \{\text{tsls}, \text{liml}, \text{full}, \text{ba}\}$, compute the null-restricted residual

$$\hat{\varepsilon}_{i,j}^r = y_{i,j} - X_{i,j}^\top \hat{\beta}_L^r - W_{i,j}^\top \hat{\gamma}_L^r,$$

where $\hat{\beta}_L^r$ and $\hat{\gamma}_L^r$ are null-restricted k-class IV estimators of β and γ from $(y_{i,j}, X_{i,j}^\top, W_{i,j}^\top, Z_{i,j}^\top)^\top$,[4] and the unrestricted residual

$$\hat{\varepsilon}_{i,j} = y_{i,j} - X_{i,j}^\top \hat{\beta}_L - W_{i,j}^\top \hat{\gamma}_L, \tag{6}$$

where $\hat{\beta}_L$ and $\hat{\gamma}_L$ are defined in (3) with $\hat{\kappa}_L$.

Step 2: Construct $\overline{Z}_{i,j}$ as

$$\overline{Z}_{i,j} = \left(\widetilde{Z}_{i,j}^\top 1\{j=1\}, ..., \widetilde{Z}_{i,j}^\top 1\{j=q\}\right)^\top,$$

where $\widetilde{Z}_{i,j}$ is the residual of regressing $Z_{i,j}$ on $W_{i,j}$ using the entire sample.

Step 3: Compute the first-stage residual

$$\tilde{v}_{i,j} = X_{i,j} - \widetilde{\Pi}_{\overline{Z}}^\top \overline{Z}_{i,j} - \widetilde{\Pi}_w^\top W_{i,j}, \tag{7}$$

where $\widetilde{\Pi}_{\overline{Z}}$ and $\widetilde{\Pi}_w$ are the OLS coefficients of $\overline{Z}_{i,j}$ and $W_{i,j}$ from regressing $X_{i,j}$ on $(\overline{Z}_{i,j}^\top, W_{i,j}^\top, \hat{\varepsilon}_{i,j})^\top$ using the entire sample.

Step 4: Let $\mathbf{G} = \{-1, 1\}^q$ and for any $g = (g_1, ..., g_q) \in \mathbf{G}$ generate

$$X_{i,j}^*(g) = \widetilde{\Pi}_{\overline{Z}}^\top \overline{Z}_{i,j} + \widetilde{\Pi}_w^\top W_{i,j} + g_j \tilde{v}_{i,j}, \quad y_{i,j}^*(g) = X_{i,j}^{*\top}(g)\hat{\beta}_L^r + W_{i,j}^\top \hat{\gamma}_L^r + g_j \hat{\varepsilon}_{i,j}^r.$$

For each $g = (g_1, ..., g_q) \in \mathbf{G}$, compute $\hat{\beta}_{L,g}^*$ and $\hat{\gamma}_{L,g}^*$, the analogues of the estimators $\hat{\beta}_L$ and $\hat{\gamma}_L$ using $\left(y_{i,j}^*(g), X_{i,j}^{*\top}(g)\right)^\top$ in place of $\left(y_{i,j}, X_{i,j}^\top\right)^\top$ and the same $(Z_{i,j}^\top, W_{i,j}^\top)^\top$. Compute the bootstrap analogue of the Wald statistic:[5]

$$T_n^*(g) = ||\sqrt{n}(\lambda_\beta^\top \hat{\beta}_{L,g}^* - \lambda_0)||_{\hat{A}_r}, \quad T_{CR,n}^*(g) = ||\sqrt{n}(\lambda_\beta^\top \hat{\beta}_{L,g}^* - \lambda_0)||_{\hat{A}_{r,CR,g}^*} \tag{8}$$

where $\hat{A}_{r,CR,g}^*$ is defined in (15).

[4] The null-restricted k-class estimator is defined as

$$\hat{\beta}_L^r = \hat{\beta}_L - \left(X^\top P_{\tilde{Z}} X - \hat{\mu}_L X^\top M_{\tilde{Z}} X\right)^{-1} \lambda_\beta \left(\lambda_\beta^\top (X^\top P_{\tilde{Z}} X - \hat{\mu}_L X^\top M_{\tilde{Z}} X)^{-1} \lambda_\beta\right)^{-1} (\lambda_\beta^\top \hat{\beta}_L - \lambda_0),$$
$$\hat{\gamma}_L^r = (W^\top W)^{-1} W^\top (Y - X\hat{\beta}_L^r), \text{ where } \hat{\mu}_L = \hat{\kappa}_L - 1 \text{ and } \tilde{Z} = M_W Z,$$

for $L \in \{\text{tsls}, \text{liml}, \text{full}, \text{ba}\}$; e.g., see Appendix B of Roodman et al. (2019) for a general formula.

[5] Let $X^*(g)$ and $Y^*(g)$ be the $n \times d_x$ matrix constructed using $X_{i,j}^*(g)$ and the $n \times 1$ vector constructed using $Y_{i,j}^*(g)$, respectively. For $L \in \{\text{tsls}, \text{liml}, \text{full}, \text{ba}\}$ and $g \in \mathbf{G}$,

$$\hat{\beta}_{L,g}^* = \left(X^{*\top}(g) P_{\tilde{Z}} X^*(g) - \hat{\mu}_{L,g}^* X^{*\top}(g) M_{\tilde{Z}} X^*(g)\right)^{-1} \left(X^{*\top}(g) P_{\tilde{Z}} Y^*(g) - \hat{\mu}_{L,g}^* X^{*\top}(g) M_{\tilde{Z}} Y^*(g)\right), \text{ where } \hat{\mu}_{L,g}^* = \hat{\kappa}_{L,g}^* - 1.$$

Step 5: To obtain the critical values, we compute the $1-\alpha$ quantiles of $\{T_n^*(g) : g \in \mathbf{G}\}$ and $\{T_{CR,n}^*(g) : g \in \mathbf{G}\}$:

$$\hat{c}_n(1-\alpha) = \inf\left\{x \in \mathbf{R} : \frac{1}{|\mathbf{G}|}\sum_{g \in \mathbf{G}} 1\{T_n^*(g) \leq x\} \geq 1-\alpha\right\},$$

$$\hat{c}_{CR,n}(1-\alpha) = \inf\left\{x \in \mathbf{R} : \frac{1}{|\mathbf{G}|}\sum_{g \in \mathbf{G}} 1\{T_{CR,n}^*(g) \leq x\} \geq 1-\alpha\right\},$$

where $1\{E\}$ equals one whenever the event E is true and equals zero otherwise. The bootstrap test for $\mathcal{H}_0$ rejects whenever $T_{CR,n}$ exceeds $\hat{c}_{CR,n}(1-\alpha)$ and T_n exceeds $\hat{c}_n(1-\alpha)$ for Wald statistics with and without CCE, respectively.

Four remarks are in order. First, Step 1 imposes null when computing the residuals in the structural equation (1), which is advocated by Cameron et al. (2008), Davidson and MacKinnon (2010), MacKinnon et al. (2022), and Canay et al. (2021), among others. Second, the estimators $\widetilde{\Pi}_{\overline{Z}}$ and $\widetilde{\Pi}_w$ in Step 3 are similar to the efficient reduced-form estimators in the WREC bootstrap procedures advocated by Finlay and Magnusson (2014, 2019), Davidson and MacKinnon (2010), Roodman et al. (2019), and MacKinnon (2021), which have superior finite sample performance for IV regressions, even when the instruments are rather weak. In this paper, we focus on extending the WREC procedure because (1) we find the resulting bootstrap also has excellent finite sample performance for IV regressions with a small number of clusters, and (2) we want to be consistent with Davidson and MacKinnon's (2010) suggestions. Third, to adapt to the current framework, we modify the original WREC procedure and use the fully interacted $\widetilde{Z}_{i,j}$ in Step 3, which is crucial to guarantee that the bootstrap Jacobian matrix $\widehat{Q}^*_{\tilde{Z}X}(g) := \frac{1}{n}\sum_{j \in J}\sum_{i \in I_{n,j}} \tilde{Z}_{i,j} X_{i,j}^{*\top}(g)$ is asymptotically equivalent to the original Jacobian $\widehat{Q}_{\tilde{Z}X}$. Notice that the fully interacted IVs (i.e., $\overline{Z}_{i,j}$) are only needed to construct $X^*_{i,j}(g)$, and we still use the uninteracted IVs (i.e., $Z_{i,j}$) when computing $(\hat{\beta}_L^\top, \hat{\gamma}_L^\top)^\top$ in (3) and their null-restricted and bootstrap counterparts (i.e., $(\hat{\beta}_L^{r\top}, \hat{\gamma}_L^{r\top})^\top$ in Step 1 and $(\hat{\beta}_{L,g}^{*\top}, \hat{\gamma}_{L,g}^{*\top})^\top$ in Step 4). Last, when regressing $X_{i,j}$ on $(\overline{Z}_{i,j}^\top, W_{i,j}^\top, \hat{\varepsilon}_{i,j})^\top$ in Step 3, we need to use the unrestricted residuals $\hat{\varepsilon}_{i,j}$ instead of the null-restricted residuals $\hat{\varepsilon}^r_{i,j}$. This modification is required to establish the power results under few clusters.

2.3.2 Inference Procedure by Weak-instrument-robust Statistics

In this section, we describe a wild bootstrap inference procedure for Anderson-Rubin (AR) type Weak-instrument-robust Statistics with or without CCE. Recall that $\beta_n = \beta_0 + \mu_\beta/\sqrt{n}$. Under the null, we have $\mu_\beta = 0$, or equivalently, $\beta_n = \beta_0$. First, define the AR statistic without CCE as

$$AR_n = \|\sqrt{n}\widehat{f}\|_{\hat{A}_z}, \quad \widehat{f} = n^{-1}\sum_{j\in J}\sum_{i\in I_{n,j}} f_{i,j},$$

where $\hat{A}_z$ is a $d_z \times d_z$ weighting matrix with an asymptotically deterministic limit specified in Assumption 5 below, $f_{i,j} = \widetilde{Z}_{i,j}\bar{\varepsilon}^r_{i,j}$, $\bar{\varepsilon}^r_{i,j} = y_{i,j} - X_{i,j}^\top\beta_0 - W_{i,j}^\top\bar{\gamma}^r$, and $\bar{\gamma}^r$ is the null-restricted ordinary least squares (OLS) estimator of γ:

$$\bar{\gamma}^r = \left(\sum_{j\in J}\sum_{i\in I_{n,j}} W_{i,j}W_{i,j}^\top\right)^{-1}\sum_{j\in J}\sum_{i\in I_{n,j}} W_{i,j}(y_{i,j} - X_{i,j}^\top\beta_0).$$

Second, we also define the AR statistic with the (null-imposed) CCE as

$$AR_{CR,n} = \|\sqrt{n}\widehat{f}\|_{\hat{A}_{CR}}, \quad \hat{A}_{CR} = \left(n^{-1}\sum_{j\in J}\sum_{i\in I_{n,j}}\sum_{k\in I_{n,j}} f_{i,j}f_{k,j}^\top\right)^{-1}.$$

Our wild bootstrap procedure for the AR statistics is defined as follows.

Step 1: Compute the null-restricted residual $\bar{\varepsilon}^r_{i,j} = y_{i,j} - X_{i,j}^\top\beta_0 - W_{i,j}^\top\bar{\gamma}^r$.

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

$$\widehat{f}^*_g = n^{-1}\sum_{j\in J}\sum_{i\in I_{n,j}} f^*_{i,j}(g_j), \quad \text{and} \quad f^*_{i,j}(g_j) = \widetilde{Z}_{i,j}\varepsilon^*_{i,j}(g_j),$$

where $\varepsilon^*_{i,j}(g_j) = g_j\bar{\varepsilon}^r_{i,j}$. Compute the bootstrap statistics:

$$AR^*_n(g) = \|\sqrt{n}\widehat{f}^*_g\|_{\hat{A}_z} \quad \text{and} \quad AR^*_{CR,n}(g) = \|\sqrt{n}\widehat{f}^*_g\|_{\hat{A}_{CR}}.$$

Step 3: Let $\hat{c}_{AR,n}(1-\alpha)$ and $\hat{c}_{AR,CR,n}(1-\alpha)$ denote the $(1-\alpha)$-th quantile of $\{AR^*_n(g)\}_{g\in\mathbf{G}}$ and $\{AR^*_{CR,n}(g)\}_{g\in\mathbf{G}}$, respectively.

Unlike the $T_{CR,n}$-based Wald test in Section 2.3.1, we do not need to bootstrap the CCE for the $AR_{CR,n}$ test even though $\hat{A}_{CR}$ also admits a random limit. This is because $\hat{A}_{CR}$ is invariant to the sign changes.

3 Main Assumptions and Several Examples

In this section, we introduce the assumptions that will be used in our analysis of the wild bootstrap tests under a small number of clusters in Sections 4-6. For the rest of the paper, we define $\widetilde{Z}_{i,j}$ as the residual of regressing $Z_{i,j}$ on $W_{i,j}$ using the entire sample, and that for any random vectors $U_{i,j}$ and $V_{i,j}$, $\widehat{Q}_{UV,j} = \frac{1}{n_j}\sum_{i\in I_{n,j}} U_{i,j}V_{i,j}^\top$ and $\widehat{Q}_{UV} = \frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}} U_{i,j}V_{i,j}^\top$. Further define $\widehat{Q} = \widehat{Q}_{\widetilde{Z}X}^\top \widehat{Q}_{\widetilde{Z}\widetilde{Z}}^{-1}\widehat{Q}_{\widetilde{Z}X}$, and Q as the probability limits of $\widehat{Q}$.

Assumption 1. *The following statements hold: (i) For each $j \in J$, either (1) $\widehat{Q}_{\widetilde{Z}W,j} = 0$, or (2) $\widehat{Q}_{\widetilde{Z}W,j} = o_p(1)$ and $\frac{1}{\sqrt{n}}\sum_{j\in J}\sum_{i\in I_{n,j}} W_{i,j}\varepsilon_{i,j} = O_p(1)$.*
(ii) There exists a collection of independent random variables $\{\mathcal{Z}_j : j \in J\}$, where $\mathcal{Z}_j \sim N(0,\Sigma_j)$ with Σ_j positive definite for all $j \in J$, such that

$$\left\{\frac{1}{\sqrt{n_j}}\sum_{i\in I_{n,j}} \widetilde{Z}_{i,j}\varepsilon_{i,j} : j \in J\right\} \xrightarrow{d} \{\mathcal{Z}_j : j \in J\}.$$

(iii) For each $j \in J$, $n_j/n \to \xi_j > 0$.
(iv) $\frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}} W_{i,j}W_{i,j}^\top$ is invertible.

Several remarks are in order. First, we have $\widehat{Q}_{\widetilde{Z}W,j} = 0$ if $W_{i,j}$ contains the interactions between baseline exogenous regressors and cluster dummies or $Z_{i,j}$ is constructed as the residual from the cluster-level projection of original IVs onto the linear space spanned by $W_{i,j}$.[6]

Second, when some cluster contains small number of observations, such a cluster-level projection is numerically unstable (this is just a finite sample issue as asymptotically, we assume the number of clusters is fixed and the cluster size diverges to infinity). In this case, researchers may prefer to use baseline exogenous regressors without interacting them with cluster dummies. To accommodate such a practice, we give another set of conditions in Assumption 1(i)(2). Specifically, we require $\widehat{Q}_{\widetilde{Z}W,j} = o_p(1)$ if

$$\frac{1}{n_j}\sum_{i\in I_{n,j}} \left\| W_{i,j}^\top \left(\widehat{\Gamma}_n - \widehat{\Gamma}_{n,j}\right)\right\|^2 \xrightarrow{p} 0, \tag{9}$$

where $\widehat{\Gamma}_n$ and $\widehat{\Gamma}_{n,j}$ are the $d_w \times d_z$ matrices that satisfy the following orthogonality conditions:

[6] Specifically, suppose $\mathscr{Z}_{i,j}$ are the base IVs. We can construct $Z_{i,j}$ as $Z_{i,j} = \mathscr{Z}_{i,j} - \hat{\chi}_j^\top W_{i,j}$, where $\hat{\chi}_j = \widehat{Q}_{WW,j}\widehat{Q}_{WW,j}^-\widehat{Q}_{WW,j}^-\widehat{Q}_{W\mathscr{Z},j}$, and A^- denotes the pseudo inverse of the positive semidefinite matrix A. It is possible to show that $\widetilde{Z}_{i,j} = Z_{i,j}$ and $\frac{1}{n_j}\sum_{i\in I_{n,j}} \widetilde{Z}_{i,j}W_{i,j}^\top = 0$.

$\sum_{j\in J}\sum_{i\in I_{n,j}} W_{i,j}(Z_{i,j} - \widehat{\Gamma}_n^\top W_{i,j})^\top = 0$, and $\sum_{i\in I_{n,j}} W_{i,j}(Z_{i,j} - \widehat{\Gamma}_{n,j}^\top W_{i,j})^\top = 0$. Canay et al. (2021, Assumption 2(iv) in Section A) imposed the same condition as (9) and pointed out that it holds whenever the distributions of $(Z_{i,j}^\top, W_{i,j}^\top)^\top$ are the same across clusters. The condition that

$$\frac{1}{\sqrt{n}}\sum_{j\in J}\sum_{i\in I_{n,j}} W_{i,j}\varepsilon_{i,j} = O_p(1)$$

is similar to Assumption 1(i) in Canay et al. (2021) and rules out the specification in which $\varepsilon_{i,j}$ follows an error-component model, i.e.,

$$\varepsilon_{i,j} = \eta_j + e_{i,j}, \tag{10}$$

where η_j is a cluster-wide shock for cluster j, $e_{i,j}$ is an idiosyncratic shock for observation i, and the two shocks are independent. We emphasize again that if a full set of interactions between baseline exogenous regressors and cluster dummies are used as W_{ij}, then Assumption 1(i)(1) holds automatically and we do not need to Assumption 1(i)(2).

Third, Assumption 1(ii) is reasonable because $\widetilde{Z}_{i,j}$ is exogenous. It is satisfied whenever the within-cluster dependence is sufficiently weak to permit the application of a suitable central limit theorem and the data are independent across clusters. Assumption 1(iii) gives the restriction on cluster sizes, and Assumption 1(iv) ensures $\widehat{\Gamma}_n$ is uniquely defined.

Assumption 2. *The following statements hold: (i) The quantities $\widehat{Q}_{\widetilde{Z}X,j}$, $\widehat{Q}_{\widetilde{Z}\widetilde{Z},j}$, $\widehat{Q}_{\widetilde{Z}X}$, and $\widehat{Q}_{\widetilde{Z}\widetilde{Z}}$ converge in probability to deterministic matrices, which are denoted as $Q_{\widetilde{Z}X,j}$, $Q_{\widetilde{Z}\widetilde{Z},j}$, $Q_{\widetilde{Z}X}$, and $Q_{\widetilde{Z}\widetilde{Z}}$, respectively.*
(ii) The matrices $Q_{\widetilde{Z}\widetilde{Z},j}$ is invertible for $j \in J$.
(iii) For all $j \in J$, $\widehat{Q}_{XX,j} = O_p(1)$, $\widehat{Q}_{X\varepsilon,j} = O_p(1)$, and $\widehat{Q}_{\dot{\varepsilon}\dot{\varepsilon},j} \geq c > 0$ for constant c with probability approaching one, where $\dot{\varepsilon}_{i,j}$ is the residual from the cluster-level projection of $\varepsilon_{i,j}$ on $W_{i,j}$.

Assumption 2(i) holds if the dependence of units within clusters is weak enough to render some type of LLN to hold. Assumption 2(ii) is standard in the literature and holds regardless of the IVs' strength. However, it rules out the case that IVs are constructed as the interaction between baseline IVs and cluster dummies, as discussed below.

We conclude this section with several examples.

Example 3.1 (Cluster-level Exogenous Variable and Fixed Effects)**.** *We allow for cluster-level* *exogenous variables and fixed effects. Suppose*

$$y_{i,j} = X_{i,j}^\top \beta + B_{1,j}^\top \theta_{1,j} + B_{2,i,j}^\top \theta_{2,j} + \eta_j + e_{i,j}, \tag{11}$$

where $y_{i,j}$ is the outcome variable, $X_{i,j}$ contains the endogenous variables, $B_{1,j}$ contains the cluster-level baseline exogenous variables including the intercept, $B_{2,i,j}$ contains the individual-level baseline exogenous variables, η_j is the unobserved cluster-level fixed effect, $e_{i,j}$ is the individual-level idiosyncratic error, and the linear coefficients $(\theta_{1,j}, \theta_{2,j})$ for the baseline covariates $B_i = (B_{1,j}, B_{2,i,j})$ are allowed to be heterogeneous across clusters. Further denote $W_{i,j}$ as the full interaction between $\widetilde{B}_{2,i,j} = (1, B_{2,i,j}^\top)^\top$ with cluster dummies and define it as

$$W_{i,j} = (\widetilde{B}_{2,i,j}^\top 1\{j=1\}, \widetilde{B}_{2,i,j}^\top 1\{j=2\}, \cdots, \widetilde{B}_{2,i,j}^\top 1\{j=q\})^\top. \tag{12}$$

Then, (11) can be rewritten as (1) with $\varepsilon_{i,j} = \eta_j + e_{i,j}$ and $\gamma = (\gamma_1^\top, \cdots, \gamma_q^\top)^\top$, where $\gamma_j = (B_{1,j}^\top \theta_{1,j}, \theta_{2,j}^\top)^\top$. As $W_{i,j}$ contains full interaction between $\widetilde{\beta}_{2,i,j}$ and cluster dummies, $\widetilde{Z}_{i,j}$ is numerically the same as the residual from the cluster-level projection of $Z_{i,j}$ on $\widetilde{\beta}_{2,i,j}$, which implies $\widehat{Q}_{\widetilde{Z}W,j} = 0$ and Assumption 1(1) holds. We further have $\sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} \eta_j = 0$ so that Assumption 1(2) reduces to

$$\left\{ \frac{1}{\sqrt{n_j}} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} e_{i,j} : j \in J \right\} \xrightarrow{d} \{\mathcal{Z}_j : j \in J\},$$

which holds if $Z_{i,j}$ and $B_{2,i,j}$ are exogenous and the the within-cluster dependence is sufficiently weak.

Example 3.2 (Heterogeneous IV Strength across Clusters)**.** *We allow for cluster-level heterogeneity with regard to IV strength. Let $\Pi_{z,j,n}$ and $\Pi_{w,j,n}$ be the coefficients of $Z_{i,j}$ and $W_{i,j}$, respectively, via the cluster-level population projection of $X_{i,j}$ on $Z_{i,j}$ and $W_{i,j}$, for each $j \in J$.*[7] *Then, our model (1) allows for both $\Pi_{z,j,n}$ and $\Pi_{w,j,n}$ to vary across clusters. In particular, we allow for some of $\Pi_{z,j,n}$ to decay to or be zero for the bootstrap Wald tests and all $\Pi_{z,j,n}$ to decay to or be zero for the bootstrap AR tests. We will come back to this point in Sections 4-6 with more details.*

Example 3.3 (Homogeneous Slope for the Endogenous Variable)**.** *Similar to Canay et al. (2021),*

[7] We note $\Pi_{z,j,n}$ and $\Pi_{w,j,n}$ depend on the sample size because the underlying distribution is indexed by n.

we are unable to allow for β to be heterogeneous across clusters. As a stylized example, let

$$y_{i,j} = X_{i,j}^\top \beta_j + W_{i,j} + e_{i,j}, \tag{13}$$

where $W_{i,j}$ is just the cluster dummies. For β equal to weighted average of β_j's, we may rewrite (13) as (1) with $\varepsilon_{i,j} = X_{i,j}^\top(\beta_j - \beta) + e_{i,j}$, which implies

$$\frac{1}{\sqrt{n_j}} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} \varepsilon_{i,j} = \frac{1}{\sqrt{n_j}} \sum_{i \in I_{n,j}} (Z_{i,j} - \bar{Z}_j)(X_{i,j}^\top(\beta_j - \beta) + e_{i,j}).$$

We then see that Assumption 1(2) is violated unless $\beta_j = \beta$.

Example 3.4 (Difference-in-Difference and Cluster Randomization)**.** *The Wald tests are unable to allow for cluster-level IVs, which usually occur for difference-in-difference analysis and cluster randomization with imperfect compliance. In those settings, the treatment status and assignments are interpreted as our $X_{i,j}$ and $Z_{i,j}$, respectively, and they are different due to imperfect compliance. However, if the cluster is assigned as a control group, then all $Z_{i,j}$'s for such a cluster take value 0. When $Z_{i,j}$ is invariant within some cluster j_0, then $\frac{1}{\sqrt{n_{j_0}}} \sum_{i \in I_{n,j_0}} \widetilde{Z}_{i,j_0} \varepsilon_{i,j_0}$ must be degenerate under Assumption 1(i), which violates Assumption 1(ii). Following the suggestion by Canay et al. (2017) and Canay et al. (2021), it is possible to merge treated and control clusters to form a more coarse cluster.*

Example 3.5 (Cluster-level Endogenous Variables)**.** *If $X_{i,j}$ is a cluster-level variable (say, X_j), then the resulting within-cluster limiting Jacobian matrix $Q_{\widetilde{Z}X,j}$ may be random and potentially correlated with the within-cluster score component $\mathcal{Z}_j$ as X_j is endogenous, which violates Assumption 2(i). We notice that similar issues can arise for the approaches by Bester et al. (2011), Hwang (2021), IM, and CRS. Our wild bootstrap AR tests (AR_n and $AR_{CR,n}$) only requires Assumption 1 but not Assumption 2, and then, remain valid.*

Example 3.6 (Interacting IVs with Cluster Dummies)**.** *We require the IVs $Z_{i,j}$ to be the baseline instruments which are not interacted with cluster dummies. In fact, the condition that $\mathcal{Z}_j$ have full rank covariance matrices in Assumption 1(ii) rules out the case in which $Z_{i,j}$ are constructed by interacting the baseline IVs with the cluster dummies. To see this, we consider the simplest case that $W_{i,j}$ only contains the intercept and $Z_{i,j}$ is constructed as the interactions of a scalar baseline*

IV $\mathscr{Z}_{i,j}$ with cluster dummies. Then, for the last cluster (the q-th cluster), we have

$$\frac{1}{\sqrt{n_q}} \sum_{i \in I_{n,q}} \widetilde{Z}_{i,q} \varepsilon_{i,q} = \frac{1}{\sqrt{n_q}} \sum_{i \in I_{n,q}} (-\xi_1 \overline{Z}_1 \varepsilon_{i,q}, \cdots, -\xi_{q-1} \overline{Z}_{q-1} \varepsilon_{i,q}, (\mathscr{Z}_{i,q} - \xi_q \overline{Z}_q)\varepsilon_{i,q})^\top,$$

where $\overline{Z}_j = \frac{1}{n_j} \sum_{i \in I_{n,j}} \mathscr{Z}_{i,j}$. Clearly, $\frac{1}{\sqrt{n_q}} \sum_{i \in I_{n,q}} \widetilde{Z}_{i,q} \varepsilon_{i,q}$ is linearly dependent, which implies Σ_q is degenerated, and thus, Assumption 1(ii) is violated.

We notice that Abadie et al. (2019) interact the baseline IVs with subgroup dummies such as those for states, gender, or race. However, Abadie et al. (2019) use an analytical covariance matrix estimator for independent and conditionally homoskedastic observations for inference. In contrast, as we allow for heteroskedasiticity and are agnostic about the within-cluster dependence, it is difficult, if not impossible, to derive a consistent estimator of the covariance matrix in our setting without imposing additional restrictions. Instead, following the lead of Canay et al. (2021), we rely on the connection between the wild bootstrap and the randomization test to avoid the consistent estimation of the covariance matrix. Assumption 1(ii) is crucial for such a connection to hold.

4 Asymptotic Results for the Wald Tests without CCE

For the Wald test, we further assume the following assumption.

Assumption 3. *(i) $Q_{\widetilde{Z}X}$ is of full column rank.*
(ii) One of the following two conditions holds: (1) $d_x = 1$, and define $a_j = Q^{-1} Q_{\widetilde{Z}X}^\top Q_{\widetilde{Z}\widetilde{Z}}^{-1} Q_{\widetilde{Z}X,j}$, where $Q = Q_{\widetilde{Z}X}^\top Q_{\widetilde{Z}\widetilde{Z}}^{-1} Q_{\widetilde{Z}X}$, or (2) there exists a scalar a_j for each $j \in J$ such that $Q_{\widetilde{Z}X,j} = a_j Q_{\widetilde{Z}X}$.

