8-2021

Wild bootstrap for instrumental variable regressions with weak and few clusters

Wenjie WANG

Yichong ZHANG
Singapore Management University, yczhang@smu.edu.sg

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

Wenjie Wang† and Yichong Zhang‡

October 5, 2021

Abstract

We study the wild bootstrap inference for instrumental variable (quantile) regressions in the framework of a small number of large clusters in which the number of clusters is viewed as fixed and the number of observations for each cluster diverges to infinity. For the subvector inference, we show that the wild bootstrap Wald test, with or without using the cluster-robust covariance matrix, 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 (AR) test for the full-vector inference and show that it controls size asymptotically up to a small error even under weak or partial identification for all clusters. We illustrate the good finite-sample performance of the new inference methods 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, Instrumental Variable Quantile Regression.

JEL codes: C12, C26, C31

1 Introduction

Various recent surveys in leading economic journals suggest that weak instruments remain important concerns for empirical practice. For instance, Andrews, Stock, and Sun (2019) survey 230 instrumental variable (IV) regressions from 17 papers published in the *American Economic Review* (AER). They find that many of the first-stage F-statistics (and nonhomoskedastic generalizations) are in a range that raises such concerns, and virtually all of these papers report at least one first-stage F with a value smaller than 10. Brodeur, Cook, and Heyes

*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 and the 2021 Econometric Society European Meeting for their valuable comments. 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.

(2020) investigate p-hacking and publication bias among over 21,000 hypothesis tests in 25 leading economic journals. They notice, in IV regressions, a sizable over-representation of first-stage F just over 10 (also observed in Andrews et al. (2019)), and studies with relatively weak instruments have a much higher proportion of second-stage t-statistics that are barely significant around 1.65 and 1.96.

The issue of weak instruments is further complicated by the fact that in many empirical settings, observations are clustered and the number of clusters is small (e.g., when clustering is based on states, provinces, neighboring countries, or industries). In such case, the commonly employed cluster-robust covariance estimator (CCE) is no longer consistent even under strong instruments, and the cluster-robust first-stage F test (Olea and Pflueger, 2013) cannot be directly applied as its critical values are obtained under the asymptotic framework with a large number of clusters. Recently, Young (2021) analyzes 1,359 IV regressions in 31 papers published by the American Economic Association (AEA) and highlights that many findings rest on unusually large values of test statistics (rather than coefficient estimates) due to *inaccurate estimates of covariance matrices* and are *highly sensitive to influential clusters or observations*: with the removal of just one cluster or observation, in the average paper, 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, with the average p-value rising from 0.029 to 0.154.

Motivated by these issues, we study the wild bootstrap inference for linear IV and IV quantile regressions (IVQR) with a small number of clusters by exploiting its connection with a randomization test based on the group of sign changes, following the lead of Canay, Santos, and Shaikh (2021). First, for both IV and IVQR, we show that the subvector inference based on a wild bootstrap Wald test, with or without the 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. We further establish conditions under which they have power against local alternatives (e.g., at least 5 strong clusters are required when the total number of clusters equals 10 and the nominal level α equals 10%). Second, for IV and IVQR, 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. Additionally, for IV regressions, the wild bootstrap Wald test without the CCE is numerically equivalent to a certain wild bootstrap AR test in the empirically prevalent case with single IV, implying that it is robust to weak identification. Third, for IV regressions we establish the validity result for bootstrapping weak-IV-robust tests other than the AR test with at least one strong cluster. Fourth, to provide subvector and full-vector inferences for IVQR with a small number of clusters, we propose a novel two-step gradient wild bootstrap procedure.

Our approach has several empirically relevant advantages. First, it enhances practitioners' toolbox by providing a reliable inference for IV and IVQR with a small number of clusters, as illustrated by our study on the average and distributional effects of Chinese imports on local labor markets in different US regions, following Autor, Dorn, and Hanson (2013). Second, it is flexible with IV strength: by allowing for cluster-level heterogeneity in the first stage, the bootstrap Wald test is robust to influential clusters, while its AR counterpart is fully robust to weak instruments. Third, different from widely used heteroskedasticity and autocorrelation consistent (HAC) estimators, our approach is agnostic about the within-cluster dependence structure and thus avoids the use of tuning parameters to estimate the covariance matrix.

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. (2019), 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 et al. (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) depend on strong identification for all clusters. Our bootstrap Wald test is more flexible as it does not require an equal cluster size and only needs strong identification in one of the clusters.

Second, our paper is related to the literature on weak IVs, and 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 normal approximation with influential observations or clusters. When implemented appropriately, bootstrap may substantially improve the inference for IV regressions,[2] and Moreira et al. (2009) establish the validity of bootstrap weak-IV-robust tests under weak IVs and homoskedasticity. While it is possible to extend their results by allowing the number of clusters to tend to infinity, such a framework could be unreasonable with influential clusters. We com-

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

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

plement this approach by establishing validity for these tests under the alternative asymptotics.

Third, our paper is related to the literature on IVQR.[3] Although IVQR can be estimated via GMM, for computational feasibility, we follow Chernozhukov and Hansen (2006) and implement a profiled estimation procedure, which falls outside the scope of wild bootstrap for GMM outlined by Canay et al. (2021). For the subvector inference, our gradient wild bootstrap procedure is similar to the one implemented by Hagemann (2017) for a linear quantile regression with clustered data and Jiang, Liu, Phillips, and Zhang (2021) for quantile treatment effect estimation in randomized control trials, but it imposes a null hypothesis and further involves a profiled minimization to obtain the bootstrap IVQR estimator. Also different from these two papers, we show bootstrap validity by connecting it to the randomization test. For the full-vector inference, we follow the idea by Chernozhukov and Hansen (2008a) and complement their results by considering a setup with a small number of large clusters and proposing a wild bootstrap AR test for IVQR.

Finally, we notice 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 in the homoskedastic case. Instead, we focus on developing inference methods that are robust to the first-stage heterogeneity for both IV and IVQR with clustered data.

The remainder of this paper is organized as follows. Sections 2 and 3 present the main results for IV and IVQR, respectively. Section 4 discusses cluster-level variables, whose values are invariant within each cluster. Section 5 provides simulation results and practical recommendations. Section 6 presents the empirical application.

2 Linear IV Regression

In this section, we consider the setup of a linear IV model with clustered data,

$$y_{i,j} = X_{i,j}^\top \beta + W_{i,j}^\top \gamma + \varepsilon_{i,j}, \quad X_{i,j} = Z_{i,j}^\top \Pi_{z,j} + W_{i,j}^\top \Pi_w + v_{i,j}, \tag{1}$$

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. The parameters $\beta \in \mathbf{R}^{d_x}$ and $\gamma \in \mathbf{R}^{d_w}$ are unknown structural parameters, while $\Pi_{z,j} \in \mathbf{R}^{d_z \times d_x}$ and $\Pi_w \in \mathbf{R}^{d_z \times d_w}$ are unknown parameters of the first stage. We allow for cluster-level heterogeneity with regard to IV strength by allowing $\Pi_{z,j}$ in (1) to vary across clusters.

[3] See, for example, Chernozhukov and Hansen (2004, 2005, 2006, 2008a), Su and Hoshino (2016), de Castro, Galvao, Kaplan, and Liu (2019), Wüthrich (2019), Chernozhukov, Hansen, and Wüthrich (2020), Alejo, Galvao, and Montes-Rojas (2021), and Kaido and Wüthrich (2021) for a comprehensive literature review.

This setting is motivated by the fact that in empirical studies, IVs are often strong for some subgroups and weak for some others, such as ethnic groups and geographic regions,[4] and, as noted previously, many TSLS estimates and first-stage Fs are highly sensitive to influential clusters. In experimental economics with clustered randomized trials, subjects' compliance with the assignment may also have substantial variations among clusters, resulting in heterogeneous IV strength.[5]

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

Assumption 1. *The following statements hold: (i) The quantities*

$$\frac{1}{\sqrt{n}}\sum_{j\in J}\sum_{i\in I_{n,j}}\begin{pmatrix} Z_{i,j}\varepsilon_{i,j} \\ W_{i,j}\varepsilon_{i,j}\end{pmatrix} \quad \text{and} \quad \frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}}\begin{pmatrix} Z_{i,j}Z_{i,j}^\top & Z_{i,j}W_{i,j}^\top \\ W_{i,j}Z_{i,j}^\top & W_{i,j}W_{i,j}^\top\end{pmatrix}$$

converges in distribution and converges in probability to a positive-definite matrix, respectively. (ii) The quantity

$$\frac{1}{n}\sum_{j\in J}\sum_{i\in I_{n,j}}\begin{pmatrix} Z_{i,j}X_{i,j}^\top \\ W_{i,j}X_{i,j}^\top\end{pmatrix}$$

converges in probability to a full rank matrix.

Assumption 2. *The following statements hold:*
(i) There exists a collection of independent random variables $\{\mathcal{Z}_j : j \in J\}$, where $\mathcal{Z}_j = [\mathcal{Z}_{\varepsilon,j} : \mathcal{Z}_{v,j}]$ with $\mathcal{Z}_{\varepsilon,j} \in \boldsymbol{R}^{d_z}$ and $\mathcal{Z}_{v,j} \in \boldsymbol{R}^{d_z\times d_x}$, and $vec(\mathcal{Z}_j) \sim N(0, \Sigma_j)$ with Σ_j positive definite for all $j \in J$, such that

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

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

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

[4] For instance, as remarked by Abadie et al. (2019), the return-to-schooling literature has often used compulsory schooling year as an IV for years of schooling while minimum school-leaving age is determined by state-specific laws, and in studies that use exogenous shock to oil or coal supply as an IV, states with large shares of oil or coal industries typically have strong first stage.

[5] One example is Muralidharan, Niehaus, and Sukhtankar (2016)'s evaluation of a smartcard payment system. In some villages, 90% or more of the recipients complied with the treatment assignment, while in many villages, less than 10% complied.

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

Remark 1. The above assumptions are similar to those imposed in Canay et al. (2021). Assumption 1 ensures that the TSLS estimators (and their null-restricted counterparts) are well behaved. Assumption 2(i) 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. The assumption that $\mathcal{Z}_j$ have full rank covariance matrices requires that $Z_{i,j}$ can not be expressed as a linear combination of $W_{i,j}$ within each cluster j. Assumption 2(ii) gives the restriction on cluster sizes. Assumption 2(iii) ensures $\widehat{\Gamma}_n$ is uniquely defined. Assumption 2(iv) is the sufficient condition for $\frac{1}{n_j}\sum_{i\in I_{n,j}} \tilde{Z}_{i,j}W_{i,j}^\top = o_p(1)$, which is needed in our proof. It holds under cluster homogeneity. As pointed out by Canay et al. (2021), this assumption is satisfied whenever the distributions of $(Z_{i,j}^\top, W_{i,j}^\top)^\top$ are the same across clusters. Furthermore, Assumption 2(iv) holds automatically if $W_{i,j}$ includes cluster dummies and their interactions with all other control variables. If $W_{i,j}$ includes cluster-level variables, then including cluster dummies would violate Assumption 2(iii). In Appendix A, we propose a cluster-level projection procedure, which ensures $\frac{1}{n_j}\sum_{i\in I_{n,j}} \tilde{Z}_{i,j}W_{i,j}^\top = 0$ for $j \in J$.