Several remarks are in order. First, Assumption 3(i) requires (overall) strong identification for β_n. Second, Assumption 3(ii)(1) states that if there is only *one endogenous variable*, no further restrictions are required as we can always define a scalar $a_j = Q^{-1} Q_{\widetilde{Z}X}^\top Q_{\widetilde{Z}\widetilde{Z}}^{-1} Q_{\widetilde{Z}X,j}$ when $d_x = 1$. A single endogenous variable is the leading case in empirical applications involving IV regressions. For example, 101 out of 230 specifications in Andrews et al. (2019)'s sample and 1,087 out of 1,359 in Young (2021)'s sample has one endogenous regressor and one IV. Lee, McCrary, Moreira, and Porter (2021) found that among 123 papers published in AER between 2013 and 2019 that include IV regressions, 61 employ single instrumental variable (just-identified) regressions. They pointed out that the single-IV case "includes applications such as randomized trials with imperfect compliance (estimation of LATE, Imbens and Angrist (1994)), fuzzy regression discontinuity designs (see discussion in Lee and Lemieux (2010)), and fuzzy regression kink designs (see discussion in Card, Lee, Pei, and Weber (2015))". Angrist and Kolesár (2021) also pointed out that "most studies using IV (including Angrist (1990) and Angrist and Krueger (1991)) report just-identified IV estimates computed with a single instrument".[8] In addition, Assumption 3(ii)(1) further allows for the case of single endogenous regressor and multiple IVs. Third, Assumption 3(ii)(2) is needed if we have *multiple endogenous variables*. The condition is similar to that in Canay et al. (2021, Assumption 2(iii)), which restricts the type of heterogeneity of the within-cluster Jacobian matrices. However, it is still weaker than restrictions assumed in the literature for cluster-robust Wald tests under a small number of clusters. For example, Bester et al. (2011) and Hwang (2021) provide asymptotic approximations that are based on t and F distributions for the Wald statistics with CCE. The conditions in their papers require the within-cluster Jacobian matrices to have the same limit for all clusters (i.e., Assumption 3(ii)(2) to hold with $a_j = 1$ for all $j \in J$).[9] They also impose that the cluster sizes are approximately equal for all clusters and the cluster-level scores in Assumption 1(ii) have the same normal limiting distribution for all clusters, which are not necessary for the wild bootstrap. Finally, Assumption 3 will not be needed for the bootstrap AR tests in Section 6, as they require neither strong identification nor homogeneity conditions.

To further clarify our setting, we can relate the Jacobian matrices with the first-stage projection coefficient. Specifically, recall $\Pi_{z,j,n}$ is the coefficient of $Z_{i,j}$ via the cluster-level population projection of $X_{i,j}$ on $Z_{i,j}$ and $W_{i,j}$. Then, we have $\lim_{n\to\infty} \Pi_{z,j,n} = \Pi_{z,j} := Q^{-1}_{\widetilde{Z}\widetilde{Z},j} Q_{\widetilde{Z}X,j}$ under our framework. Also define $\Pi_z = Q^{-1}_{\widetilde{Z}\widetilde{Z}} Q_{\widetilde{Z}X}$. Assumption 3(i) ensures that overall we have strong identification as $Q_{\widetilde{Z}X}$ (and Π_z) is of full column rank. Furthermore, we call the clusters in which $Q_{\widetilde{Z}X,j}$ (and $\Pi_{z,j}$) are of full column rank the strong clusters, i.e., β_n is strongly identified in these clusters. On the other hand, strong identification for β_n is not ensured in the rest of the clusters. Given the number of clusters is fixed, only one strong cluster is needed for Assumption 3(i) to hold. Two additional remarks are in order for the case with multiple endogenous variables: (1) Assumption 3(ii)(2) implies that when $a_j \neq 0$, $Q_{\widetilde{Z}X,j}$ (and $\Pi_{z,j}$) is of full column rank, so that the j-th cluster is a strong cluster, and (2) Assumption 3(ii)(2) excludes the case that $Q_{\widetilde{Z}X,j}$ is of a reduced rank but is not a zero matrix. It is possible to select out the clusters with Jacobian matrices of reduced rank (Robin and Smith, 2000; Kleibergen and Paap, 2006; Chen and Fang, 2019). We leave this

[8] In our empirical application, we revisit the influential study by Autor et al. (2013), which also has only one IV.
[9] E.g., see Bester et al. (2011, Assumptions 3 and 4) and Hwang (2021, Assumptions 4 and 5) for details.

investigation for future research.

Assumption 4. *Suppose* $\hat{A}^*_{r,g}$ *equals* $\hat{A}_r$ *and* $||\hat{A}_r - A_r||_{op} = o_p(1)$*, where* A_r *is a* $d_r \times d_r$ *symmetric deterministic weighting matrix such that* $0 < c \leq \lambda_{min}(A_r) \leq \lambda_{max}(A_r) \leq C < \infty$ *for some constants* c *and* C*,* $\lambda_{\min}(A)$ *and* $\lambda_{\max}(A)$ *are the minimum and maximum eigenvalues of the generic matrix* A*, and* $||A||_{op}$ *denotes the operator norm of the matrix* A*.*

Assumption 4 requires the weighting matrix $\hat{A}_r$ in (4) has a deterministic limit and the bootstrap weighting matrix $\hat{A}^*_{r,g}$ in (8) equals $\hat{A}_r$. It rules out the case that $\hat{A}_r$ equals the inverse of CCE, which has a random limit under a small number of clusters. We will discuss the bootstrap Wald test with CCE in Section 5.

Theorem 4.1. *Suppose that Assumptions 1-4 hold. Then under* $\mathcal{H}_0$*, for all four estimation methods (namely, TSLS, LIML, FULL, and BA),*

$$\alpha - \frac{1}{2^{q-1}} \leq \liminf_{n\to\infty} \mathbb{P}\{T_n > \hat{c}_n(1-\alpha)\} \leq \limsup_{n\to\infty} \mathbb{P}\{T_n > \hat{c}_n(1-\alpha)\} \leq \alpha + \frac{1}{2^{q-1}}.$$

Several remarks are in order. First, Theorem 4.1 states that as long as there exists at least one strong cluster, the T_n-based wild bootstrap test has limiting null rejection probability no greater than $\alpha + 1/2^{q-1}$ and no smaller than $\alpha - 1/2^{q-1}$. The error $1/2^{q-1}$ can be viewed as the upper bound for the asymptotic size distortion, which decreases exponentially with the total number of clusters rather than the number of strong clusters. Intuitively, although the weak clusters do not contribute to the identification of β_n, the scores of such clusters still contribute to the limiting distributions of the IV estimators, which in turn determines the total number of sign changes in the bootstrap Wald statistics. We note that $1/2^{q-1}$ equals 1.56% and 0.2% when $q = 7$ and 10, respectively. If such an distortion is still of concern, researchers can replace α in our context by $\alpha - 1/2^{q-1}$ to ensure null rejection rate.

Second, the wild bootstrap test has resemblance to the group-based t-test in Ibragimov and Müller (2010, hereafter IM) and the randomization test with sign changes in Canay et al. (2017, hereafter CRS). However, we notice that for IV regressions, the size properties of the two approaches can be rather different from that of the wild bootstrap. More specifically, IM and CRS approaches separately estimate the parameters using the samples in each cluster (say, $\hat{\beta}_1, ..., \hat{\beta}_q$), and therefore requires β_n to be strongly identified in all clusters. This would rule out weak clusters in the sense of

Staiger and Stock (1997), where $\Pi_{z,j,n}$ has the same order of magnitude as $n_j^{-1/2}$.[10] In contrast, the size result in Theorem 4.1 holds even with only one strong cluster, and the wild bootstrap is thus more robust to cluster heterogeneity in IV strength. On the other hand, if β_n is strongly identified in all clusters and the cluster-level IV estimators have minimal finite sample bias, IM and CRS have an advantage over the wild bootstrap when there are multiple endogenous variables as they do not require Assumption 3(ii). The two types of approaches could therefore be considered as complements, and practitioners may choose between them according to the characteristics of their data and models.

Third, it is well known that estimators such as LIML and FULL have reduced finite sample bias relative to TSLS in the over-identified case, especially when the IVs are not strong. Since the validity of the randomization with sign changes requires a distributional symmetry around zero, the LIML and FULL-based bootstrap Wald tests may therefore achieve better finite sample size control than that based on TSLS. This is confirmed by the simulation experiments in Section 7.[11]

We next examine the power of the wild bootstrap test against local alternatives.

Theorem 4.2. *Suppose that Assumptions 1-4 hold. Further suppose that there exists a subset J_s of J such that $a_j > 0$ for each $j \in J_s$, $a_j = 0$ for $j \in J \backslash J_s$, and $\lceil |\boldsymbol{G}|(1-\alpha) \rceil \leq |\boldsymbol{G}| - 2^{q-q_s+1}$, where $|\boldsymbol{G}| = 2^q$, $q_s = |J_s|$, and a_j is defined in Assumption 3. Then under $\mathcal{H}_{1,n}$, for all four estimation methods (namely, TSLS, LIML, FULL, and BA),*

$$\lim_{||\mu||_2 \to \infty} \liminf_{n \to \infty} \mathbb{P}\{T_n > \hat{c}_n(1-\alpha)\} = 1.$$

Two remarks are in order. First, to establish the power of the T_n-based wild bootstrap test against $n^{-1/2}$-local alternatives, we need homogeneity of the signs of Jacobians for the strong clusters (i.e., $a_j > 0$ for each $j \in J_s$). For example, in the case with a single IV, it requires $\Pi_{z,j}$ to have the same sign across all the strong clusters. We notice that this condition is not needed for the bootstrap Wald test described in Section 5. Second, we also need a sufficient number of strong clusters. For instance, if q equals 10, then $|\mathbf{G}| = 1024$ and the condition $\lceil |\mathbf{G}|(1-\alpha) \rceil \leq |\mathbf{G}| - 2^{q-q_s+1}$ requires

[10] The cluster-level IV estimators of such weak clusters would become inconsistent and have highly nonstandard limiting distributions. Also, if there exist both strong and "semi-strong" clusters, in which the (unknown) convergence rates of IV estimators can vary among clusters and be slower than $\sqrt{n_j}$ (Andrews and Cheng, 2012), then the estimators with the slowest convergence rate will dominate in the test statistics that are based on the cluster-level estimators.

[11] To theoretically document the asymptotic bias due to the dimensionality of IVs, one needs to consider an alternative framework in which the number of clusters is fixed but the number of IVs tends to infinity, following the literature on many/many weak instruments (Bekker, 1994; Chao and Swanson, 2005; Mikusheva and Sun, 2022). We leave this direction of investigation for future research.

that $q_s \geq 5$ and $q_s \geq 6$ for $\alpha = 10\%$ and 5%, respectively. Theorem 4.2 suggests that although the size of the wild bootstrap test is well controlled even with only one strong cluster, its power depends on the number of strong clusters.

5 Asymptotic Results for the Wald Tests with CCE

Now we consider a wild bootstrap test for the Wald statistic with CCE when the weighting matrix $\hat{A}_{r,CR}$, the inverse of the CCE, is defined as

$$\hat{A}_{r,CR} = \left(\lambda_\beta^\top \widehat{V} \lambda_\beta\right)^{-1}, \quad \widehat{V} = \widehat{Q}^{-1}\widehat{Q}_{\widetilde{Z}X}^\top \widehat{Q}_{\widetilde{Z}\widetilde{Z}}^{-1}\widehat{\Omega}_{CR}\widehat{Q}_{\widetilde{Z}\widetilde{Z}}^{-1}\widehat{Q}_{\widetilde{Z}X}\widehat{Q}^{-1}, \quad \widehat{Q} = \widehat{Q}_{\widetilde{Z}X}^\top \widehat{Q}_{\widetilde{Z}\widetilde{Z}}^{-1}\widehat{Q}_{\widetilde{Z}X}, \tag{14}$$

$\widehat{\Omega}_{CR} = n^{-1}\sum_{j\in J}\sum_{i\in I_{n,j}}\sum_{k\in I_{n,j}} \widetilde{Z}_{i,j}\widetilde{Z}_{k,j}^\top \hat{\varepsilon}_{i,j}\hat{\varepsilon}_{k,j}$, and $\hat{\varepsilon}_{i,j}$ is the unrestricted residual defined in (6). The corresponding Wald statistic with CCe is denoted as

$$T_{CR,n} = ||\sqrt{n}(\lambda_\beta^\top \hat{\beta} - \lambda_0)||_{\hat{A}_{r,CR}}.$$

To obtain wild bootstrap critical value for the Wald statistic with CCE, the bootstrap weighting matrix $\hat{A}^*_{r,CR,g}$ in (8) is defined as

$$\hat{A}^*_{r,CR,g} \;=\; \left(\lambda_\beta^\top \widehat{V}_g^* \lambda_\beta\right)^{-1}, \quad \widehat{V}_g^* = \widehat{Q}_g^{*-1}\widehat{Q}_{\widetilde{Z}X}^{*\top}(g)\widehat{Q}_{\widetilde{Z}\widetilde{Z}}^{-1}\widehat{\Omega}^*_{CR,g}\widehat{Q}_{\widetilde{Z}\widetilde{Z}}^{-1}\widehat{Q}^*_{\widetilde{Z}X}(g)\widehat{Q}_g^{*-1}, \tag{15}$$

where for $L \in \{\text{tsls}, \text{liml}, \text{full}, \text{ba}\}$,

$$\widehat{\Omega}^*_{CR,g} = \frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}}\sum_{k\in I_{n,j}} \widetilde{Z}_{i,j}\widetilde{Z}_{k,j}^\top \hat{\varepsilon}^*_{i,j}(g)\hat{\varepsilon}^*_{k,j}(g), \quad \widehat{Q}^*_{\widetilde{Z}X}(g) = \frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}} \widetilde{Z}_{i,j}X_{i,j}^{*\top}(g),$$
$$\hat{\varepsilon}^*_{i,j}(g) = y^*_{i,j}(g) - X_{i,j}^{*\top}(g)\hat{\beta}^*_{g,L} - W_{i,j}^\top \hat{\gamma}^*_{g,L}, \quad \text{and} \quad \widehat{Q}^*_g = \widehat{Q}_{\widetilde{Z}X}^{*\top}(g)\widehat{Q}_{\widetilde{Z}\widetilde{Z}}^{-1}\widehat{Q}^*_{\widetilde{Z}X}(g). \tag{16}$$

Then, we denote the corresponding bootstrap Wald statistic with CCE as

$$T^*_{CR,n}(g) = ||\sqrt{n}(\lambda_\beta^\top \hat{\beta}^*_g - \lambda_0)||_{\hat{A}^*_{r,CR,g}},$$

and the bootstrap critical value $\hat{c}_{CR,n}(1-\alpha)$ as the $1-\alpha$ quantile of $\{T^*_{CR,n}(g) : g \in \mathbf{G}\}$. Unlike the Wald statistic without CCE considered in the previous section, here we need to bootstrap the weighting matrix because under our asymptotic framework with a small number of clusters $\hat{A}_{r,CR}$ has a random limit, which depends on the limits of the scores and the IV estimators $(\hat{\beta}_L, \hat{\gamma}_L)$.

Theorem 5.1. *Suppose that Assumptions 1-3 hold, and $q > d_r$. Then under $\mathcal{H}_0$, for all four estimation methods (namely, TSLS, LIML, FULL, and BA),*

$$\alpha - \frac{1}{2^{q-1}} \leq \liminf_{n\to\infty} \mathbb{P}\{T_{CR,n} > \hat{c}_{CR,n}(1-\alpha)\} \\ \leq \limsup_{n\to\infty} \mathbb{P}\{T_{CR,n} > \hat{c}_{CR,n}(1-\alpha)\} \leq \alpha + \frac{1}{2^{q-1}}.$$

We require $q > d_r$ because otherwise CCE and its bootstrap counterpart are not invertible. Theorem 5.1 states that with at least one strong cluster, the $T_{CR,n}$-based wild bootstrap test controls size asymptotically up to a small error. Next, we turn to the local power.

Theorem 5.2. *(i) Suppose that Assumptions 1-3 hold, and $q > d_r$. Suppose that there exists a subset J_s of J such that $\min_{j\in J_s} |a_j| > 0$, $a_j = 0$ for each $j \in J\backslash J_s$, and $\lceil |\boldsymbol{G}|(1-\alpha)\rceil \leq |\boldsymbol{G}| - 2^{q-q_s+1}$, where $|\boldsymbol{G}| = 2^q$, $q_s = |J_s|$, and a_j is defined in Assumption 3. Then under $\mathcal{H}_{1,n}$, for all four estimation methods (namely, TSLS, LIML, FULL, and BA),*

$$\lim_{||\mu||_2\to\infty} \liminf_{n\to\infty} \mathbb{P}\{T_{CR,n} > \hat{c}_{CR,n}(1-\alpha)\} = 1.$$

(ii) Further suppose that $d_r = 1$. Then under $\mathcal{H}_{1,n}$, for any $e > 0$, there exists a constant $c_\mu > 0$ such that when $||\mu||_2 > c_\mu$,

$$\liminf_{n\to\infty} \mathbb{P}(\phi_n^{cr} \geq \phi_n) \geq 1 - e,$$

where $\phi_n^{cr} = 1\{T_{CR,n} > \hat{c}_{CR,n}(1-\alpha)\}$ and $\phi_n = 1\{T_n > \hat{c}_n(1-\alpha)\}$.

Several remarks are in order. First, different from Theorem 4.2, the power result in Theorem 5.2 does not require the homogeneity condition on the sign of first-stage coefficients for the strong clusters (i.e., it only requires $\min_{j\in J_s} |a_j| > 0$). Such difference originates from certain good property of $\hat{c}_{CR,n}(1-\alpha)$. More precisely, although both T_n and $T_{CR,n}$ diverge as $||\mu||_2 \to \infty$, their corresponding bootstrap critical values have different behaviors, with $\hat{c}_n(1-\alpha) \xrightarrow{p} \infty$ while $\hat{c}_{CR,n}(1-\alpha) = O_p(1)$, which translates into relatively good power properties of the bootstrap Wald test with CCE. The behaviours of the test statistics and their bootstrap critical values under the distant alternative are summarized in Table 1.

Second, we further establish in Theorem 5.2(ii) that in the case of a t-test (i.e., when the null hypothesis involves one restriction), the rejection of the $T_{CR,n}$-based bootstrap test dominates that based on T_n with large probability under the distant alternative. Intuitively, we have imposed the null when generating the bootstrap pseudo-data, making $\hat{\varepsilon}^*_{i,j}(g)$ and, subsequently, $\lambda_\beta^\top \widehat{V}_g^* \lambda_\beta$ dependent on μ. Therefore, although $|\sqrt{n}(\lambda_\beta^\top \hat{\beta}_g^* - \beta_0)|$ diverges to infinity when $\|\mu\|_2 \to \infty$, $\lambda_\beta^\top \widehat{V}_g^* \lambda_\beta$ also diverges so that $\hat{c}_{CR,n}(1-\alpha)$ remains bounded in probability. Now, let $\tilde{c}_{CR,n}(1-\alpha)$ denote the $(1-\alpha)$ quantile of $\left\{ |\sqrt{n}(\lambda_\beta^\top \hat{\beta}_g^* - \lambda_0)| / \sqrt{\lambda_\beta^\top \widehat{V} \lambda_\beta} : g \in \mathbf{G} \right\}$, in which $|\sqrt{n}(\lambda_\beta^\top \hat{\beta}_g^* - \lambda_0)|$ is studentized by the original CCE instead of the bootstrap CCE, and notice that $1\{T_n > \hat{c}_n(1-\alpha)\} = 1\{T_{CR,n} > \tilde{c}_{CR,n}(1-\alpha)\}$. The result of Theorem 5.2(ii) follows since $\sqrt{\lambda_\beta^\top \widehat{V} \lambda_\beta}$ does not diverge with $||\mu||_2$ so that $\tilde{c}_{CR,n}(1-\alpha) > \hat{c}_{CR,n}(1-\alpha)$ with large probability as $||\mu||_2$ becomes sufficiently large.

Third, these power properties carry over to the linear regression model and the (single-equation) wild bootstrap procedure studied by Canay et al. (2021), as linear regression with exogenous regressors is a special case of the IV regression.

	Test Statistics	Bootstrap Critical Values
Wald Tests	$T_n \xrightarrow{p} \infty$	$\hat{c}_n(1-\alpha) \xrightarrow{p} \infty$
	$T_{CR,n} \xrightarrow{p} \infty$	$\hat{c}_{CR,n}(1-\alpha) = O_p(1)$
AR Tests	$AR_n \xrightarrow{p} \infty$	$\hat{c}_{AR,n}(1-\alpha) \xrightarrow{p} \infty$
	$AR_{CR,n} = O_p(1)$	$\hat{c}_{AR,CR,n}(1-\alpha) = O_p(1)$

Table 1: Test Statistics and Bootstrap Critical Values under Distant Alternatives

Note: This table summarizes the properties of the Wald and AR statistics and bootstrap critical values when the assumptions in Theorems 5.2(i) and 6.2 hold. T_n and $T_{CR,n}$ denote the Wald statistics without and with CCE, respectively, and $\hat{c}_n(1-\alpha)$ and $\hat{c}_{CR,n}(1-\alpha)$ denote their corresponding bootstrap critical values. AR_n and $AR_{CR,n}$ denote the AR statistics without and with CCE, respectively, and $\hat{c}_{AR,n}(1-\alpha)$ and $\hat{c}_{AR,CR,n}(1-\alpha)$ denote their corresponding bootstrap critical values.

6 Asymptotic Results for AR Tests

The size control of the bootstrap Wald tests with or without CCE relies on Assumption 3(i), which rules out overall weak identification in which all clusters are weak. In the case that the parameter of interest may be weakly identified in all clusters, we may consider the inference on the full vector of β_n. In this section, we consider bootstrap AR tests, with or without CCE, as defined in Section 2.3.2.

Assumption 5. $||\hat{A}_z - A_z||_{op} = o_p(1)$, *where* A_z *is a* $d_z \times d_z$ *symmetric deterministic weighting matrix such that* $0 < c \leq \lambda_{min}(A_z) \leq \lambda_{max}(A_z) \leq C < \infty$ *for some constants* c *and* C.

Theorem 6.1 below shows that, in the general case with multiple IVs, the limiting null rejection probability of the AR_n-based bootstrap test does not exceed the nominal level α, and that of the $AR_{CR,n}$ test does not exceed α by more than $1/2^{q-1}$ when $q > d_z$, irrespective of IV strength.

Theorem 6.1. *Suppose Assumption 1 holds and $\beta_n = \beta_0$. For AR_n, further suppose Assumption 5 holds. For $AR_{CR,n}$, further suppose $q > d_z$. Then,*

$$
\begin{aligned}
\alpha - \frac{1}{2^{q-1}} &\leq \liminf_{n\to\infty} \mathbb{P}\{AR_n > \hat{c}_{AR,n}(1-\alpha)\} \\
&\leq \limsup_{n\to\infty} \mathbb{P}\{AR_n > \hat{c}_{AR,n}(1-\alpha)\} \leq \alpha, \text{ and} \\
\alpha - \frac{1}{2^{q-1}} &\leq \liminf_{n\to\infty} \mathbb{P}\{AR_{CR,n} > \hat{c}_{AR,CR,n}(1-\alpha)\} \\
&\leq \limsup_{n\to\infty} \mathbb{P}\{AR_{CR,n} > \hat{c}_{AR,CR,n}(1-\alpha)\} \leq \alpha + \frac{1}{2^{q-1}}.
\end{aligned}
$$

Several remarks are in order. First, for the bootstrap AR test studentized by CCE, we require the number of IVs to be smaller than the number of clusters because otherwise, the CCE may not be invertible. Second, the behavior of wild bootstrap for other weak-IV-robust statistics proposed in the literature is more complicated as they depend on an adjusted sample Jacobian matrix (e.g., see Kleibergen (2005), Andrews (2016), Andrews and Mikusheva (2016), and Andrews and Guggenberger (2019), among others). Further complication therefore arises when all the clusters are weak. For example, with few clusters this adjusted Jacobian is no longer asymptotically independent from the score. In Appendix S.H, we establish the validity of wild bootstrap for these statistics with at least one strong cluster. Third, for the weak-IV-robust subvector inference, one may use a projection approach (Dufour and Taamouti, 2005) after implementing the wild bootstrap AR tests for β_n, but the result may be conservative. Alternative subvector inference methods (e.g., see Section 5.3 in Andrews et al. (2019) and the references therein) provide a power improvement over the projection approach under the framework with a large number of observations/clusters. However, it is unclear whether they can be applied to our setting with a small number of clusters.[12]

Next, to study the power of the AR_n-based bootstrap test against the local alternative, we let $\lambda_\beta = I_{d_x}$ for $\mathcal{H}_{1,n}$ in (2) so that $\mu = \mu_\beta$, and we impose the following condition.

Assumption 6. *(i) $Q_{\widetilde{Z}X} \neq 0$. (ii) There exists a scalar a_j for each $j \in J$ such that $Q_{\widetilde{Z}X,j} = a_j Q_{\widetilde{Z}X}$.*

If $d_x = d_z = 1$, Assumption 6(i) implies strong identification. When $d_z > 1$ or $d_x > 1$, Assumption 6(i) only rules out the case that $Q_{\widetilde{Z}X}$ is a zero matrix, while allowing it to be nonzero but not of full column rank. As noted below, this means the AR test can still have local power in

[12]It is unknown whether the asymptotic critical values given by these approaches will still be valid with a small number of clusters. Also, Wang and Doko Tchatoka (2018) point out that bootstrap tests based on the subvector statistics therein may not be robust to weak IVs even under conditional homoskedasticity.

some direction even without strong identification. Assumption 6(ii) is similar to Assumption 3(ii). In particular, it holds automatically if $d_x = d_z = 1$ and Assumption 6(i) holds.

Theorem 6.2. *Suppose Assumptions 1, 5, and 6 hold. Further suppose that there exists a subset J_s of J such that $\min_{j \in J_s} a_j > 0$, $a_j = 0$ for each $j \in J \backslash J_s$, and $\lceil |\boldsymbol{G}|(1-\alpha) \rceil \leq |\boldsymbol{G}| - 2^{q-q_s+1}$, where $q_s = |J_s|$ and a_j is defined in Assumption 6. Then, under $\mathcal{H}_{1,n}$ with $\lambda_\beta = I_{d_x}$,*

$$\lim_{||Q_{\widetilde{Z}X}\mu||_2 \to \infty} \liminf_{n \to \infty} \mathbb{P}\{AR_n > \hat{c}_{AR,n}(1-\alpha)\} = 1.$$

Two remarks are in order. First, notice that the bootstrap AR test not studentized by CCE has power against local alternatives as long as $||Q_{\widetilde{Z}X}\mu||_2 \to \infty$, which may hold even when β_n is not strongly identified and $Q_{\widetilde{Z}X}$ is not of full column rank. Second, when $d_z = 1$, we have $1\{AR_n > \hat{c}_{AR,n}(1-\alpha)\} = 1\{AR_{CR,n} > \hat{c}_{AR,CR,n}(1-\alpha)\}$, which implies the AR_n and $AR_{CR,n}$-based bootstrap tests have the same power against local alternatives. However, such a power equivalence does not hold when $d_z > 1$. Indeed, unlike the Wald test with CCE in Section 5, we cannot establish the power of the AR statistics with CCE for the general case. More specifically, for the Wald statistic $T_{CR,n}$, we compute its CCE with the estimated residual $\hat{\varepsilon}_{i,j}$, which causes the Wald CCE to be bounded in probability under the local alternative, and thus, the statistic $T_{CR,n}$ to diverge with the local parameter μ. On the other hand, we need to impose the null to compute CCE for $AR_{CR,n}$, which implies that the test statistic is bounded in probability even when μ diverges. Therefore, our proving strategy for the power of $T_{CR,n}$ cannot be applied to $AR_{CR,n}$. Table 1 summarizes the properties of the AR statistics and their corresponding bootstrap critical values when the assumptions in Theorem 6.2 hold. Indeed, consistent with the theoretical difference mentioned above, we observe in Section 7 that the $AR_{CR,n}$-based bootstrap test has inferior finite sample power properties compared with its AR_n-based counterpart. Furthermore, Table 1 also suggests that the bootstrap Wald test studentized with CCE has power advantage over the other tests under strong identification.