2.1 Subvector Inference

In this section, we study the properties of the wild bootstrap Wald test with a small number of large clusters. 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. 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 can be written as

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

We consider the test statistic with a $d_r \times d_r$ weighting matrix $\hat{A}_r$, that is,

$$T_n = ||\sqrt{n}(\lambda_\beta^\top \hat{\beta} - \lambda_0)||_{\hat{A}_r}, \tag{3}$$

where $||u||_A = \sqrt{u^\top A u}$ for a generic vector u and a weighting matrix A. The wild bootstrap test is implemented as follows:

Step 1: Compute the null-restricted residual $\hat{\varepsilon}^r_{i,j} = y_{i,j} - X_{i,j}^\top \hat{\beta}^r - W_{i,j}^\top \hat{\gamma}^r$, where $\hat{\beta}^r$ and $\hat{\gamma}^r$ are null-restricted TSLS estimators of β and γ.

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

$$y^*_{i,j}(g) = X_{i,j}^\top \hat{\beta}^r + W_{i,j}^\top \hat{\gamma}^r + g_j \hat{\varepsilon}^r_{i,j}. \tag{4}$$

For each $g = (g_1, ..., g_q) \in \mathbf{G}$ compute $\hat{\beta}^*_g$ and $\hat{\gamma}^*_g$, the analogues of the TSLS estimators $\hat{\beta}$ and $\hat{\gamma}$ using $y^*_{i,j}(g)$ in place of $y_{i,j}$ and the same $(Z^\top_{i,j}, X^\top_{i,j}, W^\top_{i,j})^\top$. Compute the bootstrap analogues of the test statistic:

$$T^*_n(g) \quad = \quad ||\sqrt{n}(\lambda^\top_\beta \hat{\beta}^*_g - \lambda_0)||_{\hat{A}^*_{r,g}}, \tag{5}$$

where $\hat{A}^*_{r,g}$ denotes the bootstrap weighting matrix, which will be specified below.

Step 3: To obtain the critical value, we compute the $1-\alpha$ quantile of $\{T^*_n(g) : g \in \mathbf{G}\}$:

$$\hat{c}_n(1-\alpha) = \inf \left\{ x \in \mathbf{R} : \frac{1}{|\mathbf{G}|} \sum_{g \in \mathbf{G}} I\{T^*_n(g) \leq x\} \geq 1 - \alpha \right\}, \tag{6}$$

where $I\{E\}$ equals one whenever the event E is true and equals zero otherwise. The bootstrap test for $\mathcal{H}_0$ rejects whenever T_n exceeds $\hat{c}_n(1-\alpha)$.

Let $\widehat{Q}_{\tilde{Z}X,j} = n_j^{-1} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} X^\top_{i,j}$ and $\widehat{Q}_{\tilde{Z}X} = n^{-1} \sum_{j \in J} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} X^\top_{i,j}$, where $\tilde{Z}_{i,j}$ are the residuals from regressing $Z_{i,j}$ on $W_{i,j}$ using the full sample (i.e., $\tilde{Z}_{i,j} = Z_{i,j} - \widehat{\Gamma}^\top_n W_{i,j}$ and $\widehat{\Gamma}_n$ is defined in Assumption 2(iv)). Let $Q_{\tilde{Z}X,j}$ and $Q_{\tilde{Z}X}$ denote the probability limits of $\widehat{Q}_{\tilde{Z}X,j}$ and $\widehat{Q}_{\tilde{Z}X}$, respectively. In addition, let $\widehat{Q}_{\tilde{Z}\tilde{Z}} = n^{-1} \sum_{j \in J} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} \tilde{Z}^\top_{i,j}$, $\widehat{Q} = \widehat{Q}^\top_{\tilde{Z}X} \widehat{Q}^{-1}_{\tilde{Z}\tilde{Z}} \widehat{Q}_{\tilde{Z}X}$, and let $Q_{\tilde{Z}\tilde{Z}}$ and Q denote the probability limits of $\widehat{Q}_{\tilde{Z}\tilde{Z}}$ and $\widehat{Q}$, respectively.

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

Assumption 3(i) states that in the case with one endogenous variable (the leading case in empirical applications), no further condition is required. With multiple endogenous variables, Assumption 3(ii) requires that $Q_{\tilde{Z}X,j} = a_j Q_{\tilde{Z}X}$ for all $j \in J$, where $a_j \neq 0$ for some clusters and $a_j = 0$ otherwise. Assumption 1 ensures that overall we have strong identification, and thus, $Q_{\tilde{Z}X}$ is of full rank, then when $a_j \neq 0$, $Q_{\tilde{Z}X,j}$ is also of full rank, i.e., the coefficients of the endogenous variables β_n are strongly identified in cluster j. We call these clusters the strong clusters. On the other hand, strong identification for β_n is not ensured in the rest of the clusters. For these clusters, Assumption 3(ii) excludes the case that the Jacobian matrix $Q_{\tilde{Z}X,j}$ is of a reduced rank but is not a zero matrix.[6]

We first study the asymptotic behaviors of the wild bootstrap Wald test when the weighting matrix $\hat{A}_r$ in (3) has a deterministic limit and the bootstrap weighting matrix $\hat{A}^*_{r,g}$ equals $\hat{A}_r$.

Assumption 4. $||\hat{A}_r - A_r||_{op} = o_p(1)$, *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,* $\lambda_{\min}(A)$

[6] It is possible to select out the clusters with Jacobian matrices of reduced rank via some testing procedure (Robin and Smith, 2000; Kleibergen and Paap, 2006; Chen and Fang, 2019). We leave this investigation for future research.

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*.*

Theorem 2.1. *(i) Suppose that Assumptions 1-4 hold. Then under* $\mathcal{H}_0$*,*

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

(ii) 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* $j \in J \backslash J_s$*,* $\lceil |\boldsymbol{G}|(1-\alpha)\rceil \leq |\boldsymbol{G}| - 2^{q-q_s+1}$*, where* $q_s = |J_s|$*, and let* $a_j = Q^{-1} Q_{\tilde{Z}X}^\top Q_{\tilde{Z}\tilde{Z}}^{-1} Q_{\tilde{Z}X,j}$ *when* $d_x = 1$*. Then under* $\mathcal{H}_{1,n}$*,*

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

Remark 2. Theorem 2.1 states that as long as there exists at least one strong cluster, the T_n-based wild bootstrap test is valid in the sense that its limiting null rejection probability is no greater than the nominal level α and no smaller than $\alpha - 1/2^{q-1}$, 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 and their bootstrap counterparts (i.e., $\frac{1}{\sqrt{n_j}} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} \varepsilon_{i,j}$ and $\frac{1}{\sqrt{n_j}} \sum_{i \in I_{n,j}} g_j \tilde{Z}_{i,j} \hat{\varepsilon}_{i,j}^r$ for $j \in J \backslash J_s$) still contribute to the limiting distribution of TSLS and the randomization with sign changes.

Additionally, one might consider employing an alternative procedure (e.g., see Moreira et al. (2009), Davidson and MacKinnon (2010), Finlay and Magnusson (2019), and Young (2021)):

$$X_{i,j}^*(g) = Z_{i,j}^\top \widehat{\Pi}_z + W_{i,j}^\top \widehat{\Pi}_w + g_j \hat{v}_{i,j}, \quad y_{i,j}^*(g) = X_{i,j}^{*\top}(g) \hat{\beta}^r + W_{i,j}^\top \hat{\gamma}^r + g_j \hat{\varepsilon}_{i,j}^r, \tag{7}$$

where $\widehat{\Pi}_z$ and $\widehat{\Pi}_w$ are some estimates of the first-stage coefficients. This procedure is asymptotically equivalent to ours under the current framework.

Remark 3. For the empirically prevalent case with a single endogenous variable and single IV (e.g., 101 out of 230 specifications in Andrews et al. (2019)'s sample and 1,087 out of 1,359 in Young (2021)'s sample), the T_n-based wild bootstrap test is numerically equivalent to a certain bootstrap AR test (the AR_n-based bootstrap test in Section 2.2), which is fully robust to weak IV. However, the more widely used wild bootstrap test with the CCE (e.g., Cameron et al. (2008), Cameron and Miller (2015), and Roodman, Nielsen, MacKinnon, and Webb (2019)) is not weak-IV robust and thus may produce size distortions in this case if the IV is weak (see Section 5.1). Furthermore, the robustness of the T_n-based bootstrap test depends on our specific procedure and thus cannot be extended to alternative ones such as the commonly employed pairs cluster bootstrap (block bootstrap—for example, see Cameron et al. (2008), Angrist and Pischke (2008), Goldsmith-Pinkham, Sorkin, and Swift (2020), and Hahn and Liao (2021)—including percentile, percentile-t, and bootstrap standard error).

Remark 4. For alternative inference methods with a small number of clusters, Bester et al. (2011) and Hwang (2021) provide asymptotic approximations that are based on t and F distributions, in the spirit of Kiefer and Vogelsang (2002, 2005)'s fixed-k asymptotics. Their approach would require stronger homogeneity conditions than the wild bootstrap in the current context; for example, the cluster sizes are approximately equal for all clusters ($n_j/\bar{n} \to 1$ for $j \in J$, where $\bar{n} = q^{-1}\sum_{j\in J} n_j$), the cluster-level scores in Assumption 2(i) have the same normal limiting distribution for all clusters, and the cluster-level Jacobian matrices in Assumption 3 have the same probability limit for all strong clusters (i.e., $a_j = 1$ for $j \in J_s$).

Furthermore, the wild bootstrap test has remarkable resemblance to the Fama-MacBeth type approach in Ibragimov and Müller (2010, hereafter IM) and the randomization test with sign changes in Canay et al. (2017, hereafter CRS), which are based on the asymptotic independence of cluster-level estimators (say, $\hat{\beta}_1, ..., \hat{\beta}_q$). However, we notice that for IV regressions (and IVQR), their size properties can be very different. For instance, IM and CRS rule out weak IVs in the sense of Staiger and Stock (1997) for all clusters (i.e., $\Pi_{z,j} = n_j^{-1/2}C_j$, where C_j has a fixed full rank value), as the cluster-level IV estimators of such weak clusters would become inconsistent and have highly nonstandard limiting distributions, violating their underlying assumptions.[7] By contrast, the size result in Theorem 2.1 holds even with only one strong cluster. In this sense, the wild bootstrap is more robust to cluster heterogeneity in IV strength. However, if all clusters are strong and the cluster-level estimators have minimal finite-sample bias, IM and CRS have an advantage over the wild bootstrap as they do not require Assumption 3(ii) when $d_x > 1$. The two types of approaches could therefore be considered as complements.

Remark 5. To establish in Theorem 2.1 the power of the wild bootstrap test against $n^{-1/2}$-local alternatives, we need a sufficient number of strong clusters and homogeneity of the signs of Jacobians for these strong clusters (i.e., $a_j > 0$ for each $j \in J_s$).[8] For instance, if q equals 10, the condition $\lceil|\mathbf{G}|(1-\alpha)\rceil \leq |\mathbf{G}| - 2^{q-q_s+1}$ requires that $q_s \geq 5$ and $q_s \geq 6$ for $\alpha = 10\%$ and 5%, respectively. Theorem 2.1 suggests that although the size of the wild bootstrap test is well controlled even when some clusters are weak, to enhance power, it might be beneficial to merge some less influential clusters.[9]

Remark 6. Theorem 2.1 can also be shown for other IV estimators such as the limited information maximum likelihood (LIML) and jackknife IV estimators (Phillips and Hale, 1977; Angrist, Imbens, and Krueger, 1999; Hausman, Newey, Woutersen, Chao, and Swanson, 2012).