Next, we show that in the specific case with one endogenous regressor and one IV (i.e., $d_x = 1$ and $d_z = 1$), if the wild bootstrap procedure described in this section is applied to the unstudentized Wald statistic T_n, then the resulting test will be asymptotically equivalent to the AR_n-based bootstrap test, both under the null and the alternative. More precisely, for $T_n = |\sqrt{n}(\hat{\beta} - \beta_0)|$, where $\hat{\beta}$ is the TSLS estimator, the wild bootstrap generates

$$T_n^{s*}(g) = \left| \widehat{Q}_{\widetilde{Z}X}^{-1} \frac{1}{\sqrt{n}} \sum_{j \in J} \sum_{i \in I_{n,j}} g_j \widetilde{Z}_{i,j} \hat{\varepsilon}_{i,j}^r \right|. \quad (17)$$

Notice that in this case, the restricted TSLS estimator $\hat{\gamma}^r$ and the restricted OLS estimator $\bar{\gamma}^r$ are the same, which implies the restricted residuals $\hat{\varepsilon}_{i,j}^r$ and $\bar{\varepsilon}_{i,j}^r$ defined in Sections 4 and 6, respectively, are the same. Let $\hat{c}_n^s(1-\alpha)$ denote the $(1-\alpha)$-th quantile of $\{T_n^{s*}(g)\}_{g \in \mathbf{G}}$. We show this equivalence result in Theorem 6.3.

Theorem 6.3. *Suppose that $d_x = d_z = 1$, $\liminf_{n \to \infty} \mathbb{P}(\widehat{Q}_{\widetilde{Z}X} \neq 0) = 1$, $\hat{\beta}$ is the TSLS estimator, and Assumption 1(iv) holds. Then,*

$$\liminf_{n \to \infty} \mathbb{P}\{\phi_n^s = \phi_n^{ar}\} = 1,$$

where $\phi_n^s = 1\{T_n > \hat{c}_n^s(1-\alpha)\}$ and $\phi_n^{ar} = 1\{AR_n > \hat{c}_n^{ar}(1-\alpha)\}$.

Several remarks are in order. First, with one endogenous variable and one IV, the Jacobian matrix $\widehat{Q}_{\widetilde{Z}X}$ is just a scalar, which shows up in both the Wald statistic T_n and the bootstrap critical value. After the cancellation of $\widehat{Q}_{\widetilde{Z}X}$, T_n and its critical value are numerically the same as their AR counterparts, which leads to Theorem 6.3. Second, for $\widehat{Q}_{\widetilde{Z}X}$ to be cancelled, we only need $\liminf_{n \to \infty} \mathbb{P}\{\widehat{Q}_{\widetilde{Z}X} \neq 0\} = 1$, which is very mild. It holds when at least one of $\widetilde{Z}_{i,j}$ and $X_{i,j}$ is continuously distributed. Even when both $\widetilde{Z}_{i,j}$ and $X_{i,j}$ are discrete, it still holds if there exists at least one strong cluster, i.e., $Q_{\widetilde{Z}X} \neq 0$, where $Q_{\widetilde{Z}X}$ is the probability limit of $\widehat{Q}_{\widetilde{Z}X}$. When both $\widetilde{Z}_{i,j}$ and $X_{i,j}$ are discrete and $Q_{\widetilde{Z}X} = 0$, this condition can still hold. For example, some type of CLT may still hold such that $\sqrt{n}\widehat{Q}_{\widetilde{Z}X} \xrightarrow{d} N(c, \sigma^2)$. As $N(c, \sigma^2)$ is continuous, the condition still holds. Third, the robustness of the T_n-based bootstrap test in (17) does not carry over to the general case with multiple IVs as T_n and its bootstrap statistic can no longer be reduced to their AR counterparts. Fourth, the robustness to weak IV cannot be extended to the $T_{CR,n}$-based bootstrap test, for which we have to further bootstrap the CCE.

In Section S.A in the Supplement, we further show that our inference procedure allows for cluster-level exogenous variables in $W_{i,j}$, which are invariant within each cluster, and that our procedure allows for the error-component model in (10). We also explain why our assumptions rule out the cases with cluster-level variables in IVs and fully-interacted IVs, which are constructed by interacting the base IVs with all the cluster dummies.

7 Monte Carlo Simulation

In this section, we investigate the finite sample performance of the wild bootstrap tests and alternative methods. We consider a simulation design similar to that in Section IV of Canay et al. (2021) and extend theirs to the IV model. The data are generated as

$$X_{i,j} = \gamma + Z_{i,j}^\top \Pi_{z,j} + \sigma(Z_{i,j})\left(a_{v,j} + v_{i,j}\right), \quad y_{i,j} = \gamma + X_{i,j}\beta + \sigma(Z_{i,j})\left(a_{\varepsilon,j} + \varepsilon_{i,j}\right),$$

for $i = 1, ..., n$ and $j = 1, ..., q$. The number of clusters q equals 10, and the cluster size n_j is set to be 50 for $j = 1, ..., 5$, and 25 for $j = 6, ..., 10$, respectively. The total sample size n therefore equals 375.[13] The disturbances $(\varepsilon_{i,j}, v_{i,j})$, cluster effects $(a_{\varepsilon,j}, a_{v,j})$, IVs $Z_{i,j}$, and $\sigma(Z_{i,j})$ are specified as follows:

$$(\varepsilon_{i,j}, u_{i,j})^\top \sim N(0, I_2), \quad v_{i,j} = \rho\varepsilon_{i,j} + (1-\rho^2)^{1/2} u_{i,j}, \quad (a_{\varepsilon,j}, a_{u,j})^\top \sim N(0, I_2),$$
$$a_{v,j} = \rho a_{\varepsilon,j} + (1-\rho^2)^{1/2} a_{u,j}, \quad Z_{i,j} \sim N(0, I_{d_z}), \quad \text{and} \quad \sigma(Z_{i,j}) = \left(\sum_{k=1}^{d_z} Z_{i,j,k}\right)^2,$$

where I_{d_z} is the $d_z \times d_z$ identity matrix, $Z_{i,j,k}$ denotes the k-th element of $Z_{i,j}$, and $\rho \in \{0, 0.1, 0.2, \cdots, 0.9, 0.99\}$ corresponds to the degree of endogeneity. We let $W_{i,j}$ be just all the cluster dummies and the first-stage coefficients $\Pi_{z,j} = (\Pi_0/\sqrt{d_z}, ..., \Pi_0/\sqrt{d_z})^\top$ for $j = 1, ..., 5$, while $\Pi_{z,j} = 0.5 \cdot (\Pi_0/\sqrt{d_z}, ..., \Pi_0/\sqrt{d_z})^\top$ for $j = 6, ..., 10$, with $\Pi_0 \in \{0.5, 1, 2\}$. Such a DGP satisfies our Assumptions. Specifically, we have $\widetilde{Z}_{i,j} = Z_{i,j} - \frac{1}{n_j}\sum_{i \in I_{n,j}} Z_{i,j}$ and $\frac{1}{\sqrt{n_j}}\sum_{i \in I_{n,j}} \widetilde{Z}_{i,j}\sigma(Z_{i,j})(a_{\varepsilon,j} + \varepsilon_{i,j})$ is asymptotically normal conditional on $a_{\varepsilon,j}$. [14] The number of Monte Carlo and bootstrap replications equal 5,000 and 500, respectively. The nominal level α is set at 10%. The values of β and γ are set at 0 and 1, respectively. For T_n

[13]We also did simulations with homogeneity in within-cluster sample size and IV strength (e.g., $n_j = 50$ and same $\Pi_{z,j}$ for all j), and the patterns are similar to those reported here. Results are omitted for brevity but are available upon request.

[14]This is because $\left(\frac{1}{\sqrt{n_j}}\sum_{i \in I_{n,j}} Z_{i,j}\left(\sum_{k=1}^{d_z} Z_{i,j,k}^2\right), \left[\frac{1}{\sqrt{n_j}}\sum_{i \in I_{n,j}} Z_{i,j}\right]\left[\frac{1}{n_j}\sum_{i \in I_{n,j}}\left(\sum_{k=1}^{d_z} Z_{i,j,k}^2\right)\right], \frac{1}{\sqrt{n_j}}\sum_{i \in I_{n,j}} Z_{i,j}\left(\sum_{k=1}^{d_z} Z_{i,j,k}^2\right)\varepsilon_{i,j}\right)$ are jointly asymptotically normal given $(a_{\varepsilon,j})_{j \in J}$. Then, we have

$$\frac{1}{\sqrt{n_j}}\sum_{i \in I_{n,j}} \widetilde{Z}_{i,j}\sigma(Z_{i,j})(a_{\varepsilon,j} + \varepsilon_{i,j}) = \frac{a_{\varepsilon,j}}{\sqrt{n_j}}\sum_{i \in I_{n,j}} Z_{i,j}\left(\sum_{k=1}^{d_z} Z_{i,j,k}^2\right) + \frac{1}{\sqrt{n_j}}\sum_{i \in I_{n,j}} Z_{i,j}\left(\sum_{k=1}^{d_z} Z_{i,j,k}^2\right)\varepsilon_{i,j}$$
$$- \left[\frac{1}{\sqrt{n_j}}\sum_{i \in I_{n,j}} Z_{i,j}\right]\left[\frac{a_{\varepsilon,j}}{n_j}\sum_{i \in I_{n,j}}\left(\sum_{k=1}^{d_z} Z_{i,j,k}^2\right)\right] + o_p(1),$$

which is asymptotically normal given $(a_{\varepsilon,j})_{j \in J}$.

and AR_n, we set the weighting matrices $\hat{A}_r = 1$ and $\hat{A}_z = I_{d_z}$.

Figure 1 reports the null empirical rejection frequencies of the tests that are based on the IV estimators, including the T_n and $T_{CR,n}$-based wild bootstrap procedures, the group-based t-tests of IM, and the randomization tests of CRS. The results of the tests are based on TSLS when $d_z = 1$, and we further report the results based on TSLS, LIML, and FULL when $d_z = 3$. Following the recommendation in the literature, we set the tuning parameter of FULL to be 1.[15] Several observations are in order. First, in general size distortions increase when the IVs become weak, the degree of endogeneity becomes high, or the number of IVs becomes large. Second, IM and CRS tests with LIML or FULL show a substantial size improvement over their counterparts with TSLS, as these tests are based on cluster-level estimates (i.e., we run the cluster-by-cluster IV regressions), which could produce serious finite sample bias if TSLS is employed for the over-identified cases. Similarly, the LIML or FULL-based bootstrap tests, no matter studentized or unstudentized, show a size improvement over their TSLS-based counterparts. Third, overall the wild bootstrap procedures compare favorably with the alternatives, and the unstudentized bootstrap Wald tests (T_n) have the smallest size distortions across different settings of IV strength, degree of endogeneity, and number of IVs. In particular, the size control of the T_n-based tests with LIML or FULL remains excellent when the degree of over-identification increases.

Figure 2 reports the null rejection frequencies of AR tests, including the $AR_{CR,n}$-based asymptotic tests, which reject the null when the square of the corresponding test statistic exceeds $\chi^2_{d_z,1-\alpha}$, the $1-\alpha$ quantile of the chi-squared distribution with d_z degrees of freedom. Additionally, it reports the rejection frequencies of the wild bootstrap AR tests in Section 6 that are based on AR_n and $AR_{CR,n}$, respectively. We notice from Figure 2 that $AR_{CR,n}$-based asymptotic tests control the size but under-reject in the over-identified cases ($d_z = 3$).[16] By contrast, the bootstrap AR tests always have rejection frequencies very close to 10%.

Figure 3 compares the power properties of the wild bootstrap tests. For the Wald test, we focus on LIML and FULL estimators as they have better size control than their TSLS counterparts in the over-identified case. We also include the $AR_{CR,n}$-based asymptotic test to compare its power with

[15]In this case FULL is best unbiased to a second order among k-class estimators under normal errors (Rothenberg, 1984).

[16]The null rejection probabilities of the $AR_{CR,n}$-based asymptotic test decrease toward zero when d_z approaches q. When d_z is equal to q, the value of $AR_{CR,n}$ will be exactly equal to d_z (or q), and thus has no variation (for $\bar{f} = \left(\bar{f}_1, ..., \bar{f}_q\right)^\top$ and $\bar{f}_j = n^{-1}\sum_{i\in I_{n,j}} f_{i,j}$, $AR_{CR,n} = \iota_q^\top \bar{f}\left(\bar{f}^\top \bar{f}\right)^{-1}\bar{f}^\top \iota_q = \iota_q^\top \iota_q = d_z$ as long as $\bar{f}$ is invertible, where ι_q denotes a q-dimensional vector of ones). By contrast, the AR_n-based bootstrap test works well even when d_z is larger than q.

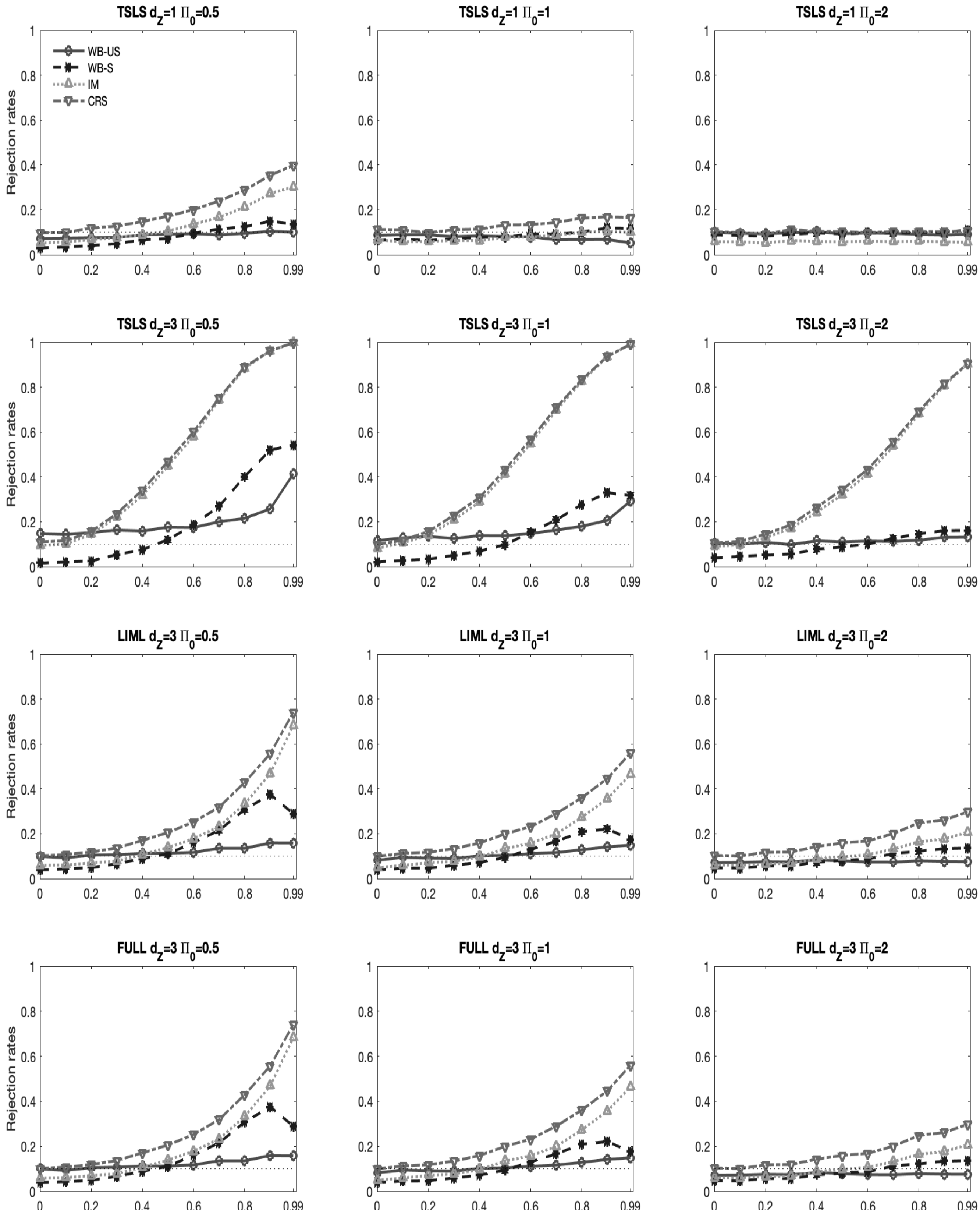

Figure 1: Null Rejection Probabilities of the Tests based on IV Estimators

Note: "WB-US" (solid line with circle), "WB-S" (dashed line with star), "IM" (dotted line with upward-pointing triangle), and "CRS" (dash-dotted line with downward-pointing triangle) denote the T_n and $T_{CR,n}$-based wild bootstrap tests, IM tests, and CRS tests, respectively.

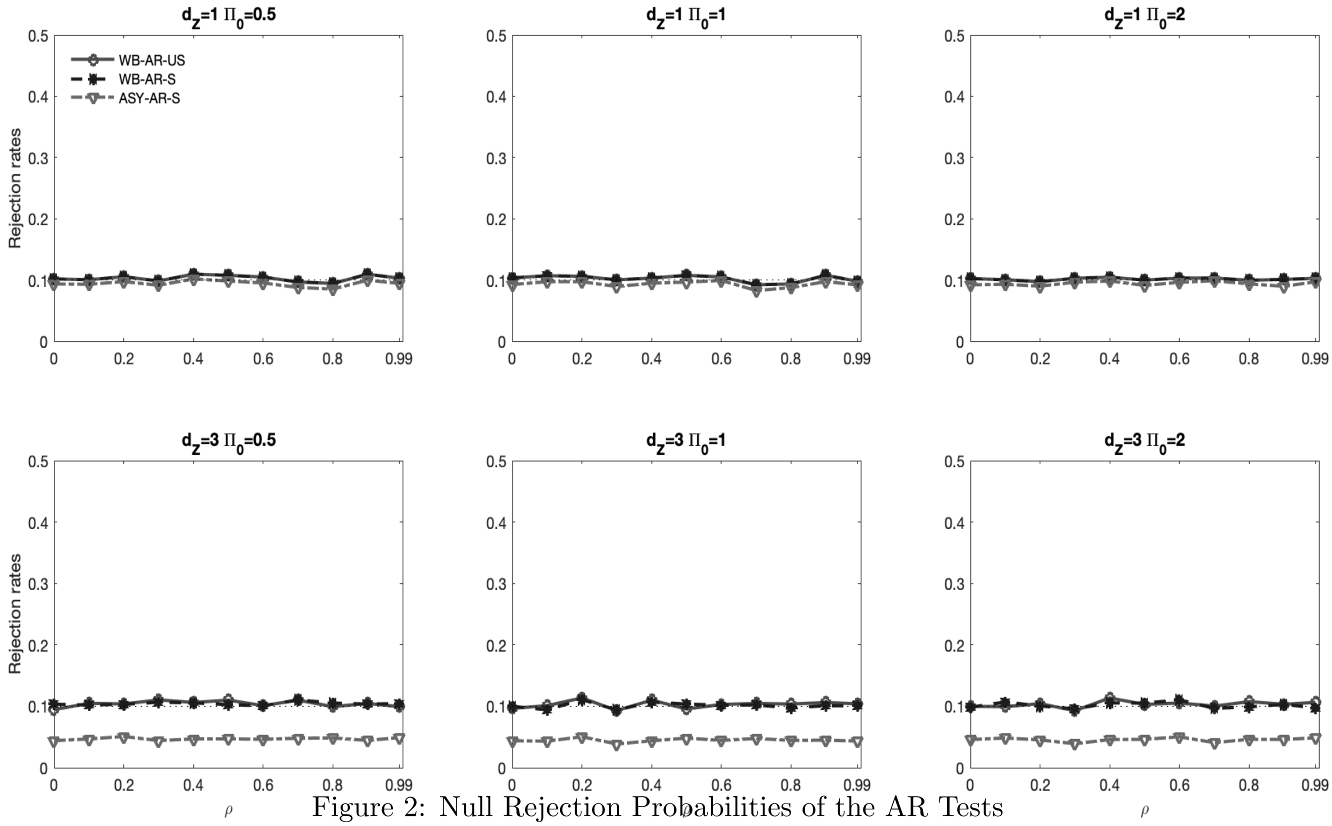

Figure 2: Null Rejection Probabilities of the AR Tests

Note: "WB-AR-US" (solid line with circle) and "WB-AR-S" (dashed line with star) denote the AR_n and $AR_{CR,n}$-based wild bootstrap tests, while "ASY-AR-S" (dash-dotted line with downward-pointing triangle) denote the $AR_{CR,n}$-based asymptotic tests, respectively.

those of the bootstrap AR tests. We let the number of IVs be 2, $\Pi_0 \in \{1, 2\}$, and $\rho \in \{0.1, 0.4, 0.7\}$.[17] First, overall the power ranking among the bootstrap tests is as follows (from the highest to the lowest): (1) the bootstrap Wald tests with CCE, (2) the bootstrap Wald tests without CCE, (3) the bootstrap AR tests without CCE, and (4) the bootstrap AR tests with CCE. Second, more specifically, we notice that among the AR tests, the AR_n-based bootstrap tests (the bootstrap AR tests without CCE) have the highest power, followed by the $AR_{CR,n}$-based bootstrap tests, which is in line with our theoretical analysis in Section 6. The $AR_{CR,n}$-based asymptotic tests have the lowest power among the AR tests, which is in line with the under-rejections found for the over-identified cases in Figure 2. Third, the T_n-based bootstrap tests with FULL have remarkable power advantage over those with LIML, especially when $\Pi_0 = 1$. This may be due to the fact that FULL has finite moments, and is thus less dispersed than LIML. Last, in line with our theory in Section 5, the $T_{CR,n}$-based bootstrap tests (the bootstrap Wald tests studentized by CCE) with both LIML and FULL are more powerful than their T_n-based counterparts and therefore may be preferred when

[17] Simulation results for other settings show similar patterns and are available upon request.

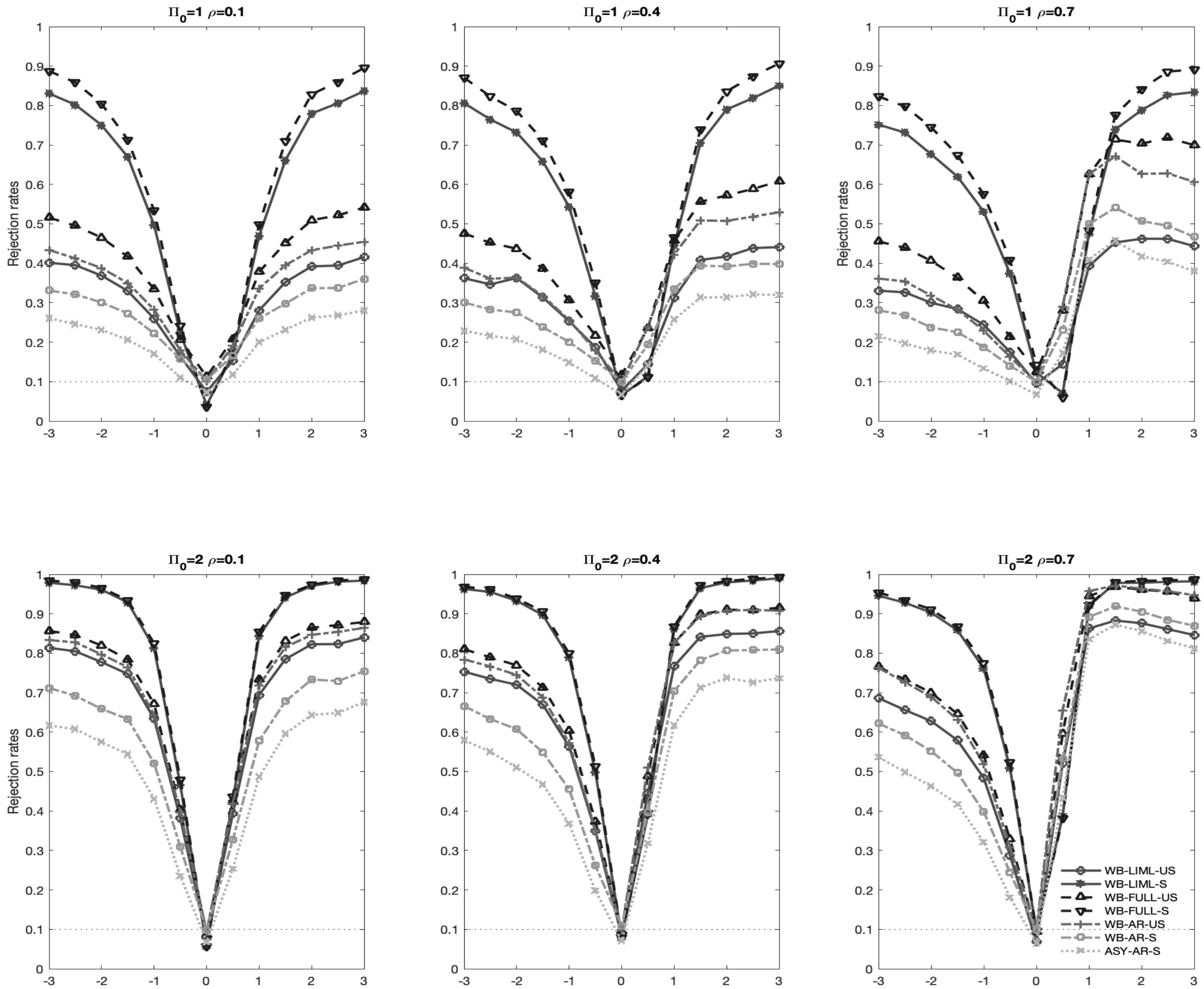

Figure 3: Power of Wild Bootstrap Wald and AR Tests for $d_z = 2$

Note: "WB-LIML-US" (solid line with circle), "WB-LIML-S" (solid line with star), "WB-FULL-US" (dashed line with upward-pointing triangle), and "WB-FULL-S" (dashed line with downward-pointing triangle) denote the T_n and $T_{CR,n}$-based wild bootstrap tests with LIML and FULL, respectively. "ASY-AR-S" (dotted line with cross) denotes the $AR_{CR,n}$-based AR tests with asymptotic CVs, while "WB-AR-US" (dash-dotted line with plus) and "WB-AR-S" (dash-dotted line with square) denote the AR_n and $AR_{CR,n}$-based wild bootstrap AR tests, respectively.

identification is strong.

8 Empirical Application

In an influential study, Autor et al. (2013) analyze the effect of rising Chinese import competition on US local labor markets between 1990 and 2007, when the share of total US spending on Chinese goods increased substantially from 0.6% to 4.6%. The dataset of Autor et al. (2013) includes 722 commuting zones (CZs) that cover the entire mainland US. In this section, we further analyze the region-wise effects of such import exposure by applying IV regression with the proposed wild bootstrap procedures to three Census Bureau-designated regions: South, Midwest and West, with 16, 12, and 11 states, respectively, in each region.[18]

We let the outcome variable ($y_{i,j}$) denote the decadal change in average individual log weekly wage in a given CZ. The endogenous variable ($X_{i,j}$) is the change in Chinese import exposure per worker in a CZ, which is instrumented ($Z_{i,j}$) by Chinese import growth in other high-income countries.[19] In addition, the exogenous variables ($W_{i,j}$) include the characteristic variables of CZs and decade specified in Autor et al. (2013) as well as state fixed effects. Our regressions are based on the CZ samples in each region, and the samples are clustered at the state level, following Autor et al. (2013). Besides the results for the full sample, we also report those for female and male samples separately.

The main result of the IV regression for the three regions is given in Table 2, with the number of observations (n) and clusters (q) for each region. As we have only one endogenous variable and one IV, the TSLS estimator is used throughout this section. We further construct the 90% bootstrap confidence sets (CSs) constructed by inverting the corresponding T_n, $T_{CR,n}$, and AR_n-based wild bootstrap tests with a 10% nominal level. The computation of the bootstrap CSs was conducted over the parameter space $[-10, 10]$ with a step size of 0.01, and the number of bootstrap draws is set at 2,000 for each step.