[7] 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 rate will dominate in their statistics.

[8] The case with a single IV requires $\Pi_{z,j}$ to have the same sign across all clusters. The case with one endogenous variable and multiple (orthogonalized) IVs requires the sign of each element of $\Pi_{z,j}$ to be the same across all clusters.

[9] To avoid double dipping, we need prior knowledge of cluster-level first-stage heterogeneity, which may be available in practice (for example, when compulsory schooling years or exogenous shock to oil or coal supply are used as an IV).

For instance, let

$$(\hat{\beta}_{liml}^\top, \hat{\gamma}_{liml}^\top)^\top = (\bar{X}^\top P_{\bar{Z}} \bar{X} - \hat{\mu} \bar{X}^\top M_{\bar{Z}} \bar{X})^{-1} (\bar{X}^\top P_{\bar{Z}} Y - \hat{\mu} \bar{X}^\top M_{\bar{Z}} Y), \tag{8}$$

where $\hat{\mu} = \min_r r^\top \bar{Y}^\top M_W Z (Z^\top M_W Z)^{-1} Z^\top M_W \bar{Y} r / (r^\top \bar{Y}^\top M_{\bar{Z}} \bar{Y} r)$, $r = (1, -\beta^\top)^\top$, $\bar{Y} = [Y : X]$, $\bar{Z} = [Z : W]$, $\bar{X} = [X : W]$, Y, ε, X, Z and W are formed by $y_{i,j}$, $\varepsilon_{i,j}$, $X_{i,j}^\top$, $Z_{i,j}^\top$ 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. Then, by the definition of $\hat{\mu}$,

$$n\hat{\mu} \leq \left(\frac{1}{\sqrt{n}} e^\top M_W Z \right) \left(\frac{1}{n} Z^\top M_W Z \right)^{-1} \left(\frac{1}{\sqrt{n}} Z^\top M_W e \right) / \left(\frac{1}{n} e^\top M_{\bar{Z}} e \right) = O_P(1), \tag{9}$$

and LIML has the same limiting distribution as TSLS under Assumptions 1 and 2. The results in Theorem 2.1 therefore hold for the wild bootstrap test with LIML as well. We omit the details for brevity, but notice that LIML is less biased than TSLS in the over-identified case,[10] and its corresponding bootstrap tests can therefore have better finite-sample size control since the randomization requires a distributional symmetry around zero (see Section 5.1).

Now we consider a wild bootstrap test for the Wald statistic when the weighting matrix $\hat{A}_r$ equals $\hat{A}_{r,CR}$, the inverse of the CCE, where

$$\hat{A}_{r,CR} = \left(\lambda_\beta^\top \widehat{V} \lambda_\beta \right)^{-1}, \tag{10}$$

$\widehat{V} = \widehat{Q}^{-1} \widehat{Q}_{\tilde{Z}X}^\top \widehat{Q}_{\tilde{Z}\tilde{Z}}^{-1} \widehat{\Omega}_{CR} \widehat{Q}_{\tilde{Z}\tilde{Z}}^{-1} \widehat{Q}_{\tilde{Z}X} \widehat{Q}^{-1}$, $\widehat{\Omega}_{CR} = n^{-1} \sum_{j \in J} \sum_{i \in I_{n,j}} \sum_{k \in I_{n,j}} \tilde{Z}_{i,j} \tilde{Z}_{k,j}^\top \hat{\varepsilon}_{i,j} \hat{\varepsilon}_{k,j}$, $\hat{\varepsilon}_{i,j} = y_{i,j} - X_{i,j}^\top \hat{\beta} - W_{i,j}^\top \hat{\gamma}$, and $\hat{\beta}$ and $\hat{\gamma}$ are the TSLS estimators. The corresponding Wald statistic is

$$T_{CR,n} = ||\sqrt{n}(\lambda_\beta^\top \hat{\beta} - \lambda_0)||_{\hat{A}_{r,CR}}. \tag{11}$$

We follow a similar wild bootstrap procedure by defining

$$T_{CR,n}^*(g) = ||\sqrt{n}(\lambda_\beta^\top \hat{\beta}_g^* - \lambda_0)||_{\hat{A}_{r,CR,g}^*}, \quad \hat{A}_{r,CR,g}^* = \left(\lambda_\beta^\top \widehat{V}_g^* \lambda_\beta \right)^{-1}, \tag{12}$$

where $\widehat{V}_g^* = \widehat{Q}^{-1} \widehat{Q}_{\tilde{Z}X}^\top \widehat{Q}_{\tilde{Z}\tilde{Z}}^{-1} \widehat{\Omega}_{CR,g}^* \widehat{Q}_{\tilde{Z}\tilde{Z}}^{-1} \widehat{Q}_{\tilde{Z}X} \widehat{Q}^{-1}$, $\widehat{\Omega}_{CR,g}^* = n^{-1} \sum_{j \in J} \sum_{i \in I_{n,j}} \sum_{k \in I_{n,j}} \tilde{Z}_{i,j} \tilde{Z}_{k,j}^\top \hat{\varepsilon}_{i,j}^*(g) \hat{\varepsilon}_{k,j}^*(g)$, and $\hat{\varepsilon}_{i,j}^*(g) = y_{i,j}^*(g) - X_{i,j}^\top \hat{\beta}_g^* - W_{i,j}^\top \hat{\gamma}_g^*$. The bootstrap critical value $\hat{c}_{CR,n}(1-\alpha)$ is the $1-\alpha$ quantile of $\{T_{CR,n}^*(g) : g \in \mathbf{G}\}$. The asymptotic behavior of this test is given as follows.