We highlight the main findings below. First, the results in Table 2 suggest that there may exist regional heterogeneity in terms of the average effect of Chinese imports on wages in local labor markets. For instance, the TSLS estimates for the South and West regions equal -0.97 and -1.05, respectively, while that for the Midwest region equals -0.025. That is, a $\$1,000$ per worker increase in a CZ's exposure to Chinese imports is estimated to reduce average weekly earnings by 0.97, 1.05, and 0.025 log points, respectively, for the three regions (the corresponding TSLS estimate in Autor et al. (2013) for the entire mainland US is -0.76). Second, only the effect on CZs in the South is significantly different from zero at the 10% level under all the three wild bootstrap CSs, while the effects on CZs in the other two regions are not. Third, we notice that the effect on West is only significant under the studentized bootstrap Wald CS (the CS is $[-1.22, -0.47]$). Fourth, compared

[18] The Northeast region is not included in the study because of the relatively small number of states (9) and small number of CZs in each state (e.g., Connecticut and Rhode Island have only 2 CZs).

[19] See Sections I.B and III.A in Autor et al. (2013) for a detailed definition of these variables.

with that for the South, the wider CSs for the Midwest and West may be due to relatively weak identification, which our bootstrap AR procedure is able to guard against. Table 2 also reports the results for female and male samples. We find that across all the regions, the effects are more substantial for the male samples. Furthermore, the effects for both female and male samples in the South are significantly different from zero. Last, the studentized bootstrap Wald CSs have the shortest length among the three types of CSs in most cases in Table 2, which is in line with our power results in Section 5.

Gender	Region	n	q	Estimate	Unstud Wald CS	Stud Wald CS	Unstud AR CS
All	South	578	16	-0.97	[-1.71, -0.58]	[-1.61, -0.43]	[-1.70, -0.58]
	Midwest	504	12	-0.025	[-0.69, 0.83]	[-0.60, 0.75]	[-0.69, 0.83]
	West	276	11	-1.05	[-1.50, 0.25]	[-1.22, -0.47]	[-1.50, 0.24]
Female	South	578	16	-0.81	[-1.48, -0.41]	[-1.40, -0.26]	[-1.48, -0.41]
	Midwest	504	12	0.024	[-0.64, 0.74]	[-0.56, 0.67]	[-0.64, 0.74]
	West	276	11	-0.61	[-1.46, 0.74]	[-0.91, 0.30]	[-1.46, 0.75]
Male	South	578	16	-1.08	[-1.90, -0.65]	[-1.78, -0.53]	[-1.91, -0.65]
	Midwest	504	12	-0.17	[-1.02, 0.80]	[-0.86, 0.77]	[-1.01, 0.82]
	West	276	11	-1.26	[-1.99, 0.68]	[-1.58, 0.42]	[-2.00, 0.69]

Table 2: IV regressions of Autor et al. (2013) with all, female, and male samples for three US regions

9 Conclusion and Practical Recommendations

In this paper, we study the wild bootstrap inference for IV regressions in the framework of a small number of clusters. For the Wald tests with and without CCE, we extend Davidson and MacKinnon (2010)'s WREC bootstrap procedure to allow for the setting of few clusters and cluster-level heterogeneity in IV strength. For the full-vector inference, we further develop wild bootstrap AR tests that control size asymptotically irrespective of IV strength.

Our results have several important implications for applied works. First, if at least one of the clusters is strong so that overall the structural parameters of interest are well identified, we recommend the bootstrap Wald test studentized by CCE because of its superior power properties. Second, for the over-identified case, instead of using the bootstrap Wald tests with TSLS, we recommend to use those based on estimators with reduced finite sample bias such as Fuller's modified LIML estimator. Third, for the full-vector inference, if researchers are concerned that all clusters are weak and thus would like to implement a weak-identification-robust procedure, then we recommend the bootstrap AR test without CCE because of its power advantage over alternative AR tests that are studentized by CCE.

Appendices

Sections A and B contain the proofs of Theorems 4.1 and 5.2, respectively. The proofs of other results in the paper are relegated to the Online Supplement.

Appendix A Proof of Theorem 4.1

Let $\mathbb{S} \equiv \mathbf{R}^{d_z \times d_x} \times \mathbf{R}^{d_z \times d_z} \times \otimes_{j \in J} \mathbf{R}^{d_z} \times \mathbf{R}^{d_r \times d_r}$ and write an element $s \in \mathbb{S}$ by $s = (s_1, s_2, \{s_{3,j} : j \in J\}, s_4)$ where $s_{3,j} \in \mathbf{R}^{d_z}$ for any $j \in J$. Define the function T: $\mathbb{S} \to \mathbf{R}$ to be given by

$$T(s) = \left\| \lambda_\beta^\top \left(s_1^\top s_2^{-1} s_1 \right)^{-1} s_1^\top s_2^{-1} \left(\sum_{j \in J} s_{3,j} \right) \right\|_{s_4} \tag{18}$$

for any $s \in \mathbb{S}$ such that s_2 and $s_1^\top s_2^{-1} s_1$ are invertible and let $T(s) = 0$ 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 $gs = (s_1, s_2, \{g_j s_{3,j} : j \in J\}, s_4)$. For any $s \in \mathbb{S}$ and $\mathbf{G'} \subseteq \mathbf{G}$, denote the ordered values of $\{T(gs) : g \in \mathbf{G'}\}$ by $T^{(1)}(s|\mathbf{G'}) \leq \ldots \leq T^{(|\mathbf{G'}|)}(s|\mathbf{G'})$. In addition, for any $\mathbf{G'} \subseteq \mathbf{G}$, denote the ordered values of $\{T_n^*(g) : g \in \mathbf{G'}\}$ by $T_n^{*(1)}(\mathbf{G'}) \leq \ldots \leq T_n^{*(|\mathbf{G'}|)}(\mathbf{G'})$.

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

$$S_n = \left(\widehat{Q}_{\widetilde{Z}X}, \widehat{Q}_{\widetilde{Z}\widetilde{Z}}, \left\{ \frac{1}{\sqrt{n}} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} \varepsilon_{i,j} : j \in J \right\}, \hat{A}_r \right), \quad \widehat{S}_n = \left(\widehat{Q}_{\widetilde{Z}X}, \widehat{Q}_{\widetilde{Z}\widetilde{Z}}, \left\{ \frac{1}{\sqrt{n}} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} \hat{\varepsilon}_{i,j}^r : j \in J \right\}, \hat{A}_r \right).$$

Let E_n denote the event $E_n = I\left\{ \widehat{Q}_{\widetilde{Z}X} \text{ is of full rank value and } \widehat{Q}_{\widetilde{Z}\widetilde{Z}} \text{ is invertible} \right\}$, and Assumptions 2-3 imply that $\liminf_{n \to \infty} \mathbb{P}\{E_n = 1\} = 1$. Also let $T_n^*(g) = 0$ if $E_n = 0$.

We first give the proof for the Wald statistic based on TSLS. Note that whenever $E_n = 1$ and $\mathcal{H}_0$ is true, the Frisch-Waugh-Lovell theorem implies that

$$\begin{aligned} T_n &= \left\| \sqrt{n} \left(\lambda_\beta^\top \hat{\beta}_{tsls} - \lambda_0 \right) \right\|_{\hat{A}_r} = \left\| \sqrt{n} \lambda_\beta^\top \left(\hat{\beta}_{tsls} - \beta_n \right) \right\|_{\hat{A}_r} \\ &= \left\| \lambda_\beta^\top \widehat{Q}^{-1} \widehat{Q}_{\widetilde{Z}X}^\top \widehat{Q}_{\widetilde{Z}\widetilde{Z}}^{-1} \sum_{j \in J} \frac{1}{\sqrt{n}} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} \varepsilon_{i,j} \right\|_{\hat{A}_r} = T(\iota S_n), \end{aligned} \tag{19}$$

where $\widehat{Q} = \widehat{Q}_{\widetilde{Z}X}^\top \widehat{Q}_{\widetilde{Z}\widetilde{Z}}^{-1} \widehat{Q}_{\widetilde{Z}X}$ and $\iota \in \mathbf{G}$ is a $q \times 1$ vector of ones.

In the following, we divide the proof into three steps. In the first step, we show

$$T_n^*(g) = T(g\widehat{S}_n) + o_p(1) \text{ for any } g \in \mathbf{G}. \tag{20}$$

In the second step, we show

$$T(g\widehat{S}_n) = T(gS_n) + o_p(1) \text{ for any } g \in \mathbf{G}. \tag{21}$$

In the last step, we prove the desired result.

Step 1. By the continuous mapping theorem, it suffices to show $\widehat{Q}^*_{\widetilde{Z}X}(g) = \widehat{Q}_{\widetilde{Z}X} + o_p(1)$. Note that

$$\widehat{Q}^*_{\widetilde{Z}X}(g) = \frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}} \widetilde{Z}_{i,j}X^{*\top}_{i,j}(g) = \frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}} \widetilde{Z}_{i,j}\left(X_{i,j} + (g_j - 1)\tilde{v}_{i,j}\right)^\top.$$

Therefore, it suffices to show $\frac{1}{n_j}\sum_{i\in I_{n,j}} \widetilde{Z}_{i,j}\tilde{v}_{i,j} = o_p(1)$ for all $j \in J$. Recall $\overline{Z}_{i,j}$ is just $\widetilde{Z}_{i,j}$ interacted with all the cluster dummies and $\widetilde{\Pi}_{\overline{Z}}$ is the OLS coefficient of $\overline{Z}_{i,j}$ defined in (7). Denote $\widetilde{\Pi}_{\overline{Z},j}$ as the j-th block of $\widetilde{\Pi}_{\overline{Z}}$, which corresponds to the OLS coefficient of $\widetilde{\Pi}_{\overline{Z},j}$. We have

$$\frac{1}{n_j}\sum_{i\in I_{n,j}} \widetilde{Z}_{i,j}\tilde{v}_{i,j} = \frac{1}{n_j}\sum_{i\in I_{n,j}}\left(\widetilde{Z}_{i,j}X_{i,j} - \widetilde{Z}_{i,j}\widetilde{Z}^\top_{i,j}\widetilde{\Pi}_{\overline{Z},j} - \widetilde{Z}_{i,j}W^\top_{i,j}\widetilde{\Pi}_w\right) = \widehat{Q}_{\widetilde{Z}X,j} - \widehat{Q}_{\widetilde{Z}\widetilde{Z},j}\widetilde{\Pi}_{\overline{Z},j} + o_p(1), \tag{22}$$

where the second equality in (22) holds by $\frac{1}{n_j}\sum_{i\in I_{n,j}} \widetilde{Z}_{i,j}W^\top_{i,j} = o_p(1)$ and $\widetilde{\Pi}_w = O_p(1)$. In particular,

$$\widetilde{\Pi}_w = \left(\widehat{Q}_{\widetilde{W}\widetilde{W}} - \widehat{Q}_{\widetilde{W}\hat{\varepsilon}}\widehat{Q}^{-1}_{\hat{\varepsilon}\hat{\varepsilon}}\widehat{Q}_{\hat{\varepsilon}\widetilde{W}}\right)^{-1}\left(\widehat{Q}_{\widetilde{W}X} - \widehat{Q}_{\widetilde{W}\hat{\varepsilon}}\widehat{Q}^{-1}_{\hat{\varepsilon}\hat{\varepsilon}}\widehat{Q}_{\hat{\varepsilon}X}\right),$$

where $\widehat{Q}_{\widetilde{W}\widetilde{W}} = \frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}} \widetilde{W}_{i,j}\widetilde{W}^\top_{i,j}$, $\widehat{Q}_{\widetilde{W}\hat{\varepsilon}} = \frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}} \widetilde{W}_{i,j}\hat{\varepsilon}_{i,j}$, $\widehat{Q}_{\hat{\varepsilon}\hat{\varepsilon}} = \frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}} \hat{\varepsilon}^2_{i,j}$, $\widetilde{W}_{i,j} = W_{i,j} - \widehat{\Gamma}^\top_{w,j}\widetilde{Z}_{i,j}$, and $\widehat{\Gamma}_{w,j} = \widehat{Q}^{-1}_{\widetilde{Z}\widetilde{Z},j}\widehat{Q}_{\widetilde{Z}W,j}$. Notice that by $\widehat{Q}_{\widetilde{Z}W,j} = o_p(1)$, $\widehat{\Gamma}_{w,j} = o_p(1)$ so that

$$\widehat{Q}_{\widetilde{W}\widetilde{W}} = \widehat{Q}_{WW} + o_p(1). \tag{23}$$

Similarly, we have

$$\widehat{Q}_{\widetilde{W}\hat{\varepsilon}} = \widehat{Q}_{W\hat{\varepsilon}} + o_p(1) = o_p(1), \tag{24}$$

where the second equality follows from $\widehat{Q}_{W\hat{\varepsilon}} = 0$ by the first-order condition of the k-class estimators. Furthermore, $\widehat{Q}_{\hat{\varepsilon}\hat{\varepsilon}} \geq c > 0$ by Assumption 2(iii) and $\widehat{Q}_{\hat{\varepsilon}\hat{\varepsilon}} \geq \frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}} \dot{\varepsilon}^2_{i,j}$, where $\dot{\varepsilon}_{i,j}$ is the residual from the cluster-level projection of $\varepsilon_{i,j}$ on $W_{i,j}$. Therefore, $\widehat{Q}_{\widetilde{W}\widetilde{W}} - \widehat{Q}_{\widetilde{W}\hat{\varepsilon}}\widehat{Q}^{-1}_{\hat{\varepsilon}\hat{\varepsilon}}\widehat{Q}_{\hat{\varepsilon}\widetilde{W}} = \widehat{Q}_{WW} + o_p(1)$ by combining (23) and (24), and further by Assumption 1(iv),

$$\left(\widehat{Q}_{\widetilde{W}\widetilde{W}} - \widehat{Q}_{\widetilde{W}\hat{\varepsilon}}\widehat{Q}^{-1}_{\hat{\varepsilon}\hat{\varepsilon}}\widehat{Q}_{\hat{\varepsilon}\widetilde{W}}\right)^{-1} = O_p(1). \tag{25}$$

Next, we define $\widehat{Q}_{\widetilde{Z}\dot{X},j} = \frac{1}{n_j}\sum_{i\in I_{n,j}} \widetilde{Z}_{i,j}\dot{X}^\top_{i,j}$ and recall $\widehat{Q}_{\widetilde{Z}\widetilde{Z},j} = \frac{1}{n_j}\sum_{i\in I_{n,j}} \widetilde{Z}_{i,j}\widetilde{Z}^\top_{i,j}$, where $\dot{X}_{i,j} = X_{i,j} - \widetilde{\Pi}^\top_w W_{i,j} - \widetilde{\Pi}^\top_{\hat{\varepsilon}}\hat{\varepsilon}_{i,j}$, and $\widetilde{\Pi}_w$ and $\widetilde{\Pi}_{\hat{\varepsilon}}$ are the OLS coefficients of $W_{i,j}$ and $\hat{\varepsilon}_{i,j}$, respectively, from regressing $X_{i,j}$ on $(\overline{Z}^\top_{i,j}, W^\top_{i,j}, \hat{\varepsilon}_{i,j})^\top$

using the entire sample. Then, we have

$$\widetilde{\Pi}_{\overline{Z},j} = \widehat{Q}^{-1}_{\widetilde{Z}\widetilde{Z},j}\widehat{Q}_{\widetilde{Z}\dot{X},j} = \widehat{Q}^{-1}_{\widetilde{Z}\widetilde{Z},j}\left[\widehat{Q}_{\widetilde{Z}X,j} - \widehat{Q}_{\widetilde{Z}W,j}\widetilde{\Pi}_w - \widehat{Q}_{\widetilde{Z}\hat{\varepsilon},j}\widetilde{\Pi}_{\hat{\varepsilon}}\right] = \widehat{Q}^{-1}_{\widetilde{Z}\widetilde{Z},j}\widehat{Q}_{\widetilde{Z}X,j} + o_p(1), \tag{26}$$

where we use the facts that $\widehat{Q}_{\widetilde{Z}W,j} = o_p(1)$, $\widetilde{\Pi}_w = O_p(1)$, $\widetilde{\Pi}_{\hat{\varepsilon}} = O_p(1)$, and $\widehat{Q}_{\widetilde{Z}\hat{\varepsilon},j} = o_p(1)$. In particular,

$$\widetilde{\Pi}_{\hat{\varepsilon}} = \left(\widehat{Q}_{\tilde{\varepsilon}\tilde{\varepsilon}} - \widehat{Q}_{\tilde{\varepsilon}W}\widehat{Q}^{-1}_{WW}\widehat{Q}_{W\tilde{\varepsilon}}\right)^{-1}\left(\widehat{Q}_{\tilde{\varepsilon}X} - \widehat{Q}_{\tilde{\varepsilon}W}\widehat{Q}^{-1}_{WW}\widehat{Q}_{WX}\right),$$

where $\widehat{Q}_{\tilde{\varepsilon}\tilde{\varepsilon}} = \frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}}\tilde{\varepsilon}^2_{i,j}$, $\widehat{Q}_{\tilde{\varepsilon}W} = \frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}}\tilde{\varepsilon}_{i,j}W^\top_{i,j}$, $\tilde{\varepsilon}_{i,j} = \hat{\varepsilon}_{i,j} - \widehat{\Gamma}^\top_{\hat{\varepsilon},j}\widetilde{Z}_{i,j}$, and $\widehat{\Gamma}_{\hat{\varepsilon},j} = \widehat{Q}^{-1}_{\widetilde{Z}\widetilde{Z},j}\widehat{Q}_{\widetilde{Z}\hat{\varepsilon},j}$. Then, by using similar arguments as those for (23), (24), and (25), we obtain

$$\widehat{Q}_{\tilde{\varepsilon}\tilde{\varepsilon}} = \widehat{Q}_{\hat{\varepsilon}\hat{\varepsilon}} + o_p(1),\ \widehat{Q}_{\tilde{\varepsilon}W} = \widehat{Q}_{\hat{\varepsilon}W} + o_p(1) = o_p(1),\ \text{and}\ \left(\widehat{Q}_{\tilde{\varepsilon}\tilde{\varepsilon}} - \widehat{Q}_{\tilde{\varepsilon}W}\widehat{Q}^{-1}_{WW}\widehat{Q}_{W\tilde{\varepsilon}}\right)^{-1} = O_p(1).$$

To see the last equality in (26), we note that

$$\begin{aligned}\widehat{Q}_{\widetilde{Z}\hat{\varepsilon},j} &= \frac{1}{n_j}\sum_{i\in I_{n,j}}\widetilde{Z}_{i,j}\hat{\varepsilon}_{i,j}\\ &= \frac{1}{n_j}\sum_{i\in I_{n,j}}\widetilde{Z}_{i,j}\varepsilon_{i,j} - \frac{1}{n_j}\sum_{i\in I_{n,j}}\widetilde{Z}_{i,j}X^\top_{i,j}(\hat{\beta}_{tsls} - \beta_n) - \frac{1}{n_j}\sum_{i\in I_{n,j}}\widetilde{Z}_{i,j}W^\top_{i,j}(\hat{\gamma}_{tsls} - \gamma) = o_p(1),\end{aligned}$$

where the last equality holds by Assumption 1(ii), Lemma S.B.3, and Lemma S.B.1. Plugging (26) into (22), we obtain the desired result that $\frac{1}{n_j}\sum_{i\in I_{n,j}}\widetilde{Z}_{i,j}\tilde{v}_{i,j} = o_p(1)$.

Step 2. We note that whenever $E_n = 1$, for every $g \in \mathbf{G}$,

$$\begin{aligned}\left|T(gS_n) - T(g\widehat{S}_n)\right| \le &\left\|\lambda^\top_\beta\widehat{Q}^{-1}\widehat{Q}^\top_{\widetilde{Z}X}\widehat{Q}^{-1}_{\widetilde{Z}\widetilde{Z}}\sum_{j\in J}\frac{n_j}{n}\frac{1}{n_j}\sum_{i\in I_{n,j}}g_j\widetilde{Z}_{i,j}X^\top_{i,j}\sqrt{n}(\beta_n - \hat{\beta}^r_{tsls})\right\|_{\hat{A}_r}\\ &+\left\|\lambda^\top_\beta\widehat{Q}^{-1}\widehat{Q}^\top_{\widetilde{Z}X}\widehat{Q}^{-1}_{\widetilde{Z}\widetilde{Z}}\sum_{j\in J}\frac{n_j}{n}\frac{1}{n_j}\sum_{i\in I_{n,j}}g_j\widetilde{Z}_{i,j}W^\top_{i,j}\sqrt{n}(\gamma - \hat{\gamma}^r_{tsls})\right\|_{\hat{A}_r}.\end{aligned} \tag{27}$$

By Lemma S.B.3, we have

$$\limsup_{n\to\infty}\mathbb{P}\left\{\left\|\lambda^\top_\beta\widehat{Q}^{-1}\widehat{Q}^\top_{\widetilde{Z}X}\widehat{Q}^{-1}_{\widetilde{Z}\widetilde{Z}}\sum_{j\in J}\frac{n_j}{n}\frac{1}{n_j}\sum_{i\in I_{n,j}}g_j\widetilde{Z}_{i,j}W^\top_{i,j}\sqrt{n}(\gamma - \hat{\gamma}^r_{tsls})\right\|_{\hat{A}_r} > \varepsilon; E_n = 1\right\} = 0. \tag{28}$$

Note under both cases in Assumption 3(ii), we have $\sum_{j\in J}\xi_j a_j = 1$ and

$$\lambda^\top_\beta Q^{-1}Q^\top_{\widetilde{Z}X}Q^{-1}_{\widetilde{Z}\widetilde{Z}}Q_{\widetilde{Z}X,j} = a_j\lambda^\top_\beta. \tag{29}$$

Then, for any $\varepsilon > 0$, we have

$$\limsup_{n\to\infty} \mathbb{P}\left\{ \left\| \lambda_\beta^\top \widehat{Q}^{-1} \widehat{Q}_{\widetilde{Z}X}^\top \widehat{Q}_{\widetilde{Z}\widetilde{Z}}^{-1} \sum_{j\in J} \frac{n_j}{n} \frac{1}{n_j} \sum_{i\in I_{n,j}} g_j \widetilde{Z}_{i,j} X_{i,j}^\top \sqrt{n}(\beta_n - \hat{\beta}_{tsls}^r) \right\|_{\hat{A}_r} > \varepsilon; E_n = 1 \right\}$$
$$= \limsup_{n\to\infty} \mathbb{P}\left\{ \left\| \sum_{j\in J} \xi_j g_j a_j \sqrt{n}(\lambda_\beta^\top \beta_n - \lambda_\beta^\top \hat{\beta}_{tsls}^r) \right\|_{\hat{A}_r} > \varepsilon; E_n = 1 \right\} = 0, \tag{30}$$

where the last equality holds because $\lambda_\beta^\top \hat{\beta}_{tsls}^r = \lambda_0$ under $\mathcal{H}_0$.

Note that $T(g\widehat{S}_n) = T(gS_n)$ whenever $E_n = 0$ as we have defined $T(s) = 0$ for any $s = (s_1, s_2, \{s_{3,j} : j \in J\}, s_4)$ whenever s_2 or $s_1^\top s_2^{-1} s_1$ is not invertible. Therefore, results in (27), (28) and (30) imply (21).

Step 3. Note that by Assumptions 1, 2, 4, and the continuous mapping theorem, we have

$$\left(\widehat{Q}_{\widetilde{Z}X}, \widehat{Q}_{\widetilde{Z}\widetilde{Z}}, \left\{ \frac{1}{\sqrt{n}} \sum_{i\in I_{n,j}} \widetilde{Z}_{i,j}\varepsilon_{i,j} : j \in J \right\}, \hat{A}_r \right) \xrightarrow{d} \left(Q_{\widetilde{Z}X}, Q_{\widetilde{Z}\widetilde{Z}}, \left\{ \sqrt{\xi_j}\mathcal{Z}_j : j \in J \right\}, A_r \right) \equiv S, \tag{31}$$

where $\xi_j > 0$ for all $j \in J$ by Assumption 1(iii). Therefore, we obtain from (20), (21), (31), and the continuous mapping theorem that

$$(T(S_n), \{T_n^*(g) : g \in \mathbf{G}\}) \xrightarrow{d} (T(S), \{T(gS) : g \in \mathbf{G}\}) .$$

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 (19) that

$$1\{T_n > \hat{c}_n(1-\alpha)\} = 1\left\{ T(S_n) > T_n^{*(k^*)}(\mathbf{G}) \right\}. \tag{32}$$

Since $\liminf_{n\to\infty} P\{E_n = 1\} = 1$, we have

$$\limsup_{n\to\infty} \mathbb{P}\{T_n > \hat{c}_n(1-\alpha))\} = \limsup_{n\to\infty} \mathbb{P}\{T_n > \hat{c}_n(1-\alpha)); E_n = 1\}$$
$$\leq \limsup_{n\to\infty} \mathbb{P}\{T(S_n) \geq T_n^{*(k^*)}(\mathbf{G}); E_n = 1\} \leq \mathbb{P}\{T(S) \geq T^{(k^*)}(S|\mathbf{G})\} \leq \alpha + 2^{1-q},$$

where the second inequality is due to the Portmanteau's theorem. To see the last inequality, we note that for all $g \in \mathbf{G}$, $\mathbb{P}\{T(gS) = T(-gS)\} = 1$, $\mathbb{P}\{T(gS) = T(\tilde{g}S)\} = 0$ for $\tilde{g} \notin \{g, -g\}$, and under the null, $T(gS)$ has the same distribution across $g \in \mathbf{G}$. Let $|\mathbf{G}| = 2^q$. Then, we have

$$|\mathbf{G}|\mathbb{E}1\{T(S) \geq T^{(k^*)}(S|\mathbf{G})\} = \mathbb{E}\sum_{g\in\mathbf{G}} 1\{T(gS) \geq T^{(k^*)}(S|\mathbf{G})\} \leq |\mathbf{G}| - k^* + 2,$$

which implies

$$\mathbb{E}1\{T(S) \geq T^{(k^*)}(S|\mathbf{G})\} \leq 1 - \frac{k^*}{|\mathbf{G}|} + \frac{2}{|\mathbf{G}|} \leq \alpha + \frac{1}{2^{q-1}}.$$

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

$$\begin{aligned}\liminf_{n\to\infty} \mathbb{P}\{T_n > \hat{c}_n(1-\alpha)\} &= \liminf_{n\to\infty} \mathbb{P}\{T(S_n) > T_n^{*(k^*)}(\mathbf{G})\} \geq \mathbb{P}\{T(S) > T^{(k^*)}(S|\mathbf{G})\} \\ &\geq \mathbb{P}\{T(S) > T^{(k^*+2)}(S|\mathbf{G})\} + \mathbb{P}\{T(S) = T^{(k^*+2)}(S|\mathbf{G})\} \geq \alpha - \frac{1}{2^{q-1}},\end{aligned}$$

where the first equality follows from (32), the first inequality follows from Portmanteau's theorem, the second inequality holds because $\mathbb{P}\{T^{(\mathbf{z}+2)}(S|\mathbf{G}) > T^{(\mathbf{z})}(S|\mathbf{G})\} = 1$ for any integer $\mathbf{z} \leq |\mathbf{G}| - 2$ by (18) and Assumption 1, and the last inequality 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.