Theorem 2.2. *(i) Suppose that Assumptions 1-3 hold, and $q > d_r$. Then under $\mathcal{H}_0$,*

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

(ii) 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 let $a_j = Q^{-1} Q_{\tilde{Z}X}^\top Q_{\tilde{Z}\tilde{Z}}^{-1} Q_{\tilde{Z}X,j}$

[10] It may be interesting to consider an alternative asymptotic framework in which the number of clusters is fixed but the number of IVs tends to infinity (Bekker, 1994). We leave this direction of investigation for future research.

when $d_x = 1$. Then under $\mathcal{H}_{1,n}$,

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

Remark 7. Theorem 2.2 states that with at least one strong cluster, the $T_{CR,n}$-based wild bootstrap test controls size asymptotically up to a small error, and it has power against local alternatives if the condition on the number of strong clusters is satisfied. Moreover, different from Theorem 2.1, the power result in Theorem 2.2 does not require the homogeneity condition on the sign of first-stage coefficients. Indeed, the critical values of the two tests have different asymptotic behaviors, with $\hat{c}_n(1-\alpha) \xrightarrow{p} \infty$ while $\hat{c}_{CR,n}(1-\alpha) = O_P(1)$, as $||\mu||_2 \to \infty$, which translates into relatively good power properties of the bootstrap test with the CCE. In particular, when $d_r = 1$, its local power is strictly higher when $||\mu||_2$ is sufficiently large.[11]

2.2 Full-Vector Inference

In general, Theorems 2.1 and 2.2 do not hold when all clusters are "weak." For instance, under the weak-IV sequence such that $\Pi_{z,j} = n_j^{-1/2} C_j$ with some fixed full rank C_j for all $j \in J$,

$$\frac{1}{\sqrt{n}} \sum_{j \in J} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} X_{i,j}^\top \xrightarrow{d} \sum_{j \in J} \sqrt{\xi_j} Q_{\tilde{Z}\tilde{Z},j} C_j + \sum_{j \in J} \sqrt{\xi_j} \mathcal{Z}_{v,j}, \tag{13}$$

where $Q_{\tilde{Z}\tilde{Z},j} = \text{plim} \frac{1}{n_j} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} \tilde{Z}_{i,j}^\top$ and $\sum_{j \in J} \sqrt{\xi_j} Q_{\tilde{Z}\tilde{Z},j} C_j$, the signal part of the first stage, is of the same order of magnitude as the noise part $\sum_{j \in J} \sqrt{\xi_j} \mathcal{Z}_{v,j}$. Then, the randomization with sign changes would be invalid because for each $j \in J$, (i) the distribution of $\sqrt{\xi_j} \left(Q_{\tilde{Z}\tilde{Z},j} C_j + \mathcal{Z}_{v,j} \right)$ is not symmetric around zero and (ii) C_j cannot be consistently estimated. For example, the bootstrap analogue of (13) under the procedure in (7) has the following limiting distribution:

$$\sum_{j \in J} \sqrt{\xi_j} Q_{\tilde{Z}\tilde{Z},j} C_j + \sum_{j \in J} \sqrt{\xi_j} g_j \mathcal{Z}_{v,j} + \sum_{j \in J} \xi_j (1 - g_j) Q_{\tilde{Z}\tilde{Z},j} Q_{\tilde{Z}\tilde{Z}}^{-1} \left(\sum_{\tilde{j} \in J} \sqrt{\xi_{\tilde{j}}} \mathcal{Z}_{v,\tilde{j}} \right), \tag{14}$$

where the second term equals the $\mathbf{G}$-transformed noise in (13) while the third is an extra term.

In the case that the parameter of interest may be weakly identified in all clusters or the homogeneity condition in Assumption 3(ii) may not hold, we may consider the inference on the full vector of β_n. Recall that $\beta_n = \beta_0 + \mu_\beta / \sqrt{n}$. Under the null, we have $\mu_\beta = 0$, or equivalently, $\beta_n = \beta_0$. Define the AR statistic with an asymptotically deterministic weighting matrix as

$$AR_n \quad = \quad \left\| \sqrt{n} \widehat{f} \right\|_{\hat{A}_z}, \tag{15}$$

where $\hat{A}_z$ is a $d_z \times d_z$ weighting matrix with an asymptotically deterministic limit, $\widehat{f} =$

[11] Note that as $||\mu||_2 \to \infty$, $\liminf_{n\to\infty} \mathbb{P}\{T_{CR,n} > \hat{c}_{CR,n}(1-\alpha)\} > \liminf_{n\to\infty} \mathbb{P}\{T_{CR,n} > \tilde{c}_n(1-\alpha)\} = \liminf_{n\to\infty} \mathbb{P}\{T_n > \hat{c}_n(1-\alpha)\}$, where $\tilde{c}_n(1-\alpha)$ denotes the $(1-\alpha)$ quantile of $\left\{ |\sqrt{n}(\lambda_\beta^\top \hat{\beta}_g^* - \lambda_0)| / \sqrt{\lambda_\beta^\top \widehat{V} \lambda_\beta} : g \in \mathbf{G} \right\}$, and the inequality holds because $\hat{c}_{CR,n}(1-\alpha) = O_P(1)$ while $\tilde{c}_n(1-\alpha) \xrightarrow{p} \infty$, since $\lambda_\beta^\top \widehat{V} \lambda_\beta = O_P(1)$.