Lemma S.B.1 and S.B.2 in the Supplement further show the other k-class estimators and their null-restricted and bootstrap counterparts are asymptotically equivalent to those of the TSLS estimator. Therefore, the results for LIML, FULL, and BA estimators can be derived in the same manner. ■

Appendix B Proof of Theorem 5.2

For the power analysis, we focus on the TSLS estimator. The results for other k-class estimators can be derived in the same manner given Lemma S.B.2. Recall a_j defined in Assumption 3. We further define

$$T_{CR,\infty}(g) = \left\| \lambda_\beta^\top \tilde{Q} \left[\sum_{j\in J} g_j \sqrt{\xi}_j \mathcal{Z}_j \right] + c_{0,g}\mu \right\|_{A_{r,CR,g}}, \tag{33}$$

where $\tilde{Q} = Q^{-1} Q_{\tilde{Z}X}^\top Q_{\tilde{Z}\tilde{Z}}^{-1}$, $Q = Q_{\tilde{Z}X}^\top Q_{\tilde{Z}\tilde{Z}} Q_{\tilde{Z}X}$, $c_{0,g} = \sum_{j\in J} \xi_j g_j a_j$, and

$$A_{r,CR,g} = \begin{pmatrix} \sum_{j\in J} \xi_j \left\{ \lambda_\beta^\top \tilde{Q} \left[g_j \mathcal{Z}_j - a_j \sqrt{\xi}_j \sum_{\tilde{j}\in J} \sqrt{\xi_{\tilde{j}}} g_{\tilde{j}} \mathcal{Z}_{\tilde{j}} \right] + \sqrt{\xi}_j (g_j - c_{0,g}) a_j \mu \right\} \\ \times \left\{ \lambda_\beta^\top \tilde{Q} \left[g_j \mathcal{Z}_j - a_j \sqrt{\xi}_j \sum_{\tilde{j}\in J} \sqrt{\xi_{\tilde{j}}} g_{\tilde{j}} \mathcal{Z}_{\tilde{j}} \right] + \sqrt{\xi}_j (g_j - c_{0,g}) a_j \mu \right\}^\top \end{pmatrix}^{-1}.$$

We order $\{T_{CR,\infty}(g)\}_{g\in\mathbf{G}}$ in ascending order: $(T_{CR,\infty})^{(1)} \leq \cdots \leq (T_{CR,\infty})^{|\mathbf{G}|}$. In the proof of Theorem 4.2, we have already shown that, under $\mathcal{H}_{1,n}$,

$$\sqrt{n}(\lambda_\beta^\top \hat{\beta}_{tsls} - \lambda_0) \xrightarrow{d} \sum_{j\in J} \left[\sqrt{\xi_j} \lambda_\beta^\top \tilde{Q} \mathcal{Z}_j \right] + \mu, \quad \sqrt{n}(\lambda_\beta^\top \hat{\beta}_{tsls,g}^* - \lambda_0) \xrightarrow{d} \sum_{j\in J} g_j \left[\sqrt{\xi_j} \lambda_\beta^\top \tilde{Q} \mathcal{Z}_j \right] + c_{0,g}\mu.$$

Next, we derive the limit of $\hat{A}_{r,CR}$ and $\hat{A}^*_{r,CR,g}$. We first note that

$$\begin{aligned}
\frac{1}{\sqrt{n}} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} \hat{\varepsilon}_{i,j} &= \frac{1}{\sqrt{n}} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} \left[\varepsilon_{i,j} - X_{i,j}^\top (\hat{\beta}_{tsls} - \beta_n) - W_{i,j}^\top (\hat{\gamma}_{tsls} - \gamma) \right] \\
&= \frac{1}{\sqrt{n}} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} \varepsilon_{i,j} - \xi_j \widehat{Q}_{\widetilde{Z}X,j} \sqrt{n} (\hat{\beta}_{tsls} - \beta_n) + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} \varepsilon_{i,j} - \xi_j \widehat{Q}_{\widetilde{Z}X,j} \widehat{Q}^{-1} \widehat{Q}_{\widetilde{Z}X} \widehat{Q}^{-1}_{\widetilde{Z}\widetilde{Z}} \sum_{\tilde{j} \in J} \frac{1}{\sqrt{n}} \sum_{i \in I_{n,\tilde{j}}} \widetilde{Z}_{i,\tilde{j}} \varepsilon_{i,\tilde{j}} + o_p(1) \\
&\xrightarrow{d} \sqrt{\xi_j} \mathcal{Z}_j - \xi_j Q_{\widetilde{Z}X,j} Q^{-1} Q_{\widetilde{Z}X} Q^{-1}_{\widetilde{Z}\widetilde{Z}} \sum_{\tilde{j} \in J} \sqrt{\xi_{\tilde{j}}} \mathcal{Z}_{\tilde{j}} + o_p(1),
\end{aligned}$$

where the first equality holds by Lemma S.B.3. This implies

$$\begin{aligned}
&\hat{\Omega}_{CR} \xrightarrow{d} \sum_{j \in J} \xi_j \left(\mathcal{Z}_j - \sqrt{\xi_j} Q_{\widetilde{Z}X,j} \tilde{Q} \sum_{\tilde{j} \in J} \sqrt{\xi_{\tilde{j}}} \mathcal{Z}_{\tilde{j}} \right) \left(\mathcal{Z}_j - \sqrt{\xi_j} Q_{\widetilde{Z}X,j} \tilde{Q} \sum_{\tilde{j} \in J} \sqrt{\xi_{\tilde{j}}} \mathcal{Z}_{\tilde{j}} \right)^\top, \\
&\hat{A}_{r,CR} \xrightarrow{d} A_{r,CR,\iota}, \quad \text{and} \quad T_{CR,n} \xrightarrow{d} T_{CR,\infty}(\iota),
\end{aligned}$$

where ι is a $q \times 1$ vector of ones. Similarly, we have

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} \hat{\varepsilon}^*_{i,j}(g) \\
&= \frac{1}{\sqrt{n}} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} \left[g_j \hat{\varepsilon}^r_{i,j} - X^{*\top}_{i,j}(g) (\hat{\beta}^*_{tsls,g} - \hat{\beta}^r_{tsls}) - W_{i,j}^\top (\hat{\gamma}^*_{tsls,g} - \hat{\gamma}^r_{tsls}) \right] \\
&= \frac{g_j}{\sqrt{n}} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} \hat{\varepsilon}^r_{i,j} - \xi_j \widehat{Q}^*_{\widetilde{Z}X,j}(g) \sqrt{n} (\hat{\beta}^*_{tsls,g} - \hat{\beta}^r_{tsls}) + o_p(1) \\
&= \frac{g_j}{\sqrt{n}} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} \varepsilon_{i,j} - \xi_j g_j \widehat{Q}_{\widetilde{Z}X,j} \sqrt{n} (\hat{\beta}^r_{tsls} - \beta_n) - \xi_j \widehat{Q}_{\widetilde{Z}X,j} \sqrt{n} (\hat{\beta}^*_{tsls,g} - \hat{\beta}^r_{tsls}) + o_p(1) \\
&= \frac{g_j}{\sqrt{n}} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} \varepsilon_{i,j} - \xi_j g_j Q_{\widetilde{Z}X,j} \sqrt{n} (\hat{\beta}^r_{tsls} - \beta_n) - \xi_j Q_{\widetilde{Z}X,j} \sqrt{n} (\hat{\beta}^*_{tsls,g} - \hat{\beta}^r_{tsls}) + o_p(1),
\end{aligned} \tag{34}$$

where $\widehat{Q}^*_{\widetilde{Z}X,j}(g) = \frac{1}{n_j} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} X^{*\top}_{i,j}(g)$, the second equality is by Lemma S.B.3, and the last equality holds because $\widehat{Q}_{\widetilde{Z}X,j} = Q_{\widetilde{Z}X,j} + o_p(1)$, $\widehat{Q}^*_{\widetilde{Z}X,j}(g) = Q_{\widetilde{Z}X,j} + o_p(1)$ proved in Step 1 of the proof of Theorem 4.1, $\sqrt{n}(\hat{\beta}^r_{tsls} - \beta_n) = O_p(1)$, and $\sqrt{n}(\hat{\beta}^*_{tsls,g} - \beta_n) = O_p(1)$. In addition, following the same arguments that lead to (S.C.2), we have

$$\begin{aligned}
\sqrt{n}(\hat{\beta}^*_{tsls,g} - \hat{\beta}^r_{tsls}) &= \tilde{Q} \sum_{\tilde{j} \in J} \sum_{i \in I_{n,\tilde{j}}} \frac{g_{\tilde{j}} \widetilde{Z}_{i,\tilde{j}} \varepsilon_{i,\tilde{j}}}{\sqrt{n}} + \tilde{Q} \sum_{\tilde{j} \in J} \xi_{\tilde{j}} g_{\tilde{j}} Q_{\widetilde{Z}X,\tilde{j}} \sqrt{n} (\beta_n - \hat{\beta}^r_{tsls}) + o_p(1) \\
&= \tilde{Q} \sum_{\tilde{j} \in J} \sum_{i \in I_{n,\tilde{j}}} \frac{g_{\tilde{j}} \widetilde{Z}_{i,\tilde{j}} \varepsilon_{i,\tilde{j}}}{\sqrt{n}} + \sum_{\tilde{j} \in J} \xi_{\tilde{j}} g_{\tilde{j}} a_{\tilde{j}} \sqrt{n} (\beta_n - \hat{\beta}^r_{tsls}) + o_p(1).
\end{aligned} \tag{35}$$

Note $\widehat{Q}_g^{*-1}\widehat{Q}_{\widetilde{Z}X}^{*\top}(g)\widehat{Q}_{\widetilde{Z}\widetilde{Z}}^{-1} \xrightarrow{p} \tilde{Q}$. Therefore, combining (34) and (35), we have

$$\begin{aligned}
&\lambda_\beta^\top \widehat{Q}_g^{*-1}\widehat{Q}_{\widetilde{Z}X}^{*\top}(g)\widehat{Q}_{\widetilde{Z}\widetilde{Z}}^{-1}\left[\frac{1}{\sqrt{n}}\sum_{i\in I_{n,j}}\widetilde{Z}_{i,j}\hat{\varepsilon}_{i,j}^*(g)\right] \\
&= \lambda_\beta^\top \tilde{Q}\left[\frac{g_j}{\sqrt{n}}\sum_{i\in I_{n,j}}\widetilde{Z}_{i,j}\varepsilon_{i,j} - \xi_j g_j Q_{\widetilde{Z}X,j}\sqrt{n}(\hat{\beta}_{tsls}^r - \beta_n) - \xi_j Q_{\widetilde{Z}X,j}\sqrt{n}(\hat{\beta}_{tsls,g}^* - \hat{\beta}_{tsls}^r)\right] + o_p(1) \\
&= \lambda_\beta^\top \tilde{Q}\left[\frac{g_j}{\sqrt{n}}\sum_{i\in I_{n,j}}\widetilde{Z}_{i,j}\varepsilon_{i,j}\right] - \xi_j g_j a_j \lambda_\beta^\top\sqrt{n}(\hat{\beta}_{tsls}^r - \beta_n) - \xi_j a_j \lambda_\beta^\top\sqrt{n}(\hat{\beta}_{tsls,g}^* - \hat{\beta}_{tsls}^r) + o_p(1) \\
&\xrightarrow{d} \lambda_\beta^\top \tilde{Q}\sqrt{\xi_j} g_j \mathcal{Z}_j - \xi_j a_j \lambda_\beta^\top \tilde{Q}\sum_{\tilde{j}\in J}\sqrt{\xi_{\tilde{j}}} g_{\tilde{j}}\mathcal{Z}_{\tilde{j}} + \xi_j(g_j - c_{0,g})a_j\mu,
\end{aligned}$$

where the second equality is by the fact that $\tilde{Q}Q_{\widetilde{Z}X,j} = a_j I_{d_x}$ and the last convergence is by the fact that $\lambda_\beta^\top\sqrt{n}(\hat{\beta}_{tsls}^r - \beta_n) = -\mu$. This implies

$$\hat{A}_{r,CR,g}^* \xrightarrow{d} A_{r,CR,g}, \quad \text{and thus,} \quad (T_{CR,n}, \{T_{CR,n}^*(g)\}_{g\in\mathbf{G}}) \xrightarrow{d} (T_{CR,\infty}(\iota), \{T_{CR,\infty}(g)\}_{g\in\mathbf{G}}).$$

By the Portmanteau theorem, we have

$$\liminf_{n\to\infty}\mathbb{P}\{T_{CR,n} > \hat{c}_{CR,n}(1-\alpha)\} \geq \mathbb{P}\left\{T_{CR,\infty}(\iota) > (T_{CR,\infty})^{(k^*)}\right\}.$$

We aim to show that, as $||\mu||_2 \to \infty$, we have

$$\mathbb{P}\left\{T_{CR,\infty}(\iota) > \max_{g\in\mathbf{G}_s} T_{CR,\infty}(g)\right\} \to 1, \tag{36}$$

where $\mathbf{G}_s = \mathbf{G}\backslash\mathbf{G}_w$, and $\mathbf{G}_w = \{g \in \mathbf{G} : g_j = g_{j'}, \forall j, j' \in J_s\}$. Then, given $|\mathbf{G}_s| = |\mathbf{G}| - 2^{q-q_s+1}$ and $k^* = \lceil|\mathbf{G}|(1-\alpha)\rceil \leq |\mathbf{G}| - 2^{q-q_s+1}$, (36) implies that as $||\mu||_2 \to \infty$,

$$\mathbb{P}\left\{T_{CR,\infty}(\iota) > (T_{CR,\infty})^{(k^*)}\right\} \geq \mathbb{P}\left\{T_{CR,\infty}(\iota) > \max_{g\in\mathbf{G}_s} T_{CR,\infty}(g)\right\} \to 1.$$

Therefore, it suffices to establish (36).

By (33), we see that

$$T_{CR,\infty}(\iota) = \left\|\lambda_\beta^\top\tilde{Q}\left[\sum_{j\in J} g_j\sqrt{\xi_j}\mathcal{Z}_j\right] + \mu\right\|_{A_{r,CR,\iota}},$$

and $A_{r,CR,\iota}$ is independent of μ as $c_{0,\iota} = 1$. In addition, we have $\lambda_{\min}(\tilde{Q}^\top\lambda_\beta^\top A_{r,CR,\iota}\lambda_\beta\tilde{Q}) > 0$ with probability one. Therefore, for any $e > 0$, we can find a sufficiently small constant $c > 0$ and a sufficiently large constant $M > 0$ such that with probability greater than $1 - e$, for any μ,

$$T_{CR,\infty}(\iota) \geq c||\mu||_2^2 - M. \tag{37}$$

On the other hand, for $g \in \mathbf{G}_s$, we can write $T_{CR,\infty}(g)$ as

$$T_{CR,\infty}(g) = \left\{ (N_{0,g} + c_{0,g}\mu)^\top \left[\sum_{j \in J} \xi_j (N_{j,g} + c_{j,g}\mu)(N_{j,g} + c_{j,g}\mu)^\top \right]^{-1} (N_{0,g} + c_{0,g}\mu) \right\},$$

where for $j \in J$, $N_{0,g} = \lambda_\beta^\top \tilde{Q} \left[\sum_{j \in J} g_j \sqrt{\xi_j} \mathcal{Z}_j \right]$, $N_{j,g} = \lambda_\beta^\top \tilde{Q} \left[g_j \mathcal{Z}_j - a_j \sqrt{\xi_j} \sum_{\tilde{j} \in J} \sqrt{\xi_{\tilde{j}}} g_{\tilde{j}} \mathcal{Z}_{\tilde{j}} \right]$, and $c_{j,g} = \sqrt{\xi_j}(g_j - c_{0,g})a_j$.

We claim that for $g \in \mathbf{G}_s$, $c_{j,g} \neq 0$ for some $j \in J_s$. Suppose it does not hold, then it implies that $g_j = c_{0,g}$ for all $j \in J_s$, i.e., for all $j \in J_s$, g_j shares the same sign, and thus, contradicts the definition of $\mathbf{G}_s$. Therefore, combining the claim with the assumption that $\min_{j \in J_s} |a_j| > 0$, we have $\min_{g \in \mathbf{G}_s} \sum_{j \in J} \xi_j c_{j,g}^2 > 0$.

In addition, we note that

$$\begin{aligned}
& \sum_{j \in J} \xi_j (N_{j,g} + c_{j,g}\mu)(N_{j,g} + c_{j,g}\mu)^\top \\
= & \sum_{j \in J} \xi_j N_{j,g} N_{j,g}^\top + \sum_{j \in J} \xi_j c_{j,g} N_{j,g} \mu^\top + \sum_{j \in J} \xi_j c_{j,g} \mu N_{j,g}^\top + (\sum_{j \in J} \xi_j c_{j,g}^2) \mu\mu^\top \\
\equiv & M_1 + M_2\mu^\top + \mu M_2^\top + \bar{c}^2 \mu\mu^\top,
\end{aligned}$$

where we denote $M_1 = \sum_{j \in J} \xi_j N_{j,g} N_{j,g}^\top$, $M_2 = \sum_{j \in J} \xi_j c_{j,g} N_{j,g}$, and $\bar{c}^2 = \sum_{j \in J} \xi_j c_{j,g}^2$. For notation ease, we suppress the dependence of $(M_1, M_2, \bar{c})$ on g. Then, we have

$$M_1 + M_2\mu^\top + \mu M_2^\top + \bar{c}^2\mu\mu^\top = M_1 - \frac{M_2 M_2^\top}{\bar{c}^2} + \left(\frac{M_2}{\bar{c}} + \bar{c}\mu \right) \left(\frac{M_2}{\bar{c}} + \bar{c}\mu \right)^\top.$$

Note for any $d_r \times 1$ vector u, by the Cauchy–Schwarz inequality,

$$u^\top \left(M_1 - \frac{M_2 M_2^\top}{\bar{c}^2} \right) u = \sum_{j \in J} \xi_j (u^\top N_{j,g})^2 - \frac{\left(\sum_{j \in J} \xi_j u^\top N_{j,g} c_{j,g} \right)^2}{\sum_{j \in J} \xi_j c_{j,g}^2} \geq 0,$$

where the equal sign holds if and only if there exist $(u, g) \in \Re^{d_r} \times \mathbf{G}_s$ such that

$$\frac{u^\top N_{1,g}}{c_{1,g}} = \cdots = \frac{u^\top N_{q,g}}{c_{q,g}},$$

which has probability zero if $q > d_r$ as $\{N_{j,g}\}_{j \in J}$ are independent and non-degenerate normal vectors. Therefore, the matrix $\mathbb{M} \equiv M_1 - \frac{M_2 M_2^\top}{\bar{c}^2}$ is invertible with probability one. Specifically, denote $\mathbb{M}$ as $\mathbb{M}(g)$ to highlight its dependence on g. We have $\max_{g \in \mathbf{G}_s} (\lambda_{\min}(\mathbb{M}(g)))^{-1} = O_p(1)$. In addition, denote $\frac{M_2}{\bar{c}} + \bar{c}\mu$ as $\mathbb{V}$, which is a $d_r \times 1$ vector. Then, we have

$$\left[\sum_{j \in J} \xi_j (N_{j,g} + c_{j,g}\mu)(N_{j,g} + c_{j,g}\mu)^\top \right]^{-1} = [\mathbb{M} + \mathbb{V}\mathbb{V}^\top]^{-1} = \mathbb{M}^{-1} - \mathbb{M}^{-1}\mathbb{V}(1 + \mathbb{V}^\top \mathbb{M}^{-1}\mathbb{V})^{-1}\mathbb{V}^\top \mathbb{M}^{-1},$$

where the second equality is due to the Sherman–Morrison–Woodbury formula.

Next, we note that

$$N_{0,g} + c_{0,g}\mu = N_{0,g} + c_{0,g}\left(\frac{\mathbb{V}}{\bar{c}} - \frac{M_2}{\bar{c}^2}\right) \equiv \mathbb{M}_0 + \frac{c_{0,g}}{\bar{c}}\mathbb{V},$$

where $\mathbb{M}_0 = N_{0,g} - \frac{c_{0,g}M_2}{\bar{c}^2} = N_{0,g} - \frac{c_{0,g}(\sum_{j\in J}\xi_j c_{j,g}N_{j,g})}{\sum_{j\in J}\xi_j c_{j,g}^2}$. With these notations, we have

$$\begin{aligned}
&(N_{0,g} + c_{0,g}\mu)^\top \left[\sum_{j\in J}\xi_j(N_{j,g} + c_{j,g}\mu)(N_{j,g} + c_{j,g}\mu)^\top\right]^{-1}(N_{0,g} + c_{0,g}\mu) \\
&= \left(\mathbb{M}_0 + \frac{c_{0,g}}{\bar{c}}\mathbb{V}\right)^\top (\mathbb{M}^{-1} - \mathbb{M}^{-1}\mathbb{V}(1 + \mathbb{V}^\top\mathbb{M}^{-1}\mathbb{V})^{-1}\mathbb{V}^\top\mathbb{M}^{-1})\left(\mathbb{M}_0 + \frac{c_{0,g}}{\bar{c}}\mathbb{V}\right) \\
&\leq 2\mathbb{M}_0^\top(\mathbb{M}^{-1} - \mathbb{M}^{-1}\mathbb{V}(1 + \mathbb{V}^\top\mathbb{M}^{-1}\mathbb{V})^{-1}\mathbb{V}^\top\mathbb{M}^{-1})\mathbb{M}_0 \\
&+ \frac{2c_{0,g}^2}{\bar{c}^2}\mathbb{V}^\top(\mathbb{M}^{-1} - \mathbb{M}^{-1}\mathbb{V}(1 + \mathbb{V}^\top\mathbb{M}^{-1}\mathbb{V})^{-1}\mathbb{V}^\top\mathbb{M}^{-1})\mathbb{V} \\
&\leq 2\mathbb{M}_0^\top\mathbb{M}^{-1}\mathbb{M}_0 + \frac{2c_{0,g}^2}{\bar{c}^2}\frac{\mathbb{V}^\top\mathbb{M}^{-1}\mathbb{V}}{1 + \mathbb{V}^\top\mathbb{M}^{-1}\mathbb{V}} \leq 2\mathbb{M}_0^\top\mathbb{M}^{-1}\mathbb{M}_0 + \frac{2c_{0,g}^2}{\bar{c}^2} \\
&\leq 2(\lambda_{\min}(\mathbb{M}))^{-1}\left\|N_{0,g} - \frac{c_{0,g}(\sum_{j\in J}\xi_j c_{j,g}N_{j,g})}{\sum_{j\in J}\xi_j c_{j,g}^2}\right\|^2 + \frac{2c_{0,g}^2}{\sum_{j\in J}\xi_j c_{j,g}^2} \equiv C(g),
\end{aligned}$$

where the first inequality is due to the fact that $(u+v)^\top A(u+v) \leq 2(u^\top Au + v^\top Av)$ for some $d_r \times d_r$ positive semidefinite matrix A and $u, v \in \Re^{d_r}$, the second inequality holds due to the fact that $\mathbb{M}^{-1}\mathbb{V}(1+\mathbb{V}^\top\mathbb{M}^{-1}\mathbb{V})^{-1}\mathbb{V}^\top\mathbb{M}^{-1}$ is positive semidefinite, the third inequality holds because $\mathbb{V}^\top\mathbb{M}^{-1}\mathbb{V}$ is a nonnegative scalar, and the last inequality holds by substituting in the expressions for $\mathbb{M}_0$ and $\bar{c}$.

Then, we have

$$\max_{g\in\mathbf{G}_s} T_{CR,\infty}(g) \leq \max_{g\in\mathbf{G}_s} C(g). \tag{38}$$

Combining (37) and (38), we have, as $||\mu||_2 \to \infty$,

$$\begin{aligned}
\liminf_{n\to\infty}\mathbb{P}\{T_{CR,n} > \hat{c}_{CR,n}(1-\alpha)\} &\geq \mathbb{P}\left\{T_{CR,\infty}(\iota_q) > (T_{CR,\infty})^{(k^*)}\right\} \\
&\geq \mathbb{P}\left\{T_{CR,\infty}(\iota_q) > \max_{g\in\mathbf{G}_s} T_{CR,\infty}(g)\right\} \\
&= 1 - \mathbb{P}\left\{T_{CR,\infty}(\iota_q) \leq \max_{g\in\mathbf{G}_s} T_{CR,\infty}(g)\right\} \\
&\geq 1 - \mathbb{P}\left\{c||\mu||_2^2 - M \leq \max_{g\in\mathbf{G}_s} C(g)\right\} - e \to 1 - e,
\end{aligned}$$

where the second inequality is by the fact that $k^* \leq |\mathbf{G}_s|$, and thus, $(T_{CR,\infty})^{(k^*)} \leq \max_{g\in\mathbf{G}_s} T_{CR,\infty}(g)$ and the last convergence holds because $\max_{g\in\mathbf{G}_s} C(g) = O_p(1)$ and does not depend on μ. As e is arbitrary, we can let $e \to 0$ and obtain the desired result.

For the proof of part (ii), let $\tilde{c}_{CR,n}(1-\alpha)$ denote the $(1-\alpha)$ quantile of

$$\left\{|\sqrt{n}(\lambda_\beta^\top \hat{\beta}^*_{tsls,g} - \lambda_0)|/\sqrt{\lambda_\beta^\top \widehat{V} \lambda_\beta} : g \in \mathbf{G}\right\},$$

i.e., the bootstrap statistic $T_n^*(g)$ studentized by the original CCE instead of the bootstrap CCE. Then, because $d_r = 1$, we have

$$1\{T_n > \hat{c}_n(1-\alpha)\} = 1\{T_{CR,n} > \tilde{c}_{CR,n}(1-\alpha)\}.$$

Therefore, it suffices to show that $\tilde{c}_{CR,n}(1-\alpha) > \hat{c}_{CR,n}(1-\alpha)$ with large probability as $||\mu||_2$ becomes sufficiently large. First, note that as $||\mu||_2 \to \infty$, we have $\tilde{c}_{CR,n}(1-\alpha) \xrightarrow{p} \infty$, since $|\sqrt{n}(\lambda_\beta^\top \hat{\beta}^*_{tsls,g} - \lambda_0)| \xrightarrow{p} \infty$ for all $g \in \mathbf{G}$, and $\left(\lambda_\beta^\top \widehat{V} \lambda_\beta\right)^{-1} = O_p(1)$ as

$$\left(\lambda_\beta^\top \widehat{V} \lambda_\beta\right)^{-1} \xrightarrow{d} A_{r,CR,\iota},$$

and $A_{r,CR,\iota}$ does not depend on μ. On the other hand, $\hat{c}_{CR,n}(1-\alpha) = O_p(1)$ as has been proved in Part (i). Therefore, we have for any $e > 0$, there exists a constant $c_\mu > 0$, such that when $||\mu||_2 > c_\mu$, $\tilde{c}_{CR,n}(1-\alpha) > \hat{c}_{CR,n}(1-\alpha)$ with probability greater than $1-e$. This concludes the proof. ■

References

Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge (2022): "When should you adjust standard errors for clustering?" Tech. rep., Working paper, MIT.

Abadie, A., J. Gu, and S. Shen (2019): "Instrumental variable estimation with first stage heterogeneity," Tech. rep., Working paper, MIT.

Acemoglu, D., D. Cantoni, S. Johnson, and J. A. Robinson (2011): "The consequences of radical reform: The French revolution," *American Economic Review*, 101, 3286–3307.

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, 46–63.

Andrews, D. W. and X. Cheng (2012): "Estimation and inference with weak, semi-strong, and strong identification," *Econometrica*, 80, 2153–2211.

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

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

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

Andrews, I. and A. Mikusheva (2016): "Conditional inference with a functional nuisance parameter," *Econometrica*, 84, 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. and M. Kolesár (2021): "One instrument to rule them all: The bias and coverage of just-id IV," Tech. rep., National Bureau of Economic Research.

Angrist, J. D. (1990): "Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records," *American Economic Review*, 80, 313–336.

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

Angrist, J. D. and J.-S. Pischke (2008): *Mostly harmless econometrics: An empiricist's companion*, Princeton university press.

Autor, D., D. Dorn, and G. H. Hanson (2013): "The China syndrome: Local labor market effects of import competition in the United States," *American Economic Review*, 103, 2121–68.

Bekker, P. (1994): "Alternative approximations to the distributions of instrumental variable estimators," *Econometrica*, 62, 657–681.

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

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, 414–427.

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

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

Canay, I. A., A. Santos, and A. M. Shaikh (2021): "The wild bootstrap with a "small" number of "large" clusters," *Review of Economics and Statistics*, 103, 346–363.

Card, D., D. S. Lee, Z. Pei, and A. Weber (2015): "Inference on causal effects in a generalized regression kink design," *Econometrica*, 83, 2453–2483.

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

Chen, Q. and Z. Fang (2019): "Improved inference on the rank of a matrix," *Quantitative Economics*, 10, 1787–1824.

Conley, T. G. and C. R. Taber (2011): "Inference with "difference in differences" with a small number of policy changes," *The Review of Economics and Statistics*, 93, 113–125.

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

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

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

Dufour, J.-M. and M. Taamouti (2005): "Projection-based statistical inference in linear structural models with possibly weak instruments," *Econometrica*, 73, 1351–1365.

Ferman, B. and C. Pinto (2019): "Inference in differences-in-differences with few treated groups and heteroskedasticity," *Review of Economics and Statistics*, 101, 452–467.

Finlay, K. and L. M. Magnusson (2014): "Bootstrap methods for inference with cluster sample IV models," Tech. rep., Working Papers 14-12, The University of Western Australia, Department of Economics.

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

Fuller, W. A. (1977): "Some properties of a modification of the limited information estimator," *Econometrica*, 45, 939–953.

Glitz, A. and E. Meyersson (2020): "Industrial espionage and productivity," *American Economic Review*, 110, 1055–1103.

Hagemann, A. (2017): "Cluster-robust bootstrap inference in quantile regression models," *Journal of the American Statistical Association*, 112, 446–456.

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

——— (2019b): "Placebo inference on treatment effects when the number of clusters is small," *Journal of Econometrics*, 213, 190–209.

——— (2020): "Inference with a single treated cluster," *arXiv preprint arXiv:2010.04076*.

Hansen, B. E. and S. Lee (2019): "Asymptotic theory for clustered samples," *Journal of Econometrics*, 210, 268–290.

Hwang, J. (2021): "Simple and trustworthy cluster-robust GMM inference," *Journal of Econometrics*, 222, 993–1023.

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

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

Imbens, G. W. and J. D. Angrist (1994): "Identification and estimation of local average treatment effects," *Econometrica*, 62, 467–475.

Imbens, G. W. and M. Kolesar (2016): "Robust standard errors in small samples: Some practical advice," *Review of Economics and Statistics*, 98, 701–712.

Kleibergen, F. (2005): "Testing parameters in GMM without assuming that they are identified," *Econometrica*, 73, 1103–1124.

Kleibergen, F. and R. Paap (2006): "Generalized reduced rank tests using the singular value decomposition," *Journal of Econometrics*, 133, 97–126.

Lee, D. S. and T. Lemieux (2010): "Regression discontinuity designs in economics," *Journal of Economic Literature*, 48, 281–355.

Lee, D. S., J. McCrary, M. J. Moreira, and J. R. Porter (2021): "Valid t-ratio inference for IV," Tech. rep., National Bureau of Economic Research.

MacKinnon, J. G. (2021): "Fast cluster bootstrap methods for linear regression models," *Econometrics and Statistics*, Forthcoming.

MacKinnon, J. G., M. Ø. Nielsen, and M. D. Webb (2021): "Wild bootstrap and asymptotic inference with multiway clustering," *Journal of Business & Economic Statistics*, 39, 505–519.

——— (2022): "Cluster-robust inference: A guide to empirical practice," *Journal of Econometrics*, Forthcoming.

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

Menzel, K. (2021): "Bootstrap with cluster-dependence in two or more dimensions," *Econometrica*, forthcoming.

Mikusheva, A. and L. Sun (2022): "Inference with many weak instruments," *Review of Economic Studies*, forthcoming.

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

Moreira, M. J., J. Porter, and G. Suarez (2009): "Bootstrap validity for the score test when instruments may be weak," *Journal of Econometrics*, 149, 52–64.

Nagar, A. L. (1959): "The bias and moment matrix of the genaralized k-class estimators of the parameters in simultaneous equations," *Econometrica*, 27, 575–595.

Robin, J.-M. and R. J. Smith (2000): "Tests of rank," *Econometric Theory*, 16, 151–175.

Rogall, T. (2021): "Mobilizing the masses for genocide," *American Economic Review*, 111, 41–72.

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, 4–60.

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, 557–586.

Stock, J. H. and J. H. Wright (2000): "GMM with weak identification," *Econometrica*, 68, 1055–1096.

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, 188–211.

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

Young, A. (2021): "Leverage, heteroskedasticity and instrumental variables in practical application," Tech. rep., London School of Economics.

Online Supplement to "Wild Bootstrap for Instrumental Variables Regressions with Weak and Few Clusters"*

Wenjie Wang† and Yichong Zhang‡

September 6, 2022

Abstract

This document gathers together all the supplementary materials to the main paper. Section S.A discuss the restrictions we impose on IVs, control variables, and unobserved cluster-level fixed effects. Section S.B establishes the equivalence between TSLS and other k-class estimators. Sections S.C–S.G contain proofs of Theorems 2.2, 2.3, 2.5–2.7, respectively. Sections S.H-S.I discuss wild bootstrap inference with other weak-IV-robust statistics.

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

JEL codes: C12, C26, C31

S.A Cluster-level Variables and Interactions

S.B Equivalence Among k-Class Estimators

We define $\hat{\beta}_L$ as the k-class estimator with $\hat{\kappa}_L$ for $L \in \{\text{tsls}, \text{liml}, \text{full}, \text{ba}\}$. Their null-restricted and bootstrap counterparts are denoted as $\hat{\beta}_L^r$ and $\hat{\beta}_{L,g}^*$, respectively, and $\hat{\gamma}_L$, $\hat{\gamma}_L^r$, and $\hat{\gamma}_{L,g}^*$ are similarly defined. In the following, we show that $\hat{\beta}_{tsls}$, $\hat{\beta}_{liml}$, $\hat{\beta}_{full}$, $\hat{\beta}_{ba}$ are asymptotically equivalent, and so be their null-restricted and bootstrap counterparts.

Lemma S.B.1 *Suppose Assumptions 1, 2, and 3(i) hold. Then, for $L \in \{liml, full, ba\}$, we have*

$$\hat{\beta}_L = \hat{\beta}_{tsls} + o_p(n^{-1/2}), \quad \hat{\beta}_{tsls} - \beta_n = O_p(n^{-1/2}),$$
$$\hat{\beta}_L^r = \hat{\beta}_{tsls}^r + o_p(n^{-1/2}), \quad \textit{and} \quad \hat{\beta}_{tsls}^r - \beta_n = O_p(n^{-1/2}).$$

Proof. First, $L \in \{\text{liml}, \text{full}, \text{ba}\}$, we have $\hat{\mu}_L = \hat{\kappa}_L - 1$ and

$$\left(\hat{\beta}_L^\top, \hat{\gamma}_L^\top\right)^\top = \left(\vec{X}^\top P_{\vec{Z}} \vec{X} - \hat{\mu}_L \vec{X}^\top M_{\vec{Z}} \vec{X}\right)^{-1} \left(\vec{X}^\top P_{\vec{Z}} Y - \hat{\mu}_L \vec{X}^\top M_{\vec{Z}} Y\right)$$

*Wang acknowledges the financial support from NTU SUG Grant No.M4082262.SS0 and Singapore Ministry of Education Tier 1 grant RG53/20. Zhang acknowledges the financial support from Singapore Ministry of Education Tier 2 grant under grant MOE2018-T2-2-169 and the Lee Kong Chian fellowship. Any possible errors are our own.

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

‡School of Economics, Singapore Management University. E-mail address: yczhang@smu.edu.sg.

$$= \left(\Upsilon^\top \vec{X}\right)^{-1} \Upsilon^\top Y, \text{ where } \Upsilon = \left[P_{\vec{Z}} X - \hat{\mu}_L M_{\vec{Z}} X : W\right].$$

Then, by applying the Frisch–Waugh–Lovell Theorem we obtain that

$$\hat{\beta}_L = \left(X^\top P_{\widetilde{Z}} X - \hat{\mu}_L X^\top M_{\vec{Z}} X\right)^{-1} \left(X^\top P_{\widetilde{Z}} Y - \hat{\mu}_L X^\top M_{\vec{Z}} Y\right), \text{ where } \tilde{Z} = M_W Z.$$

By construction, $\hat{\beta}_{tsls}$ corresponds to $\hat{\mu}_{tsls} = 0$. For the LIML estimator, we have

$$\hat{\mu}_{liml} = \min_r r^\top \vec{Y}^\top M_W Z (Z^\top M_W Z)^{-1} Z^\top M_W \vec{Y} r / (r^\top \vec{Y}^\top M_{\vec{Z}} \vec{Y} r) \quad \text{and} \quad r = (1, -\beta^\top)^\top,$$

which implies that

$$n\hat{\mu}_{liml} \leq \left(\frac{1}{\sqrt{n}} \varepsilon^\top M_W Z\right) \left(\frac{1}{n} Z^\top M_W Z\right)^{-1} \left(\frac{1}{\sqrt{n}} Z^\top M_W \varepsilon\right) / \left(\frac{1}{n} \varepsilon^\top M_{\vec{Z}} \varepsilon\right). \tag{S.B.1}$$

We note that

$$\frac{1}{\sqrt{n}} Z^\top M_W \varepsilon = \frac{1}{\sqrt{n}} \sum_{j \in J} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} \varepsilon_{i,j} = O_p(1) \quad \text{and}$$

$$\left(\frac{1}{n} Z^\top M_W Z\right)^{-1} = \left(\frac{1}{n} \sum_{j \in J} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} \widetilde{Z}_{i,j}^\top\right)^{-1} = O_p(1). \tag{S.B.2}$$

In addition, let $\hat{\varepsilon}_{i,j}$ be the residual from the full sample projection of $\varepsilon_{i,j}$ on $W_{i,j}$. Then, we have

$$\begin{aligned}
\frac{1}{n} \varepsilon^\top M_{\vec{Z}} \varepsilon &= \frac{1}{n} \varepsilon^\top \varepsilon - \frac{1}{n} \varepsilon^\top \vec{Z} (\vec{Z}^\top \vec{Z})^{-1} \vec{Z}^\top \varepsilon \\
&= \widehat{Q}_{\varepsilon\varepsilon} - \widehat{Q}_{\varepsilon\widetilde{Z}} \widehat{Q}_{\widetilde{Z}\widetilde{Z}}^{-1} \widehat{Q}_{\varepsilon\widetilde{Z}}^\top - \widehat{Q}_{\varepsilon W} \widehat{Q}_{WW}^{-1} \widehat{Q}_{\varepsilon W}^\top \\
&= \widehat{Q}_{\varepsilon\varepsilon} - \widehat{Q}_{\varepsilon W} \widehat{Q}_{WW}^{-1} \widehat{Q}_{\varepsilon W}^\top + o_p(1) \\
&= \widehat{Q}_{\tilde{\varepsilon}\tilde{\varepsilon}} + o_p(1) \\
&= \sum_{j \in J} \xi_j \widehat{Q}_{\tilde{\varepsilon}\tilde{\varepsilon},j} + o_p(1) \\
&\geq \sum_{j \in J} \xi_j \widehat{Q}_{\dot{\varepsilon}\dot{\varepsilon},j} + o_p(1) \\
&= \widehat{Q}_{\dot{\varepsilon}\dot{\varepsilon}} + o_p(1) \geq c \ w.p.a.1,
\end{aligned} \tag{S.B.3}$$

where c is a positive constant, the first inequality is by the definition that $\dot{\varepsilon}_{i,j}$ is the residual from the cluster-level projection of $\varepsilon_{i,j}$ on $W_{i,j}$, and the second inequality is by Assumption 2(iii). Combining (S.B.1)–(S.B.3), we have $\hat{\mu}_{liml} = O_p(n^{-1})$. In addition, we have $\frac{1}{n} X^\top M_{\vec{Z}} X = O_p(1)$, $\frac{1}{n} X^\top M_{\vec{Z}} Y = O_p(1)$, and

$$\left(\frac{1}{n} X^\top P_{\widetilde{Z}} X\right)^{-1} = \left(\widehat{Q}_{X\widetilde{Z}} \widehat{Q}_{\widetilde{Z}\widetilde{Z}}^{-1} \widehat{Q}_{\widetilde{Z}X}\right)^{-1} = O_p(1), \tag{S.B.4}$$

where the last equality holds by Assumptions 2(ii) and 3(i). This means

$$\hat{\beta}_{liml} = \hat{\beta}_{tsls} + o_p(n^{-1/2}).$$

In addition, we note $\hat{\mu}_{full} = \hat{\mu}_{liml} - \frac{C}{n-d_z-d_w} = O_p(n^{-1})$ and $\hat{\mu}_{ba} = O(n^{-1})$, respectively. Therefore, we have the same results for $(\hat{\beta}_{ba}, \hat{\gamma}_{ba})$.

For the second statement in the lemma, we have

$$(\hat{\beta}_{tsls} - \beta_n) = \left(X^\top P_{\widetilde{Z}} X\right)^{-1} X^\top P_{\widetilde{Z}} \varepsilon = \left(\widehat{Q}_{X\widetilde{Z}} \widehat{Q}_{\widetilde{Z}\widetilde{Z}}^{-1} \widehat{Q}_{\widetilde{Z}X}\right)^{-1} \widehat{Q}_{X\widetilde{Z}} \widehat{Q}_{\widetilde{Z}\widetilde{Z}}^{-1} \widehat{Q}_{\widetilde{Z}\varepsilon} = O_p(n^{-1/2}),$$

where the last inequality holds because $\widehat{Q}_{\widetilde{Z}\varepsilon} = \frac{1}{n} \sum_{j \in J} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} \varepsilon_{i,j} = O_p(n^{-1/2})$.

Next, we turn to the third statement in the lemma. We note that, for $L \in \{\text{tsls}, \text{liml}, \text{full}, \text{ba}\}$ and $\lambda_\beta^\top \hat{\beta}_L^r = \lambda_0$,

$$\hat{\beta}_L^r = \hat{\beta}_L - \left(X^\top (P_{\widetilde{Z}} - \hat{\mu}_L M_{\vec{Z}}) X\right)^{-1} \lambda_\beta \left(\lambda_\beta^\top (X^\top (P_{\widetilde{Z}} - \hat{\mu}_L M_{\vec{Z}}) X)^{-1} \lambda_\beta\right)^{-1} (\lambda_\beta^\top \hat{\beta}_L - \lambda_0) = O_p(1).$$

As $\hat{\mu}_L = O_p(n^{-1})$ and $\hat{\beta}_L = \hat{\beta}_{tsls} + o_p(n^{1/2})$ for $L \in \{\text{liml}, \text{full}, \text{ba}\}$, we have

$$\hat{\beta}_L^r = \hat{\beta}_{tsls} - \left(X^\top P_{\widetilde{Z}} X\right)^{-1} \lambda_\beta \left(\lambda_\beta^\top (X^\top P_{\widetilde{Z}} X)^{-1} \lambda_\beta\right)^{-1} (\lambda_\beta^\top \hat{\beta}_{tsls} - \lambda_0) + o_p(n^{-1/2}) = \hat{\beta}_{tsls}^r + o_p(n^{-1/2}).$$

For the last statement in the lemma, we note that

$$\begin{aligned} \hat{\beta}_{tsls}^r - \beta_n = (\hat{\beta}_{tsls} - \beta_n) - \left(X^\top P_{\widetilde{Z}} X\right)^{-1} \lambda_\beta \left(\lambda_\beta^\top (X^\top P_{\widetilde{Z}} X)^{-1} \lambda_\beta\right)^{-1} \lambda_\beta^\top (\hat{\beta}_{tsls} - \beta_n) \\ - \left(X^\top P_{\widetilde{Z}} X\right)^{-1} \lambda_\beta \left(\lambda_\beta^\top (X^\top P_{\widetilde{Z}} X)^{-1} \lambda_\beta\right)^{-1} (\lambda_\beta^\top \beta_n - \lambda_0) = O_p(n^{-1/2}), \end{aligned}$$

where the last equality holds because $\lambda_\beta^\top \beta_n - \lambda_0 = \mu n^{-1/2}$ by construction and $\hat{\beta}_{tsls} - \beta_n = O_p(n^{-1/2})$. ■

Lemma S.B.2 *Suppose Assumptions 1, 2, and 3(i) hold and* $\widehat{Q}_{\widetilde{Z}W,j}(\hat{\gamma}_L^r - \gamma) = o_p(n^{-1/2})$. *Then, for* $L \in \{liml, full, ba\}$ *and* $g \in \boldsymbol{G}$, *we have*

$$\hat{\beta}_{L,g}^* = \hat{\beta}_{tsls,g}^* + o_p(n^{-1/2}) \quad \text{and} \quad \hat{\beta}_{tsls,g}^* - \beta_n = O_p(n^{-1/2}). \tag{S.B.5}$$

Proof. By the same argument in the proof of Lemma S.B.1, for $L \in \{\text{tsls}, \text{liml}, \text{full}, \text{ba}\}$ and $g \in \mathbf{G}$, we have

$$\hat{\beta}_{L,g}^* = \left(X^{*\top}(g) P_{\widetilde{Z}} X^*(g) - \hat{\mu}_{L,g}^* X^{*\top}(g) M_{\vec{Z}} X^*(g)\right)^{-1} \left(X^{*\top}(g) P_{\widetilde{Z}} Y^*(g) - \hat{\mu}_{L,g}^* X^{*\top}(g) M_{\vec{Z}} Y^*(g)\right). \tag{S.B.6}$$

such that $\hat{\mu}_{tsls,g}^* = 0$, $\hat{\mu}_{full,g}^* = \hat{\mu}_{liml,g}^* - \frac{C}{n-d_z-d_x}$, $\hat{\mu}_{ba,g}^* = \hat{\mu}_{ba}$, and

$$\hat{\mu}_{liml,g}^* = \min_r r^\top \vec{Y}^{*\top}(g) M_W Z (Z^\top M_W Z)^{-1} Z^\top M_W \vec{Y}^*(g) r / (r^\top \vec{Y}^{*\top}(g) M_{\vec{Z}} \vec{Y}^*(g) r),$$

where $\vec{Y}^*(g) = [Y^*(g) : X^*(g)]$, $Y^*(g)$ is an $n \times 1$ vector formed by $Y_{i,j}^*(g)$, and $r = (1, -\beta^\top)^\top$.

Following the same argument previously, we have

$$n\hat{\mu}_{liml,g}^* \leq \left(\frac{1}{\sqrt{n}} \varepsilon_g^{*r\top} M_W Z\right) \left(\frac{1}{n} Z^\top M_W Z\right)^{-1} \left(\frac{1}{\sqrt{n}} Z^\top M_W \varepsilon_g^{*r}\right) / \left(\frac{1}{n} \varepsilon_g^{*r\top} M_{\vec{Z}} \varepsilon_g^{*r}\right), \tag{S.B.7}$$

where ε_g^{*r} is an $n \times 1$ vector formed by $g_j \hat{\varepsilon}^r_{i,j}$. We first note that

$$\begin{aligned}
\frac{1}{n}\varepsilon_g^{*r\top} M_W Z &= \frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}} g_j \widetilde{Z}_{i,j}\hat{\varepsilon}^r_{i,j} \\
&= \sum_{j\in J}\xi_j g_j \left[\frac{1}{n_j}\sum_{i\in I_{n,j}}\widetilde{Z}_{i,j}\varepsilon_{i,j} + \frac{1}{n_j}\sum_{i\in I_{n,j}}\widetilde{Z}_{i,j}X_{i,j}^\top(\beta_n - \hat{\beta}^r_L) + \frac{1}{n_j}\sum_{i\in I_{n,j}}\widetilde{Z}_{i,j}W_{i,j}^\top(\gamma - \hat{\gamma}^r_L)\right] \\
&= O_p(n^{-1/2}),
\end{aligned} \tag{S.B.8}$$

where the last line is by Assumptions 1(ii), Lemma S.B.1, and Lemma S.B.3. Second, we have

$$\begin{aligned}
\frac{1}{n}\varepsilon_g^{*r\top} M_{\vec{Z}}\varepsilon_g^{*r} &= \frac{1}{n}\varepsilon_g^{*r\top}\varepsilon_g^{*r} - \frac{1}{n}\varepsilon_g^{*r\top}\vec{Z}(\vec{Z}^\top\vec{Z})^{-1}\vec{Z}^\top\varepsilon_g^{*r} \\
&= \frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}}(g_j\hat{\varepsilon}^r_{i,j})^2 - \left[\frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}}(g_j\hat{\varepsilon}^r_{i,j}\widetilde{Z}_{i,j}^\top)\right]\widehat{Q}^{-1}_{\widetilde{Z}\widetilde{Z}}\left[\frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}}(g_j\hat{\varepsilon}^r_{i,j}\widetilde{Z}_{i,j})\right] \\
&\quad - \left[\frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}}(g_j\hat{\varepsilon}^r_{i,j}W_{i,j}^\top)\right]\widehat{Q}^{-1}_{WW}\left[\frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}}(g_j\hat{\varepsilon}^r_{i,j}W_{i,j})\right] \\
&= \frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}}(\hat{\varepsilon}^r_{i,j})^2 - \left[\frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}}(g_j\hat{\varepsilon}^r_{i,j}W_{i,j}^\top)\right]\widehat{Q}^{-1}_{WW}\left[\frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}}(g_j\hat{\varepsilon}^r_{i,j}W_{i,j})\right] + o_p(1),
\end{aligned} \tag{S.B.9}$$

where the last equality is by $\frac{1}{n_j}\sum_{i\in I_{n,j}}\hat{\varepsilon}^r_{i,j}\widetilde{Z}_{i,j} = O_p(n^{-1/2})$, as in (S.B.8). We further note that

$$\hat{\varepsilon}^r_{i,j} = \varepsilon_{i,j} - X_{i,j}^\top(\hat{\beta}^r_L - \beta_n) - W_{i,j}^\top(\hat{\gamma}^r_L - \gamma),$$

where $\hat{\beta}^r_L - \beta_n = O_p(n^{-1/2})$ and

$$\hat{\gamma}^r_L - \gamma = \widehat{Q}^{-1}_{WW}\left[\frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}} W_{i,j}\left(Y_{i,j} - X_{i,j}^\top\hat{\beta}^r_L\right)\right] - \gamma = \widehat{Q}^{-1}_{WW}\widehat{Q}_{W\varepsilon} + O_p(n^{-1/2}).$$

Therefore, we have

$$\frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}}(\hat{\varepsilon}^r_{i,j})^2 = \frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}}\mathscr{E}^2_{i,j} + o_p(1),$$

where $\mathscr{E}_{i,j} = \varepsilon_{i,j} - W_{i,j}^\top\widehat{Q}^{-1}_{WW}\widehat{Q}_{W\varepsilon}$. Similarly, we can show that

$$\begin{aligned}
&\left[\frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}}(g_j\hat{\varepsilon}^r_{i,j}W_{i,j}^\top)\right]\widehat{Q}^{-1}_{WW}\left[\frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}}(g_j\hat{\varepsilon}^r_{i,j}W_{i,j})\right] \\
&= \left[\frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}}(g_j\mathscr{E}_{i,j}W_{i,j}^\top)\right]\widehat{Q}^{-1}_{WW}\left[\frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}}(g_j\mathscr{E}_{i,j}W_{i,j})\right] + o_p(1),
\end{aligned}$$

which, combined with (S.B.9), implies

$$\begin{aligned}\frac{1}{n}\varepsilon_g^{*r\top} M_{\vec{Z}} \varepsilon_g^{*r} &= \frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}} \mathscr{E}_{i,j}^2 - \left[\frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}} (g_j \mathscr{E}_{i,j} W_{i,j}^\top)\right] \widehat{Q}_{WW}^{-1} \left[\frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}} (g_j \mathscr{E}_{i,j} W_{i,j})\right] + o_p(1) \\ &= \frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}} (\mathscr{E}_{i,j} - g_j W_{i,j}^\top \hat{\theta}_g)^2 + o_p(1),\end{aligned}$$

where $\hat{\theta}_g = \widehat{Q}_{WW}^{-1}\left[\frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}} (g_j \mathscr{E}_{i,j} W_{i,j})\right]$ and the second equality holds because

$$\widehat{Q}_{WW} = \frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}} W_{i,j} W_{i,j}^\top = \frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}} (g_j W_{i,j})(g_j W_{i,j})^\top.$$

Recall that $\dot{\varepsilon}_{i,j}$ is the residual from the cluster-level projection of $\varepsilon_{i,j}$ on $W_{i,j}$, which means there exists a vector $\hat{\theta}_j$ such that $\varepsilon_{i,j} = \dot{\varepsilon}_{i,j} + W_{i,j}^\top \hat{\theta}_j$ and $\sum_{i\in I_{n,j}} \dot{\varepsilon}_{i,j} W_{i,j} = 0$. Then, we have

$$\frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}} (\mathscr{E}_{i,j} - g_j W_{i,j}^\top \hat{\theta}_g)^2 = \frac{1}{n}\sum_{j\in J}\left[\sum_{i\in I_{n,j}} \dot{\varepsilon}_{i,j}^2 + \sum_{i\in I_{n,j}} (W_{i,j}^\top \hat{\theta}_j - g_j W_{i,j}^\top \hat{\theta}_g)^2\right] \geq \frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}} \dot{\varepsilon}_{i,j}^2 \geq c\ w.p.a.1,$$

where the last inequality is by Assumption 2(iii). This implies

$$\frac{1}{n}\varepsilon_g^{*r\top} M_{\vec{Z}} \varepsilon_g^{*r} \geq c - o_p(1). \tag{S.B.10}$$

Combining (S.B.7), (S.B.8), and (S.B.10), we obtain that $\hat{\mu}_{liml,g}^* = O_p(n^{-1})$, and thus, $\hat{\mu}_{full,g}^* = O_p(n^{-1})$. It is also obvious that $\hat{\mu}_{ba,g}^* = O_p(n^{-1})$. Given that $\hat{\mu}_{L,g}^* = O_p(n^{-1})$, to establish $\hat{\beta}_{L,g}^* = \hat{\beta}_{tsls,g}^* + o_p(n^{-1/2})$, it suffices to show $\frac{1}{n}X^{*\top}(g) M_{\vec{Z}} X^*(g) = O_p(1)$, $\frac{1}{n}X^{*\top}(g) M_{\vec{Z}} Y^*(g) = O_p(1)$, and $(\frac{1}{n}X^{*\top}(g) P_{\vec{Z}} X^*(g))^{-1} = O_p(1)$.