$n^{-1}\sum_{j\in J}\sum_{i\in I_{n,j}} f_{i,j}$, $f_{i,j} = \tilde{Z}_{i,j}\hat{\varepsilon}^r_{i,j}$, $\hat{\varepsilon}^r_{i,j} = y_{i,j} - X_{i,j}^\top\beta_0 - W_{i,j}^\top\bar{\gamma}^r$, and $\bar{\gamma}^r$ is the null-restricted ordinary least squares (OLS) estimator of γ: $\bar{\gamma}^r = \left(\sum_{j\in J}\sum_{i\in I_{n,j}} W_{i,j}W_{i,j}^\top\right)^{-1}\sum_{j\in J}\sum_{i\in I_{n,j}} W_{i,j}(y_{i,j} - X_{i,j}^\top\beta_0)$. Additionally, we define the AR statistic with the (null-imposed) CCE as

$$AR_{CR,n} \;=\; \left\|\sqrt{n}\widehat{f}\right\|_{\hat{A}_{CR}}, \quad \hat{A}_{CR} = \left(n^{-1}\sum_{j\in J}\sum_{i\in I_{n,j}}\sum_{k\in I_{n,j}} f_{i,j}f_{k,j}^\top\right)^{-1}. \tag{16}$$

Another (Wald-type) AR statistic widely applied in practice is defined as[12]

$$AR_{R,n} = \left\|\sqrt{n}\hat{\delta}\right\|_{\hat{A}_R}, \quad \hat{A}_R = \left(\widehat{Q}^{-1}_{\tilde{Z}\tilde{Z}}\widehat{\Omega}_R\widehat{Q}^{-1}_{\tilde{Z}\tilde{Z}}\right)^{-1}, \tag{17}$$

where $\widehat{\Omega}_R = n^{-1}\sum_{j\in J}\sum_{i\in I_{n,j}}\sum_{k\in I_{n,j}}\tilde{Z}_{i,j}\tilde{Z}_{k,j}^\top\hat{u}_{i,j}\hat{u}_{k,j}$, $\hat{u}_{i,j}$ is the residual of regressing $y_i - X_i^\top\beta_0$ on $Z_{i,j}$ and $W_{i,j}$, and $\hat{\delta}$ is the corresponding OLS estimate of the coefficient of $Z_{i,j}$. Our wild bootstrap procedure is defined as follows.

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

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

$$\widehat{f}^*_g = n^{-1}\sum_{j\in J}\sum_{i\in I_{n,j}} f^*_{i,j}(g_j), \quad \text{and} \quad \hat{\delta}^*_g = \widehat{Q}^{-1}_{\tilde{Z}\tilde{Z}}\widehat{f}^*_g, \tag{18}$$

where $f^*_{i,j}(g_j) = \tilde{Z}_{i,j}\varepsilon^*_{i,j}(g_j)$ and $\varepsilon^*_{i,j}(g_j) = g_j\hat{\varepsilon}^r_{i,j}$. Compute the bootstrap statistics: $AR^*_n(g) = \left\|\sqrt{n}\widehat{f}^*_g\right\|_{\hat{A}_z}$, $AR^*_{CR,n}(g) = \left\|\sqrt{n}\widehat{f}^*_g\right\|_{\hat{A}_{CR}}$, $AR^*_{R,n}(g) = \left\|\sqrt{n}\hat{\delta}^*_g\right\|_{\hat{A}^*_{R,g}}$, where $\hat{A}^*_{R,g} = \left(\widehat{Q}^{-1}_{\tilde{Z}\tilde{Z}}\widehat{\Omega}^*_{R,g}\widehat{Q}^{-1}_{\tilde{Z}\tilde{Z}}\right)^{-1}$, $\widehat{\Omega}^*_{R,g} = n^{-1}\sum_{j\in J}\sum_{i\in I_{n,j}}\sum_{k\in I_{n,j}}\tilde{Z}_{i,j}\tilde{Z}_{k,j}^\top\hat{u}^*_{i,j}(g_j)\hat{u}^*_{k,j}(g_j)$, and $\hat{u}^*_{i,j}(g_j)$ equals the residual of regressing $\varepsilon^*_{i,j}(g_j)$ on $Z_{i,j}$ and $W_{i,j}$.

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

Remark 8. Unlike the $T_{CR,n}$-based Wald test in Section 2.1, 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. On the other hand, we do need to bootstrap the CCE for the $AR_{R,n}$ test as $\hat{A}_R$, similar to that for the Wald test, is variant to the sign changes.

Theorem 2.3 shows that in the general case with multiple IVs, the AR_n-based bootstrap test is fully robust to weak IVs and few clusters, and those based on $AR_{CR,n}$ and $AR_{R,n}$ control size asymptotically up to a small error when $q > d_z$. Notice that the $AR_{R,n}$ test requires homogeneity condition similar to that imposed in Canay et al. (2021) for OLS-based Wald test.

[12] See, for example, Chernozhukov and Hansen (2008b), Finlay and Magnusson (2009), Cameron and Miller (2015), Andrews et al. (2019), and Roodman et al. (2019). This AR statistic is based on the reduced form of (1): $\tilde{y}_{i,j}(\beta_0) = Z_{i,j}^\top\delta_n + W_{i,j}^\top\theta_n + u_{i,j}$, where $\tilde{y}_{i,j}(\beta_0) = y_{i,j} - X_{i,j}^\top\beta_0$, $\delta_n = \Pi_z(\beta_n - \beta_0)$, $\theta_n = \Pi_w(\beta_n-\beta_0)+\gamma$, and $u_{i,j} = v_{i,j}^\top(\beta_n-\beta_0)+\varepsilon_{i,j}$. If the IVs $Z_{i,j}$ are valid, then testing $\beta_n = \beta_0$ is equivalent to testing $\delta_n = 0$ by using a Wald test.