For $\frac{1}{n}X^{*\top}(g) M_{\vec{Z}} Y^*(g) = \frac{1}{n}X^{*\top}(g) Y^*(g) - \frac{1}{n}X^{*\top}(g) P_{\vec{Z}} Y^*(g)$, we first show

$$\begin{aligned}\frac{1}{n}X^{*\top}(g) Y^*(g) &= \frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}} X_{i,j}^*(g)\left(X_{i,j}^{*\top}(g)\hat{\beta}_n^r + W_{i,j}\hat{\gamma}_n^r + g_j \hat{\varepsilon}_{i,j}^r\right) \\ &= \widehat{Q}_{XX}^*(g)\hat{\beta}_n^r + \widehat{Q}_{XW}^*(g)\hat{\gamma}_n^r + \widehat{Q}_{X\varepsilon}^*(g) = O_p(1),\end{aligned} \tag{S.B.11}$$

where $\widehat{Q}_{XX}^*(g) = \frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}} X_{i,j}^*(g) X_{i,j}^{*\top}(g)$, $\widehat{Q}_{XW}^*(g) = \frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}} X_{i,j}^*(g) W_{i,j}^\top$, and $\widehat{Q}_{X\varepsilon}^*(g) = \frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}} g_j X_{i,j}^*(g)\hat{\varepsilon}_{i,j}^r$. Notice that

$$\widehat{Q}_{XX}^*(g) = \sum_{j\in J}\frac{n_j}{n}\frac{1}{n_j}\sum_{i\in I_{n,j}} \left(X_{i,j} + (g_j - 1)\tilde{v}_{i,j}\right)\left(X_{i,j} + (g_j - 1)\tilde{v}_{i,j}\right)^\top = O_p(1) \tag{S.B.12}$$

because $\widehat{Q}_{XX,j} = O_p(1)$, $\widehat{Q}_{X\tilde{v},j} = O_p(1)$, and $\widehat{Q}_{\tilde{v}\tilde{v},j} = O_p(1)$. To see these three relations, we note that

$$\widehat{Q}_{X\tilde{v},j} = \widehat{Q}_{XX,j} - \widehat{Q}_{X\overline{Z},j}\widetilde{\Pi}_{\overline{Z}} - \widehat{Q}_{XW,j}\widetilde{\Pi}_w = O_p(1),$$

by $\widehat{Q}_{XX,j} = O_p(1)$, $\widehat{Q}_{X\overline{Z},j} = O_p(1)$, $\widehat{Q}_{XW,j} = O_p(1)$, $\tilde{\Pi}_{\overline{Z}} = O_p(1)$, and $\widetilde{\Pi}_w = O_p(1)$, where $\widehat{Q}_{X\overline{Z},j} = \frac{1}{n_j}\sum_{i\in I_{n,j}} X_{i,j}\overline{Z}_{i,j}^\top$, and $\widehat{Q}_{XW,j} = \frac{1}{n_j}\sum_{i\in I_{n,j}} X_{i,j} W_{i,j}^\top$. Similar arguments hold for $\widehat{Q}_{\tilde{v}\tilde{v},j}$. We can also show that

$$\widehat{Q}^*_{XW}(g) = \sum_{j\in J} \frac{n_j}{n} \frac{1}{n_j} \sum_{i\in I_{n,j}} \left(X_{i,j} + (g_j - 1)\tilde{v}_{i,j}\right) W_{i,j}^\top = O_p(1), \tag{S.B.13}$$

by $\widehat{Q}_{XW,j} = O_p(1)$ and $\widehat{Q}_{\tilde{v}W,j} = O_p(1)$, where $\widehat{Q}_{XW,j} = \frac{1}{n_j}\sum_{i\in I_{n,j}} X_{i,j}W_{i,j}^\top$ and $\widehat{Q}_{\tilde{v}W,j} = \frac{1}{n_j}\sum_{i\in I_{n,j}} \tilde{v}_{i,j}W_{i,j}^\top$. In addition, we have

$$\widehat{Q}^*_{X\varepsilon}(g) = \sum_{j\in J} \frac{n_j}{n} \frac{1}{n_j} \sum_{i\in I_{n,j}} g_j \left(X_{i,j} + (g_j - 1)\tilde{v}_{i,j}\right) \hat{\varepsilon}^r_{i,j} = O_p(1), \tag{S.B.14}$$

by $\widehat{Q}_{X\hat{\varepsilon},j} = O_p(1)$ and $\widehat{Q}_{\tilde{v}\hat{\varepsilon},j} = O_p(1)$ under similar arguments as those in (S.B.8), where

$$\widehat{Q}_{X\hat{\varepsilon},j} = \frac{1}{n_j}\sum_{i\in I_{n,j}} X_{i,j}\hat{\varepsilon}^r_{i,j} \quad \text{and} \quad \widehat{Q}_{\tilde{v}\hat{\varepsilon},j} = \frac{1}{n_j}\sum_{i\in I_{n,j}} \tilde{v}_{i,j}\hat{\varepsilon}^r_{i,j}.$$

Combining (S.B.12), (S.B.13), (S.B.14), $\hat{\beta}^r_L = O_p(1)$, and $\hat{\gamma}^r_L = \widehat{Q}^{-1}_{WW}\widehat{Q}_{w\varepsilon} + o_p(1) = O_p(1)$, we obtain (S.B.11). Next, by the fact that $\frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}} \widetilde{Z}_{i,j}W_{i,j}^\top = 0$, we have

$$\frac{1}{n}X^{*\top}(g)P_{\vec{Z}}Y^*(g) = \begin{pmatrix} \widehat{Q}^{*\top}_{\widetilde{Z}X}(g) & \widehat{Q}^*_{XW}(g) \end{pmatrix} \begin{pmatrix} \widehat{Q}^{-1}_{\widetilde{Z}\widetilde{Z}} & 0 \\ 0 & \widehat{Q}^{-1}_{WW} \end{pmatrix} \begin{pmatrix} \widehat{Q}^*_{\widetilde{Z}Y}(g) \\ \widehat{Q}^*_{WY}(g) \end{pmatrix},$$

where

$$\widehat{Q}^*_{\widetilde{Z}X}(g) = \frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}} \widetilde{Z}_{i,j}X^{*\top}_{i,j}(g), \quad \widehat{Q}^*_{\widetilde{Z}Y}(g) = \frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}} \widetilde{Z}_{i,j}Y^*_{i,j}(g),$$
$$\text{and} \quad \widehat{Q}^*_{WY}(g) = \frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}} W_{i,j}Y^*_{i,j}(g).$$

Following the same lines of reasoning, we can show that, for all $g \in \mathbf{G}$, $\widehat{Q}^*_{\widetilde{Z}X}(g) = O_p(1)$, $\widehat{Q}^*_{\widetilde{Z}Y}(g) = O_p(1)$, and $\widehat{Q}^*_{WY}(g) = O_p(1)$. In addition, by Assumptions 1(iv) and 2(iii), $\widehat{Q}^{-1}_{\widetilde{Z}\widetilde{Z}} = O_p(1)$ and $\widehat{Q}^{-1}_{WW} = O_p(1)$, which further implies that $\frac{1}{n}X^{*\top}(g)M_{\vec{Z}}Y^*(g) = O_p(1)$.

By a similar argument, we can show $\frac{1}{n}X^{*\top}(g)M_{\vec{Z}}X^*(g) = O_p(1)$. Last, we have

$$\frac{1}{n}X^{*\top}(g)P_{\widetilde{Z}}X^*(g) = \widehat{Q}^{*\top}_{\widetilde{Z}X}(g)\widehat{Q}^{-1}_{\widetilde{Z}\widetilde{Z}}\widehat{Q}^*_{\widetilde{Z}X}(g) = \widehat{Q}^\top_{\widetilde{Z}X}\widehat{Q}^{-1}_{\widetilde{Z}\widetilde{Z}}\widehat{Q}_{\widetilde{Z}X} + o_p(1) \xrightarrow{p} Q^\top_{\widetilde{Z}X}Q^{-1}_{\widetilde{Z}\widetilde{Z}}Q_{\widetilde{Z}X},$$

where we use the fact that $\widehat{Q}^*_{\widetilde{Z}X}(g) = \widehat{Q}_{\widetilde{Z}X} + o_p(1)$, which is established in Step 1 in the proof of Theorem 2.1. In addition, $Q^\top_{\widetilde{Z}X}Q^{-1}_{\widetilde{Z}\widetilde{Z}}Q_{\widetilde{Z}X}$ is invertible by Assumptions 2(ii) and 3(i). This implies $(\frac{1}{n}X^{*\top}(g)P_{\widetilde{Z}}X^*(g))^{-1} = O_p(1)$, which further implies $\hat{\beta}^*_{L,g} = \hat{\beta}^*_{tsls,g} + o_p(n^{-1/2})$.

For the second result, we note that

$$\begin{aligned}\hat{\beta}^*_{tsls,g} - \beta_n &= \left[\widehat{Q}^{*\top}_{\widetilde{Z}X}(g)\widehat{Q}^{-1}_{\widetilde{Z}\widetilde{Z}}\widehat{Q}^*_{\widetilde{Z}X}(g)\right]^{-1}\widehat{Q}^{*\top}_{\widetilde{Z}X}(g)\widehat{Q}^{-1}_{\widetilde{Z}\widetilde{Z}}\left[\frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}}\widetilde{Z}_{i,j}y^*_{i,j}(g)\right] - \beta_n \\ &= \hat{\beta}^r_{tsls} - \beta_n + \left[\widehat{Q}^{*\top}_{\widetilde{Z}X}(g)\widehat{Q}^{-1}_{\widetilde{Z}\widetilde{Z}}\widehat{Q}^*_{\widetilde{Z}X}(g)\right]^{-1}\widehat{Q}^{*\top}_{\widetilde{Z}X}(g)\widehat{Q}^{-1}_{\widetilde{Z}\widetilde{Z}}\left[\frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}}g_j\widetilde{Z}_{i,j}\hat{\varepsilon}^r_{i,j}\right],\end{aligned}$$

where the second equality holds because $\sum_{j \in J} \sum_{i \in I_{n,j}} g_j \widetilde{Z}_{i,j} W_{i,j}^\top = 0$. In addition, note that

$$\frac{1}{n_j} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} \hat{\varepsilon}_{i,j}^r = \frac{1}{n_j} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} \varepsilon_{i,j} - \widehat{Q}_{\widetilde{Z}X,j}(\hat{\beta}_{tsls}^r - \beta_n) - \widehat{Q}_{\widetilde{Z}W,j}(\hat{\gamma}_{tsls}^r - \gamma) = O_p(n^{-1/2}).$$

Combining this with the fact that $\hat{\beta}_{tsls}^r - \beta_n = O_p(n^{-1/2})$ as shown in Lemma S.B.1, we have $\hat{\beta}_{tsls,g}^* - \beta_n = O_p(n^{-1/2})$. ■

Lemma S.B.3 *Suppose Assumptions 1 and 2 hold. Then, we have, for $j \in J$,*

$$\widehat{Q}_{\widetilde{Z}W,j}(\bar{\gamma}^r - \gamma) = o_p(n^{-1/2}).$$

If, in addition, Assumption 3(i) holds, then we have, For $j \in J$, $L \in \{tsls, liml, full, ba\}$, and $g \in \boldsymbol{G}$,

$$\widehat{Q}_{\widetilde{Z}W,j}(\hat{\gamma}_L - \gamma) = o_p(n^{-1/2}), \quad \widehat{Q}_{\widetilde{Z}W,j}(\hat{\gamma}_L^r - \gamma) = o_p(n^{-1/2}), \quad \text{and} \quad \widehat{Q}_{\widetilde{Z}W,j}(\hat{\gamma}_{L,g}^* - \gamma) = o_p(n^{-1/2}).$$

Proof. When $\widehat{Q}_{\widetilde{Z}W,j} = 0$, all the results hold trivially. We now assume $\widehat{Q}_{\widetilde{Z}W,j} = o_p(1)$ and $\frac{1}{\sqrt{n}} \sum_{j \in J} \sum_{i \in I_{n,j}} W_{i,j} \varepsilon_{i,j} = O_p(1)$. The first statement holds because

$$\begin{aligned} \bar{\gamma}^r - \gamma &= \widehat{Q}_{WW}^{-1} \left[\frac{1}{n} \sum_{j \in J} \sum_{i \in I_{n,j}} W_{i,j}(Y_{i,j} - X_{i,j}^\top \beta_0) \right] - \gamma \\ &= \widehat{Q}_{WW}^{-1} \left[\frac{1}{n} \sum_{j \in J} \sum_{i \in I_{n,j}} W_{i,j} \varepsilon_{i,j} \right] - \widehat{Q}_{WW}^{-1} \widehat{Q}_{WX}(\beta_n - \beta_0) = O_p(n^{-1/2}), \end{aligned} \tag{S.B.15}$$

Next, if Assumption 3(i) also holds, then

$$\begin{aligned} \hat{\gamma}_L - \gamma &= \widehat{Q}_{WW}^{-1} \left[\frac{1}{n} \sum_{j \in J} \sum_{i \in I_{n,j}} W_{i,j}(Y_{i,j} - X_{i,j}^\top \hat{\beta}_L) \right] - \gamma \\ &= \widehat{Q}_{WW}^{-1} \left[\frac{1}{n} \sum_{j \in J} \sum_{i \in I_{n,j}} W_{i,j} \varepsilon_{i,j} \right] - \widehat{Q}_{WW}^{-1} \widehat{Q}_{WX}(\hat{\beta}_L - \beta_n) = O_p(n^{-1/2}), \end{aligned} \tag{S.B.16}$$

where the last equality holds by Lemma S.B.1. This implies $\widehat{Q}_{\widetilde{Z}W,j}(\hat{\gamma}_L - \gamma) = o_p(n^{-1/2})$. In the same manner, we can show $\widehat{Q}_{\widetilde{Z}W,j}(\hat{\gamma}_L^r - \gamma) = o_p(n^{-1/2})$. Last, given Assumptions 1, 2, 3(i), and the fact that $\widehat{Q}_{\widetilde{Z}W,j}(\hat{\gamma}_L^r - \gamma) = o_p(n^{-1/2})$, Lemma S.B.2 shows $\hat{\beta}_{L,g}^* - \beta_n = (\hat{\beta}_{L,g}^* - \hat{\beta}_{tsls,g}^*) + (\hat{\beta}_{tsls,g}^* - \beta_n) = O_p(n^{-1/2})$. Then, following the same argument in (S.B.16), we can show $\hat{\gamma}_{L,g}^* - \gamma = O_p(n^{-1/2})$, which leads to the desired result. ■

S.C Proof of Theorem 2.2

For the power of the T_n-based wild bootstrap test, we focus on the TSLS estimator. As shown in the end of the proof of Theorem 2.1, the test statistics constructed using LIML, FULL, and BA estimators and their bootstrap counterparts are asymptotically equivalent to those constructed based on TSLS estimator, which leads to the desired result. Note that

$$
\begin{aligned}
||\sqrt{n}(\lambda_\beta^\top \hat{\beta}_{tsls} - \lambda_0)||_{\hat{A}_r} &= ||\sqrt{n}\lambda_\beta^\top(\hat{\beta}_{tsls} - \beta_n) + \sqrt{n}\lambda_\beta^\top(\beta_n - \hat{\beta}^r_{tsls})||_{\hat{A}_r} \\
&= \left\| \lambda_\beta^\top \widehat{Q}^{-1} \widehat{Q}^\top_{\widetilde{Z}X} \widehat{Q}^{-1}_{\widetilde{Z}\widetilde{Z}} \sum_{j\in J} \sum_{i \in I_{n,j}} \frac{\widetilde{Z}_{i,j}\varepsilon_{i,j}}{\sqrt{n}} + \sqrt{n}\lambda_\beta^\top(\beta_n - \hat{\beta}^r_{tsls}) \right\|_{\hat{A}_r} \\
&= \left\| \lambda_\beta^\top \widehat{Q}^{-1} \widehat{Q}^\top_{\widetilde{Z}X} \widehat{Q}^{-1}_{\widetilde{Z}\widetilde{Z}} \sum_{j\in J} \sum_{i \in I_{n,j}} \left(\frac{\widetilde{Z}_{i,j}\varepsilon_{i,j}}{\sqrt{n}} + \frac{\widetilde{Z}_{i,j}X_{i,j}^\top}{n} \sqrt{n}(\beta_n - \hat{\beta}^r_{tsls}) \right) \right\|_{\hat{A}_r}.
\end{aligned}
$$

Notice that Assumptions 1, 2, 3(i), and Lemma S.B.1 imply that $\sqrt{n}(\hat{\beta}^r_{tsls} - \beta_n)$ is bounded in probability. This implies

$$
T_n = ||\sqrt{n}(\lambda_\beta^\top \hat{\beta}_{tsls} - \lambda_0)||_{\hat{A}_r} \xrightarrow{d} \left\| \sum_{j\in J} \left[\sqrt{\xi_j} \lambda_\beta^\top Q^{-1} Q^\top_{\widetilde{Z}X} Q^{-1}_{\widetilde{Z}\widetilde{Z}} \mathcal{Z}_j \right] + \mu \right\|_{A_r}. \tag{S.C.1}
$$

Recall $\widehat{Q}$ and $\widehat{Q}^*_g$ defined in (8) and (9), respectively. We have $\widehat{Q}^{*-1}_g = \widehat{Q}^{-1} + o_p(1)$ and

$$
\begin{aligned}
&||\sqrt{n}\lambda_\beta^\top(\hat{\beta}^*_{tsls,g} - \hat{\beta}^r_{tsls})||_{\hat{A}_r} \\
&= \left\| \lambda_\beta^\top \widehat{Q}^{*-1}_g \widehat{Q}^{*\top}_{\widetilde{Z}X}(g) \widehat{Q}^{-1}_{\widetilde{Z}\widetilde{Z}} \sum_{j\in J} \sum_{i \in I_{n,j}} g_j \left(\frac{\widetilde{Z}_{i,j}\varepsilon_{i,j}}{\sqrt{n}} + \frac{\widetilde{Z}_{i,j}X_{i,j}^\top}{n} \sqrt{n}(\beta_n - \hat{\beta}^r_{tsls}) + \frac{\widetilde{Z}_{i,j}W_{i,j}^\top}{n} \sqrt{n}(\gamma - \hat{\gamma}^r_{tsls}) \right) \right\|_{\hat{A}_r} \\
&= \left\| \lambda_\beta^\top \widehat{Q}^{-1} \widehat{Q}^\top_{\widetilde{Z}X} \widehat{Q}^{-1}_{\widetilde{Z}\widetilde{Z}} \sum_{j\in J} \sum_{i \in I_{n,j}} g_j \left(\frac{\widetilde{Z}_{i,j}\varepsilon_{i,j}}{\sqrt{n}} + \frac{\widetilde{Z}_{i,j}X_{i,j}^\top}{n} \sqrt{n}(\beta_n - \hat{\beta}^r_{tsls}) \right) \right\|_{\hat{A}_r} + o_p(1),
\end{aligned}
\tag{S.C.2}
$$

where the last equality follows from Lemma S.B.3 and $\widehat{Q}^*_{\widetilde{Z}X}(g) = \widehat{Q}_{\widetilde{Z}X} + o_p(1)$. Furthermore, we notice that

$$
\begin{aligned}
\hat{\beta}^r_{tsls} &= \hat{\beta}_{tsls} - \widehat{Q}^{-1}\lambda_\beta \left(\lambda_\beta^\top \widehat{Q}^{-1}\lambda_\beta\right)^{-1} \left(\lambda_\beta^\top \hat{\beta}_{tsls} - \lambda_0\right) \\
&= \hat{\beta}_{tsls} - \widehat{Q}^{-1}\lambda_\beta \left\{ \left(\lambda_\beta^\top \widehat{Q}^{-1}\lambda_\beta\right)^{-1} \lambda_\beta^\top(\hat{\beta}_{tsls} - \beta_n) + \left(\lambda_\beta^\top \widehat{Q}^{-1}\lambda_\beta\right)^{-1} (\lambda_\beta^\top \beta_n - \lambda_0) \right\}.
\end{aligned}
\tag{S.C.3}
$$

Therefore, employing (S.C.3) with $\sqrt{n}(\lambda_\beta^\top \beta_n - \lambda_0) = \lambda_\beta^\top \mu_\beta$, we conclude that whenever $E_n = 1$,

$$
\begin{aligned}
\sum_{i \in I_{n,j}} \frac{\widetilde{Z}_{i,j}X_{i,j}^\top}{n} \sqrt{n}(\beta_n - \hat{\beta}^r_{tsls}) = \sum_{i \in I_{n,j}} \frac{\widetilde{Z}_{i,j}X_{i,j}^\top}{n} &\left\{ \left(I_{d_x} - \widehat{Q}^{-1}\lambda_\beta \left(\lambda_\beta^\top \widehat{Q}^{-1}\lambda_\beta\right)^{-1} \lambda_\beta^\top \right) \sqrt{n}(\beta_n - \hat{\beta}_{tsls}) \right. \\
&\left. + \widehat{Q}^{-1}\lambda_\beta \left(\lambda_\beta^\top \widehat{Q}^{-1}\lambda_\beta\right)^{-1} \lambda_\beta^\top \mu_\beta \right\}.
\end{aligned}
$$

Together with $(S.C.2)$, this implies that

$$
\begin{aligned}
&\limsup_{n\to\infty} \mathbb{P} \left(\begin{array}{c} \left\| \sqrt{n}\lambda_\beta^\top(\hat{\beta}^*_{tsls,g} - \hat{\beta}^r_{tsls}) - \lambda_\beta^\top \widehat{Q}^{-1} \widehat{Q}^\top_{\widetilde{Z}X} \widehat{Q}^{-1}_{\widetilde{Z}\widetilde{Z}} \right. \\ \left. \times \sum_{j\in J} \sum_{i \in I_{n,j}} g_j \left(\frac{\widetilde{Z}_{i,j}\varepsilon_{i,j}}{\sqrt{n}} + \xi_j \widehat{Q}_{\widetilde{Z}X,j} \widehat{Q}^{-1}\lambda_\beta \left(\lambda_\beta^\top \widehat{Q}^{-1}\lambda_\beta\right)^{-1} \lambda_\beta^\top \mu_\beta \right) \right\|_{\hat{A}_r} > \varepsilon; E_n = 1 \end{array} \right) \\
&= \limsup_{n\to\infty} \mathbb{P} \left(\begin{array}{c} \left\| \sqrt{n}\lambda_\beta^\top(\hat{\beta}^*_{tsls,g} - \hat{\beta}^r_{tsls}) - \lambda_\beta^\top Q^{-1} Q^\top_{\widetilde{Z}X} Q^{-1}_{\widetilde{Z}\widetilde{Z}} \right. \\ \left. \times \sum_{j\in J} g_j \left[\sqrt{\xi_j} \mathcal{Z}_j + \xi_j Q_{\widetilde{Z}X,j} Q^{-1}\lambda_\beta \left(\lambda_\beta^\top Q^{-1}\lambda_\beta\right)^{-1} \mu \right] \right\|_{\hat{A}_r} > \varepsilon; E_n = 1 \end{array} \right)
\end{aligned}
$$

$$= \limsup_{n\to\infty} \mathbb{P}\Big(\quad \Big\| \sqrt{n}\lambda_\beta^\top(\hat{\beta}^*_{tsls,g} - \hat{\beta}^r_{tsls}) - \textstyle\sum_{j\in J} g_j \left[\sqrt{\xi_j}\lambda_\beta^\top Q^{-1} Q^\top_{\widetilde{Z}X} Q^{-1}_{\widetilde{Z}\widetilde{Z}} \mathcal{Z}_j + \xi_j a_j \mu\right] \Big\|_{\hat{A}_r} > \varepsilon; E_n = 1\Big) = 0.$$

This implies

$$T^*_n(g) = \left\| \sqrt{n}\lambda_\beta^\top(\hat{\beta}^*_{tsls,g} - \hat{\beta}^r_{tsls})\right\|_{\hat{A}_r} \xrightarrow{d} \Big\| \sum_{j\in J} g_j \left[\sqrt{\xi_j}\lambda_\beta^\top Q^{-1} Q^\top_{\widetilde{Z}X} Q^{-1}_{\widetilde{Z}\widetilde{Z}} \mathcal{Z}_j + \xi_j a_j \mu\right] \Big\|_{A_r}. \quad \text{(S.C.4)}$$

In addition, let $\mathbf{G}_s = \mathbf{G}\backslash\mathbf{G}_w$, where $\mathbf{G}_w = \{g \in \mathbf{G} : g_j = g_{j'}, \forall j, j' \in J_s\}$. We note that $|\mathbf{G}_s| = |\mathbf{G}| - 2^{q-q_s+1} \geq k^*$. Therefore, based on (S.C.1) and (S.C.4), to establish the desired result, it suffices to show that as $||\mu||_2 \to \infty$,

$$\liminf_{n\to\infty} \mathbb{P}\{T_n > \max_{g\in\mathbf{G}_s} T^*_n(g)\} \to 1,$$

which follows under similar arguments as those employed in the proof of Theorem 3.2 in Canay, Santos, and Shaikh (2021). ■

S.D Proof of Theorem 2.3

Following the same argument in the proof of Theorem 2.4, we can show that

$$(T_{CR,n}, \{T^*_{CR,n}(g)\}_{g\in\mathbf{G}}) \xrightarrow{d} (T_{CR,\infty}(\iota), \{T_{CR,\infty}(g)\}_{g\in\mathbf{G}})$$

where $T_{CR,n}(g)$ is defined in the proof of Theorem 2.4 with $\mu = 0$ as we are under the null. Then, the rest of the proof is similar to Step 3 in the proof of Theorem 2.1. We omit the detail for brevity. ■

S.E Proof of Theorem 2.5

The proof for the AR_n-based wild bootstrap test follows similar arguments as those in Theorem 2.1, and thus we keep exposition more concise. Let $\mathbb{S} \equiv \otimes_{j\in J}\mathbf{R}^{d_z} \times \mathbf{R}^{d_z\times d_z}$ and write an element $s \in \mathbb{S}$ by $s = (\{s_{1j} : j \in J\}, s_2)$ where $s_{1j} \in \mathbf{R}^{d_z}$ for any $j \in J$. Define the function T_{AR}: $\mathbb{S} \to \mathbf{R}$ to be given by

$$T_{AR}(s) = \Big\| \sum_{j\in J} s_{1j} \Big\|_{s_2}. \quad \text{(S.E.1)}$$

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

$$S_n = \left\{ \frac{\sqrt{n_j}}{\sqrt{n}} \frac{1}{\sqrt{n_j}} \sum_{i\in I_{n,j}} \widetilde{Z}_{i,j}\varepsilon_{i,j} : j \in J, \hat{A}_z \right\}, \quad \widehat{S}_n = \left\{ \frac{\sqrt{n_j}}{\sqrt{n}} \frac{1}{\sqrt{n_j}} \sum_{i\in I_{n,j}} \widetilde{Z}_{i,j}\bar{\varepsilon}^r_{i,j} : j \in J, \hat{A}_z \right\},$$

where $\bar{\varepsilon}^r_{i,j} = y_{i,j} - X^\top_{i,j}\beta_0 - W^\top_{i,j}\bar{\gamma}^r$. Note that by the Frisch-Waugh-Lovell theorem,

$$AR_n = \Big\| \sum_{j\in J} \frac{1}{\sqrt{n}} \sum_{i\in I_{n,j}} \widetilde{Z}_{i,j}\varepsilon_{i,j} \Big\|_{\hat{A}_z} = T_{AR}(S_n). \quad \text{(S.E.2)}$$

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

$$AR_n^*(g) = \left\| \sum_{j \in J} \frac{1}{\sqrt{n}} \sum_{i \in I_{n,j}} g_j \widetilde{Z}_{i,j} \bar{\varepsilon}_{i,j}^r \right\|_{\hat{A}_z} = T_{AR}(g\widehat{S}_n). \tag{S.E.3}$$

Therefore, letting $k^* \equiv \lceil |\mathbf{G}|(1-\alpha) \rceil$, we obtain from (S.E.2)-(S.E.3) that

$$1\{AR_n > \hat{c}_{AR,n}(1-\alpha)\} = 1\left\{T_{AR}(S_n) > T_{AR}^{(k^*)}(\widehat{S}_n|\mathbf{G})\right\}.$$

Furthermore, note that we have

$$\begin{aligned} T_{AR}\left(-\iota\widehat{S}_n\right) = T_{AR}\left(\iota\widehat{S}_n\right) &= \left\| \sum_{j \in J} \frac{1}{\sqrt{n}} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} \left(y_{i,j} - X_{i,j}^\top \beta_0 - W_{i,j}^\top \bar{\gamma}^r\right) \right\|_{\hat{A}_z} \\ &= \left\| \sum_{j \in J} \frac{1}{\sqrt{n}} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} \left(\varepsilon_{i,j} - W_{i,j}^\top(\bar{\gamma}^r - \gamma)\right) \right\|_{\hat{A}_z} = T_{AR}\left(S_n\right), \end{aligned} \tag{S.E.4}$$

where the third equality follows from $\sum_{j \in J} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} W_{i,j}^\top = 0$. (S.E.4) implies that if $k^* > |\mathbf{G}| - 2$, then $1\{T_{AR}(S_n) > T_{AR}^{(k^*)}(\widehat{S}_n|\mathbf{G})\} = 0$, and this gives the upper bound in Theorem 2.5. We therefore assume that $k^* \leq |\mathbf{G}| - 2$, in which case

$$\begin{aligned} \limsup_{n \to \infty} \mathbb{P}\{T_{AR}(S_n) > T_{AR}^{(k^*)}(\widehat{S}_n|\mathbf{G})\} &= \limsup_{n \to \infty} \mathbb{P}\{T_{AR}(S_n) > T_{AR}^{(k^*)}(\widehat{S}_n|\mathbf{G} \setminus \{\pm\iota_q\})\} \\ &\leq \limsup_{n \to \infty} \mathbb{P}\{T_{AR}(S_n) \geq T_{AR}^{(k^*)}(\widehat{S}_n|\mathbf{G} \setminus \{\pm\iota_q\})\}. \end{aligned} \tag{S.E.5}$$

Then, to examine the right hand side of (S.E.5), first note that by Assumptions 1 and 5, and the continuous mapping theorem we have

$$\left\{ \frac{\sqrt{n_j}}{\sqrt{n}} \frac{1}{\sqrt{n_j}} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} \varepsilon_{i,j} : j \in J, \hat{A}_z \right\} \xrightarrow{d} \left\{ \sqrt{\xi_j} \mathcal{Z}_j : j \in J, A_z \right\} \equiv S, \tag{S.E.6}$$

where $\xi_j > 0$ for all $j \in J$. Furthermore, by Assumptions 1(i), Lemma S.B.3, and $\beta_n = \beta_0$, we have

$$\frac{1}{\sqrt{n}} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} \bar{\varepsilon}_{i,j}^r = \frac{1}{\sqrt{n}} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} \varepsilon_{i,j} - \frac{1}{n} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} W_{i,j}^\top \sqrt{n}(\bar{\gamma}^r - \gamma) = \frac{1}{\sqrt{n}} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} \varepsilon_{i,j} + o_p(1),$$

and thus, for every $g \in \mathbf{G}$,

$$T_{AR}(g\widehat{S}_n) = T_{AR}(gS_n) + o_p(1). \tag{S.E.7}$$

We thus obtain from results in (S.E.6)-(S.E.7) and the continuous mapping theorem that

$$\left(T_{AR}(S_n), \left\{T_{AR}(g\widehat{S}_n) : g \in \mathbf{G}\right\}\right) \xrightarrow{d} \left(T_{AR}(S), \{T_{AR}(gS) : g \in \mathbf{G}\}\right).$$

Then, by the Portmanteau's theorem and the properties of randomization tests, we have

$$\limsup_{n \to \infty} \mathbb{P}\left\{AR_n > \hat{c}_{AR,n}(1-\alpha)\right\} \leq \mathbb{P}\left\{T_{AR}(S) \geq T_{AR}^{k^*}(\mathrm{G} \setminus \{\pm\iota_q\})\right\} = \mathbb{P}\left\{T_{AR}(S) > T_{AR}^{k^*}(\mathrm{G})\right\} \leq \alpha.$$

The lower bounds follow by applying similar arguments as those for Theorem 2.1.

For the $AR_{CR,n}$-based wild bootstrap test, define the statistics $S_{CR,n}, \widehat{S}_{CR,n} \in \mathbb{S}$ as

$$S_{CR,n} = \left\{ \frac{\sqrt{n_j}}{\sqrt{n}} \frac{1}{\sqrt{n_j}} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} \varepsilon_{i,j} : j \in J, \hat{A}_{CR} \right\}, \widehat{S}_{CR,n} = \left\{ \frac{\sqrt{n_j}}{\sqrt{n}} \frac{1}{\sqrt{n_j}} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} \bar{\varepsilon}^r_{i,j} : j \in J, \hat{A}_{CR} \right\}.$$

Notice that different from AR_n and $AR^*_n(g)$, we cannot establish that $T_{AR}(\iota \widehat{S}_{CR,n}) = T_{AR}(S_{CR,n})$ for $T_{AR}(s)$ defined in (S.E.1), as $n^{-1} \sum_{j \in J} \sum_{i \in I_{n,j}} \sum_{k \in I_{n,j}} \widetilde{Z}_{i,j} \widetilde{Z}^\top_{k,j} \bar{\varepsilon}^r_{i,j} \bar{\varepsilon}^r_{k,j}$ may be different from

$$n^{-1} \sum_{j \in J} \sum_{i \in I_{n,j}} \sum_{k \in I_{n,j}} \widetilde{Z}_{i,j} \widetilde{Z}^\top_{k,j} \varepsilon_{i,j} \varepsilon_{k,j}.$$

We set $E_n \in \mathbf{R}$ to equal $E_n \equiv 1 \left\{ n^{-1} \sum_{j \in J} \sum_{i \in I_{n,j}} \sum_{k \in I_{n,j}} \widetilde{Z}_{i,j} \widetilde{Z}^\top_{k,j} \bar{\varepsilon}^r_{i,j} \bar{\varepsilon}^r_{k,j} \text{ is invertible} \right\}$, and have

$$\liminf_{n \to \infty} \mathbb{P}\{E_n = 1\} = 1. \tag{S.E.8}$$

In addition, similar to the case with AR_n and $AR^*_n(g)$, we have under $\beta_n = \beta_0$,

$$T_{AR}\left(g \widehat{S}_{CR,n}\right) = T_{AR}\left(g S_{CR,n}\right) + o_p(1) \text{ for every } g \in \mathbf{G},$$

$$S_{CR,n} \xrightarrow{d} \left\{ \sqrt{\xi_j} \mathcal{Z}_j : j \in J, A_{CR} \right\} \equiv S_{CR},$$

where $A_{CR} = \sum_{j \in J} \xi_j \mathcal{Z}_j \mathcal{Z}_j^\top$, and

$$\left(T_{AR}(S_{CR,n}), \{T_{AR}(g \widehat{S}_{CR,n}) : g \in \mathbf{G}\} \right) \xrightarrow{d} \left(T_{AR}(S_{CR}), \{T_{AR}(g S_{CR}) : g \in \mathbf{G}\} \right). \tag{S.E.9}$$

Therefore, we have

$$\begin{aligned} \limsup_{n \to \infty} \mathbb{P}\left\{ AR_{CR,n} > \hat{c}_{AR,CR,n}(1-\alpha) \right\} &\leq \limsup_{n \to \infty} \mathbb{P}\left\{ AR_{CR,n} \geq \hat{c}_{AR,CR,n}(1-\alpha); E_n = 1 \right\} \\ &\leq \mathbb{P}\left\{ T_{AR}(S_{CR}) \geq T^{(k^*)}_{AR}(S_{CR}|\mathbf{G}) \right\}, \end{aligned}$$

which follows from (S.E.8), (S.E.9), the continuous mapping theorem and Portmanteau's theorem. The claim of the upper bound in the theorem then follows from similar arguments as those in Theorem 2.1. ■

S.F Proof of Theorem 2.6

Define $AR_\infty(g) = || \sum_{j \in J} g_j \sqrt{\xi_j} \mathcal{Z}_j + \sum_{j \in J} \xi_j g_j a_j Q_{\widetilde{Z}X} \mu ||_{A_z}$. In particular, notice that $AR_\infty(\iota) = || \sum_{j \in J} \sqrt{\xi_j} Z_j + Q_{\widetilde{Z}X} \mu ||_{A_z}$ since $\sum_{j \in J} \xi_j a_j = 1$. Following same arguments in the proof of Theorem 2.5, we can show that, under $\mathcal{H}_{1,n}$ with $\lambda_\beta = I_{d_x}$,

$$(AR_n, \{AR^*_n(g)\}_{g \in \mathbf{G}}) \xrightarrow{d} (AR_\infty(\iota), \{AR_\infty(g)\}_{g \in \mathbf{G}}).$$

Similar to the proofs of Theorems 2.2 and 2.4, in order to establish Theorem 2.6, it suffices to show that as $||Q_{\widetilde{Z}X}\mu||_2 \to \infty$,

$$\mathbb{P}\{AR_\infty(\iota) > \max_{g\in\mathbf{G}_s} AR_\infty(g)\} \to 1.$$

By Triangular inequality, we have $AR_\infty(\iota) \geq ||Q_{\widetilde{Z}X}\mu||_{A_z} - O_p(1)$, and

$$\max_{g\in\mathbf{G}_s} AR_\infty(g) \leq \max_{g\in\mathbf{G}_s} |\sum_{j\in J} g_j\xi_j a_j| ||Q_{\widetilde{Z}X}\mu||_{A_z} + O_p(1).$$

In addition, $\max_{g\in\mathbf{G}_s} |\sum_{j\in J} g_j\xi_j a_j| < 1$ so that as $||Q_{\widetilde{Z}X}\mu||_2 \to \infty$, $||Q_{\widetilde{Z}X}\mu||_{A_z} \to \infty$ and

$$||Q_{\widetilde{Z}X}\mu||_{A_z} - \max_{g\in\mathbf{G}_s} |\sum_{j\in J} g_j\xi_j a_j| ||Q_{\widetilde{Z}X}\mu||_{A_z} \to \infty.$$

This concludes the proof. ■

S.G Proof of Theorem 2.7

Notice that when $d_x = d_z = 1$, $\widehat{Q}_{\widetilde{Z}X}$ is a scalar, and the restricted TSLS estimator $\hat{\gamma}^r$ is equivalent to the restricted OLS estimator $\bar{\gamma}^r$, which is well defined by Assumption 1(iv), so that $\hat{\varepsilon}^r_{i,j} = \bar{\varepsilon}^r_{i,j}$. Therefore,

$$AR_n = \left|\frac{1}{\sqrt{n}}\sum_{j\in J}\sum_{i\in I_{n,j}} \widetilde{Z}_{i,j}\bar{\varepsilon}^r_{i,j}\right| = \left|\frac{1}{\sqrt{n}}\sum_{j\in J}\sum_{i\in I_{n,j}} \widetilde{Z}_{i,j}\hat{\varepsilon}^r_{i,j}\right|,$$

and whenever $\widehat{Q}_{\widetilde{Z}X} \neq 0$,

$$\begin{aligned}
T_n &= |\sqrt{n}(\hat{\beta} - \beta_n) + \sqrt{n}(\beta_n - \beta_0)| \\
&= \left|\widehat{Q}^{-1}_{\widetilde{Z}X}\frac{1}{\sqrt{n}}\sum_{j\in J}\sum_{i\in I_{n,j}} \left[\widetilde{Z}_{i,j}\varepsilon_{i,j} - \widetilde{Z}_{i,j}X_{i,j}(\hat{\beta}^r - \beta_n)\right]\right| \\
&= \left|\widehat{Q}^{-1}_{\widetilde{Z}X}\frac{1}{\sqrt{n}}\sum_{j\in J}\sum_{i\in I_{n,j}} \left[\widetilde{Z}_{i,j}\varepsilon_{i,j} - \widetilde{Z}_{i,j}X_{i,j}(\hat{\beta}^r - \beta_n) - \widetilde{Z}_{i,j}W^\top_{i,j}(\hat{\gamma}^r - \gamma)\right]\right| \\
&= \left|\widehat{Q}^{-1}_{\widetilde{Z}X}\frac{1}{\sqrt{n}}\sum_{j\in J}\sum_{i\in I_{n,j}} \widetilde{Z}_{i,j}\hat{\varepsilon}^r_{i,j}\right| = \left|\widehat{Q}^{-1}_{\widetilde{Z}X}\right| AR_n,
\end{aligned}$$

by $\beta_0 = \hat{\beta}^r$, $\sum_{j\in J}\sum_{i\in I_{n,j}} \widetilde{Z}_{i,j}W^\top_{i,j} = 0$, and $\hat{\varepsilon}^r_{i,j} = \varepsilon_{i,j} - X^\top_{i,j}(\hat{\beta}^r - \beta_n) - W^\top_{i,j}(\hat{\gamma}^r - \gamma)$.

In addition, for the bootstrap statistics we have

$$AR^*_n(g) = \left|\frac{1}{\sqrt{n}}\sum_{j\in J}\sum_{i\in I_{n,j}} g_j\widetilde{Z}_{i,j}\bar{\varepsilon}^r_{i,j}\right| = \left|\frac{1}{\sqrt{n}}\sum_{j\in J}\sum_{i\in I_{n,j}} g_j\widetilde{Z}_{i,j}\hat{\varepsilon}^r_{i,j}\right|$$

and whenever $\widehat{Q}_{\tilde{Z}X} \neq 0$,

$$T_n^{s*}(g) = \left| \widehat{Q}_{\widetilde{Z}X}^{-1} \frac{1}{\sqrt{n}} \sum_{j \in J} \sum_{i \in I_{n,j}} g_j \widetilde{Z}_{i,j} \hat{\varepsilon}_{i,j}^r \right| = \left| \widehat{Q}_{\widetilde{Z}X}^{-1} \right| AR_n^*(g).$$

Therefore, $1\{T_n > \hat{c}_n^s(1-\alpha)\}$ is equal to $1\{AR_n > \hat{c}_{AR,n}(1-\alpha)\}$ whenever $\widehat{Q}_{\widetilde{Z}X} \neq 0$. We conclude that $\liminf_{n\to\infty} \mathbb{P}\{\phi_n^s = \phi_n^{ar}\} = 1$ because $\liminf_{n\to\infty} \mathbb{P}\left\{\widehat{Q}_{\widetilde{Z}X} \neq 0\right\} = 1$. ■

S.H Wild Bootstrap for Other Weak-IV-Robust Statistics

In this section, we discuss wild bootstrap inference with other weak-IV-robust statistics. To introduce the test statistics, we define the sample Jacobian as

$$\widehat{G} = \left(\widehat{G}_1, ..., \widehat{G}_{d_x}\right) \in \mathbf{R}^{d_z \times d_x}, \quad \widehat{G}_l = n^{-1} \sum_{j \in J} \sum_{i \in I_{n,j}} \widetilde{Z}_{i,j} X_{i,j,l}, \text{ for } l = 1, ..., d_x,$$

and define the orthogonalized sample Jacobian as

$$\widehat{D} = \left(\widehat{D}_1, ..., \widehat{D}_{d_x}\right) \in \mathbf{R}^{d_z \times d_x}, \quad \widehat{D}_l = \widehat{G}_l - \widehat{\Gamma}_l \widehat{\Omega}^{-1} \widehat{f} \in \mathbf{R}^{d_z},$$

where $\widehat{\Omega} = n^{-1} \sum_{j \in J} \sum_{i \in I_{n,j}} \sum_{k \in I_{n,j}} f_{i,j} f_{k,j}^\top$, and $\widehat{\Gamma}_l = n^{-1} \sum_{j \in J} \sum_{i \in I_{n,j}} \sum_{k \in I_{n,j}} \left(\widetilde{Z}_{i,j} X_{i,j,l}\right) f_{k,j}^\top$, for $l = 1, ..., d_x$. Therefore, under the null $\beta_n = \beta_0$ and the framework where the number of clusters tends to infinity, $\widehat{D}$ equals the sample Jacobian matrix $\widehat{G}$ adjusted to be asymptotically independent of $\widehat{f}$.

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

$$LM_n = n \widehat{f}^\top \widehat{\Omega}^{-1/2} P_{\widehat{\Omega}^{-1/2} \widehat{D}} \widehat{\Omega}^{-1/2} \widehat{f}.$$

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

$$LR_n = \frac{1}{2} \left(AR_{CR,n} - rk_n + \sqrt{(AR_{CR,n} - rk_n)^2 + 4LM_n \cdot rk_n} \right),$$

where rk_n is a conditioning statistic and we let $rk_n = n \widehat{D}^\top \widehat{\Omega}^{-1} \widehat{D}$.[1]

The wild bootstrap procedure for the LM and CQLR tests is as follows. We compute

$$\widehat{D}_g^* = \left(\widehat{D}_{1,g}^*, ..., \widehat{D}_{d_x,g}^*\right), \quad \widehat{D}_{l,g}^* = \widehat{G}_l - \widehat{\Gamma}_{l,g}^* \widehat{\Omega}^{-1} \widehat{f}_g^*,$$
$$\widehat{\Gamma}_{l,g}^* = n^{-1} \sum_{j \in J} \sum_{i \in I_{n,j}} \sum_{k \in I_{n,j}} \left(\widetilde{Z}_{i,j} X_{i,j,l}\right) f_{k,j}^*(g_j)^\top, \; l = 1, ..., d_x,$$

for any $g = (g_1, ..., g_q) \in \mathbf{G}$, where the definition of $\widehat{f}_g^*$ and $f_{k,j}^*(g_j)$ is the same as that in Section 2.3.

[1]This choice follows Newey and Windmeijer (2009). Kleibergen (2005) uses alternative formula for rk_n, and Andrews and Guggenberger (2019) introduce alternative CQLR test statistic.

Then, we compute the bootstrap analogues of the test statistics as

$$LM_n^*(g) = n(\widehat{f}_g^*)^\top \widehat{\Omega}^{-1/2} P_{\widehat{\Omega}^{-1/2}\widehat{D}_g^*} \widehat{\Omega}^{-1/2} \widehat{f}_g^*,$$

$$LR_n^*(g) = \frac{1}{2}\left(AR_{CR,n}^*(g) - rk_n + \sqrt{\left(AR_{CR,n}^*(g) - rk_n\right)^2 + 4LM_n^*(g) \cdot rk_n} \right).$$

Let $\hat{c}_{LM,n}(1-\alpha)$ and $\hat{c}_{LR,n}(1-\alpha)$ denote the $(1-\alpha)$-th quantile of $\{LM_n^*(g)\}_{g\in\mathbf{G}}$ and $\{LR_n^*(g)\}_{g\in\mathbf{G}}$, respectively. We notice that with at least one strong cluster,

$$LM_n \xrightarrow{d} \left\| \left(\widetilde{D}^\top \left(\sum_{j\in J} \xi_j \mathcal{Z}_{\varepsilon,j} \mathcal{Z}_{\varepsilon,j}^\top \right)^{-1} \widetilde{D} \right)^{-1/2} \widetilde{D}^\top \left(\sum_{j\in J} \xi_j \mathcal{Z}_{\varepsilon,j} \mathcal{Z}_{\varepsilon,j}^\top \right)^{-1} \sum_{j\in J} \sqrt{\xi_j} \mathcal{Z}_{\varepsilon,j} \right\|^2,$$

where $\widetilde{D} = \left(\widetilde{D}_1, ..., \widetilde{D}_{d_x}\right)$, and for $l = 1, ..., d_x$,

$$\widetilde{D}_l = Q_{\widetilde{Z}X} - \left\{ \sum_{j\in J} \left(\xi_j Q_{\widetilde{Z}X,j,l}\right) \left(\sqrt{\xi_j} \mathcal{Z}_{\varepsilon,j}\right) \right\} \left\{ \sum_{j\in J} \xi_j \mathcal{Z}_{\varepsilon,j} \mathcal{Z}_{\varepsilon,j}^\top \right\}^{-1} \sum_{j\in J} \sqrt{\xi_j} \mathcal{Z}_{\varepsilon,j}.$$

Although the limiting distribution is nonstandard, we are able to establish the validity results by connecting the bootstrap LM test with the randomization test and by showing the asymptotic equivalence of the bootstrap LM and CQLR tests in this case. We conjecture that similar results can also be established for other weak-IV-robust statistics proposed in the literature.

Theorem S.H.1 *Suppose Assumptions 1, 2(i), and 3 hold, $\beta_n = \beta_0$, and $q > d_z$, then*

$$\alpha - \frac{1}{2^{q-1}} \leq \liminf_{n\to\infty} \mathbb{P}\{LM_n > \hat{c}_{LM,n}(1-\alpha)\} \leq \limsup_{n\to\infty} \mathbb{P}\{LM_n > \hat{c}_{LM,n}(1-\alpha)\} \leq \alpha + \frac{1}{2^{q-1}};$$

$$\alpha - \frac{1}{2^{q-1}} \leq \liminf_{n\to\infty} \mathbb{P}\{LR_n > \hat{c}_{LR,n}(1-\alpha)\} \leq \limsup_{n\to\infty} \mathbb{P}\{LR_n > \hat{c}_{LR,n}(1-\alpha)\} \leq \alpha + \frac{1}{2^{q-1}}.$$

S.I Proof of Theorem S.H.1

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

$$T_{LM}(s) \equiv \left\| \left(D(s)^\top \left(\sum_{j\in J} s_{2,j} s_{2,j}^\top \right)^{-1} D(s) \right)^{-1/2} D(s)^\top \left(\sum_{j\in J} s_{2,j} s_{2,j}^\top \right)^{-1} \sum_{j\in J} s_{2,j} \right\|^2, \tag{S.I.1}$$

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

$$D(s) \equiv (D_1(s), ..., D_{d_x}(s)),$$

$$D_l(s) \equiv \sum_{j\in J} s_{1,j,l} - \left(\sum_{j\in J} s_{1,j,l}s_{2,j}^\top\right)\left(\sum_{j\in J} s_{2,j}s_{2,j}^\top\right)^{-1}\sum_{j\in J} s_{2,j}, \tag{S.I.2}$$

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

$$S_n \equiv \left(\left\{\frac{1}{n}\sum_{i\in I_{n,j}} \tilde{Z}_{i,j}X_{i,j}^\top : j\in J\right\}, \left\{\frac{1}{\sqrt{n}}\sum_{i\in I_{n,j}} \tilde{Z}_{i,j}\varepsilon_{i,j} : j\in J\right\}\right), \tag{S.I.3}$$

Note that for $l = 1, ..., d_x$ and $j \in J$, by Assumptions 1(iii) and 2(i) we have

$$\frac{1}{n}\sum_{i\in I_{n,j}} \tilde{Z}_{i,j}X_{i,j,l} \xrightarrow{p} \xi_j Q_{\tilde{Z}X,j,l}, \tag{S.I.4}$$

where $Q_{\tilde{Z}X,j,l}$ denotes the l-th column of the $d_z \times d_x$-dimensional matrix $Q_{\tilde{Z}X,j}$. Then, by Assumptions 1(ii) and 1(iii), 3(i) and 3(ii), and the continuous mapping theorem we have

$$S_n \xrightarrow{d} \left(\left\{\xi_j a_j Q_{\tilde{Z}X} : j\in J\right\}, \left\{\sqrt{\xi_j}\mathcal{Z}_j : j\in J\right\}\right) \equiv S, \tag{S.I.5}$$

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

$$\widehat{D}_l = \sum_{j\in J}\left(\frac{1}{n}\sum_{i\in I_{n,j}}\tilde{Z}_{i,j}X_{i,j,l}\right) - \left(\sum_{j\in J}\left(\frac{1}{n}\sum_{i\in I_{n,j}}\tilde{Z}_{i,j}X_{i,j,l}\right)\left(\frac{1}{\sqrt{n}}\sum_{k\in I_{k,j}}\tilde{Z}_{k,j}\bar{\varepsilon}_{k,j}^r\right)^\top\right)$$
$$\cdot\left(\sum_{j\in J}\left(\frac{1}{\sqrt{n}}\sum_{i\in I_{i,j}}\tilde{Z}_{i,j}\bar{\varepsilon}_{i,j}^r\right)\left(\frac{1}{\sqrt{n}}\sum_{k\in I_{k,j}}\tilde{Z}_{k,j}\bar{\varepsilon}_{k,j}^r\right)^\top\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i\in I_{i,j}}\tilde{Z}_{i,j}\bar{\varepsilon}_{i,j}^r, \tag{S.I.6}$$

and $\frac{1}{\sqrt{n}}\sum_{i\in I_{n,j}}\tilde{Z}_{i,j}\bar{\varepsilon}_{i,j}^r = \frac{1}{\sqrt{n}}\sum_{i\in I_{n,j}}\tilde{Z}_{i,j}\varepsilon_{i,j} + o_p(1)$ by $\beta_n = \beta_0$ and Lemma S.B.3. In addition, we set $A_n \in \mathbf{R}$ to equal

$$A_n \equiv I\left\{\widehat{D} \text{ is of full rank value and } \widehat{\Omega} \text{ is invertible}\right\}, \tag{S.I.7}$$

and we have

$$\liminf_{n\to\infty}\mathbb{P}\{A_n = 1\} = 1, \tag{S.I.8}$$

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 column rank for all $j \in J_s$, by Assumptions 1(ii), 3(i), and 3(ii).

It follows that whenever $A_n = 1$,

$$(LM_n, \{LM_n^*(g) : g\in\mathbf{G}\}) = (T_{LM}(S_n), \{T_{LM}(gS_n) : g\in\mathbf{G}\}) + o_p(1). \tag{S.I.9}$$

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

$$T_{LM}^{(1)}(s|\mathbf{G}) \le ... \le T_{LM}^{|\mathbf{G}|}(s|\mathbf{G}). \tag{S.I.10}$$

Next, we have

$$\begin{aligned}
&\limsup_{n\to\infty} \mathbb{P}\left\{LM_n > \hat{c}_{LM,n}(1-\alpha)\right\} \\
&\leq \limsup_{n\to\infty} \mathbb{P}\left\{LM_n \geq \hat{c}_{LM,n}(1-\alpha); A_n = 1\right\} \\
&\leq \mathbb{P}\left\{T_{LM}(S) \geq \inf\left\{u \in \mathbf{R} : \frac{1}{|\mathbf{G}|}\sum_{g\in\mathbf{G}} I\{T_{LM}(gS) \leq u\} \geq 1-\alpha\right\}\right\}, \qquad \text{(S.I.11)}
\end{aligned}$$

where the final inequality follows from (S.I.3), (S.I.5), (S.I.8), (S.I.9), the continuous mapping theorem and Portmanteau's theorem. Therefore, setting $k^* \equiv \lceil |\mathbf{G}|(1-\alpha)\rceil$, we can obtain from (S.I.11) that

$$\begin{aligned}
&\limsup_{n\to\infty} \mathbb{P}\left\{LM_n > \hat{c}_{LM,n}(1-\alpha)\right\} \\
&\leq \mathbb{P}\left\{T_{LM}(S) > T_{LM}^{(k^*)}(S|\mathbf{G})\right\} + \mathbb{P}\left\{T_{LM}(S) = T_{LM}^{(k^*)}(S|\mathbf{G})\right\} \\
&\leq \alpha + \mathbb{P}\left\{T_{LM}(S) = T_{LM}^{(k^*)}(S|\mathbf{G})\right\}, \qquad \text{(S.I.12)}
\end{aligned}$$

where the final inequality follows by $gS \overset{d}{=} S$ for all $g \in \mathbf{G}$ and the properties of randomization tests. Then, we notice that for all $g \in \mathbf{G}$, $T_{LM}(gS) = T_{LM}(-gS)$ with probability 1, and $P\{T_{LM}(gS) = T_{LM}(\tilde{g}S)\} = 0$ for $\tilde{g} \notin \{g, -g\}$. Therefore,

$$\mathbb{P}\left\{T_{LM}(S) = T_{LM}^{(k^*)}(S|\mathbf{G})\right\} = \frac{1}{2^{q-1}}. \qquad \text{(S.I.13)}$$

The claim of the upper bound in the theorem then follows from (S.I.12) and (S.I.13). 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

$$\begin{aligned}
LR_n &= \frac{1}{2}\left\{AR_{CR,n} - rk_n + \sqrt{(AR_{CR,n} - rk_n)^2 + 4\cdot LM_n \cdot rk_n}\right\} \\
&= \frac{1}{2}\left\{AR_{CR,n} - rk_n + |AR_{CR,n} - rk_n|\sqrt{1 + \frac{4\cdot LM_n\cdot rk_n}{(AR_{CR,n} - rk_n)^2}}\right\} \\
&= \frac{1}{2}\left\{AR_{CR,n} - rk_n + |AR_{CR,n} - rk_n|\left(1 + 2\cdot LM_n \frac{rk_n}{(AR_{CR,n} - rk_n)^2}(1 + o_p(1))\right)\right\} \\
&= LM_n \frac{rk_n}{rk_n - AR_{CR,n}}(1 + o_p(1)) = LM_n + o_p(1), \qquad \text{(S.I.14)}
\end{aligned}$$

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 $AR_{CR,n} - rk_n < 0$ w.p.a.1 since $AR_{CR,n} = O_p(1)$ while $rk_n \to \infty$ w.p.a.1 under Assumption 3(i). Using arguments similar to those in (S.I.14), we obtain that for each $g \in \mathbf{G}$,

$$LR_n^*(g) = LM_n^*(g)\frac{rk_n}{rk_n - AR_{CR,n}^*(g)}(1 + o_p(1)) = LM_n^*(g) + o_p(1), \qquad \text{(S.I.15)}$$

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

$$(LR_n, \{LR_n^*(g) : g \in \mathbf{G}\}) = (T_{LM}(S_n), \{T_{LM}(gS_n) : g \in \mathbf{G}\}) + o_p(1). \tag{S.I.16}$$

Then, we obtain that

$$\begin{aligned}
&\limsup_{n\to\infty} \mathbb{P}\left\{LR_n > \hat{c}_{LR,n}(1-\alpha)\right\} \\
&\leq \limsup_{n\to\infty} \mathbb{P}\left\{LR_n \geq \hat{c}_{LR,n}(1-\alpha); A_n = 1\right\} \\
&\leq \mathbb{P}\left\{T_{LM}(S) \geq \inf\left\{u \in \mathbf{R} : \frac{1}{|\mathbf{G}|}\sum_{g\in\mathbf{G}} I\{T_{LM}(gS) \leq u\} \geq 1-\alpha\right\}\right\},
\end{aligned} \tag{S.I.17}$$

where the second inequality follows from (S.I.3), (S.I.5), (S.I.8), (S.I.16), 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. ■

References

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

Canay, I. A., A. Santos, and A. M. Shaikh (2021): "The wild bootstrap with a "small" number of "large" clusters," *Review of Economics and Statistics*, 103, 346–363.

Guggenberger, P., J. J. Ramalho, and R. J. Smith (2012): "GEL statistics under weak identification," *Journal of Econometrics*, 170, 331–349.

Kleibergen, F. (2002): "Pivotal statistics for testing structural parameters in instrumental variables regression," *Econometrica*, 70, 1781–1803.

——— (2005): "Testing parameters in GMM without assuming that they are identified," *Econometrica*, 73, 1103–1124.

Moreira, M. J. (2003): "A conditional likelihood ratio test for structural models," *Econometrica*, 71, 1027–1048.

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