Theorem 2.3. *Suppose that Assumptions 1(i) and 2 hold, and* $\beta_n = \beta_0$*. For* AR_n*, further suppose that* $||\hat{A}_z - A_z||_{op} = o_p(1)$*, 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*. For* $AR_{R,n}$*, further suppose that for each* $j \in J$*,* $\widehat{Q}_{\tilde{Z}\tilde{Z},j} \xrightarrow{p} b_j Q_{\tilde{Z}\tilde{Z}}$*, where* $b_j \neq 0$ *and* $Q_{\tilde{Z}\tilde{Z}}$ *is positive definite. Then,*

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

If further $q > d_z$*, then for* $h \in \{CR, R\}$*,*

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

Remark 9. 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), and Andrews and Guggenberger (2019)). Further complication therefore arises when all the clusters are weak for a reason similar to that noted in (14). Additionally, with few clusters this adjusted Jacobian is no longer asymptotically independent from the score. However, with at least one strong cluster, we are still able to establish the validity results. Further details are given in Appendix C .

For the weak-IV-robust subvector inference, one may use a projection approach (Dufour and Taamouti, 2005) after implementing the bootstrap AR tests for β_n, but the result may be rather conservative. Alternative subvector inference approaches (e.g., see Section 5.3 in Andrews et al. (2019)) provide a power improvement over the projection under the framework with a large number of observations/clusters, but they cannot be directly applied in the current context for a reason similar to that noted above.[13] To enhance power, we may apply the methods in Section 2.1 if we are confident that β_n is strongly identified in at least one of the clusters.

3 Linear IV Quantile Regression

In this section, we consider the linear IV quantile regression (IVQR) following the setup in Chernozhukov and Hansen (2006). We propose a gradient wild bootstrap procedure for IVQR and obtain results parallel to those in Theorems 2.1-2.3. Similar to the previous section, we allow the coefficients to shift w.r.t. the sample size to incorporate the analyses of size and local power in a concise manner. We define that $X_{i,j} \in \mathbf{R}^{d_x}$ contains endogenous covariates, whose coefficients are the parameters of interest, $W_{i,j} \in \mathbf{R}^{d_w}$ contains the exogenous control variables, and $Z_{i,j}$ contains exogenous variables that are excluded from the regression. The variable

[13]The asymptotic critical values given by these approaches will no longer be valid with a small number of clusters. Also, bootstrap tests based on the subvector statistics therein may not be robust to weak IVs even under conditional homoskedasticity (Wang and Doko Tchatoka, 2018).

$\Phi_{i,j}(\tau) \in \mathbf{R}^{d_\phi}$ contains instrumental variables (IVs) that are constructed from $(W_{i,j}, Z_{i,j})$, that is, $\Phi_{i,j}(\tau) = \Phi(W_{i,j}, Z_{i,j}, \tau)$. The function $\Phi(\cdot)$ may be unknown but can be estimated as $\hat{\Phi}(\cdot)$. The corresponding feasible IVs are defined as $\hat{\Phi}_{i,j}(\tau) = \hat{\Phi}(W_{i,j}, Z_{i,j}, \tau)$. Additionally, a scalar nonnegative weight is defined as $V_{i,j}(\tau)$, which also may be unknown, and its estimator is defined as $\hat{V}_{i,j}(\tau)$. Then the DGP of our IVQR is summarized in the following assumption. Throughout this section, for a generic function g of data $D_{i,j} = (y_{i,j}, X_{i,j}, W_{i,j}, Z_{i,j})$, we let $\mathbb{P}_n g(D_{i,j}) = \frac{1}{n}\sum_{j \in J}\sum_{i \in I_{n,j}} g(D_{i,j})$, $\overline{\mathbb{P}}_n g(D_{i,j}) = \frac{1}{n}\sum_{j \in J}\sum_{i \in I_{n,j}} \mathbb{E}g(D_{i,j})$, $\mathbb{P}_{n,j} g(D_{i,j}) = \frac{1}{n_j}\sum_{i \in I_{n,j}} g(D_{i,j})$, and $\overline{\mathbb{P}}_{n,j} g(D_{i,j}) = \frac{1}{n_j}\sum_{i \in I_{n,j}} \mathbb{E}g(D_{i,j})$.

Assumption 5. *(i) Suppose* $\mathbb{P}(y_{i,j} \leq X_{i,j}^\top \beta_n(\tau) + W_{i,j}^\top \gamma_n(\tau) | W_{i,j}, Z_{i,j}) = \tau$ *for* $\tau \in \Upsilon$, *where* Υ *is a compact subset of* $(0,1)$, $\beta_n(\tau) = \beta_0(\tau) + \mu_\beta(\tau)/\sqrt{n}$, *and* $\gamma_n(\tau) = \gamma_0(\tau) + \mu_\gamma(\tau)/\sqrt{n}$.

(ii) Suppose $\sup_{\tau \in \Upsilon} (||\mu_\gamma(\tau)||_2 + ||\mu_\beta(\tau)||_2 + ||\beta_0(\tau)||_2 + ||\gamma_0(\tau)||_2) \leq C < \infty$.

(iii) For all $\tau \in \Upsilon$, $\beta_n(\tau) \in int(\mathcal{B})$, *where* $\mathcal{B}$ *is compact and convex.*

(iv) Suppose $\max_{i \in [n_j], j \in J} \sup_{y \in \boldsymbol{R}} f_{y_{i,j}|W_{i,j}, X_{i,j}, Z_{i,j}}(y) < C$ *for some constant* $C \in (0, \infty)$, *where* $f_{y_{i,j}|W_{i,j}, X_{i,j}, Z_{i,j}}(\cdot)$ *denotes the conditional density of* $Y_{i,j}$ *given* $W_{i,j}, X_{i,j}, Z_{i,j}$.

(v) Denote $\Pi(b, r, t, \tau) = \overline{\mathbb{P}}_n(\tau - 1\{y_{i,j} < X_{i,j}^\top b + W_{i,j}^\top r + \Phi_{i,j}^\top(\tau) t\})\Psi_{i,j}(\tau)$, *where* $\Psi_{i,j}(\tau) = V_{i,j}(\tau) \cdot [W_{i,j}^\top, \Phi_{i,j}^\top(\tau)]^\top$. *Then, there are compact subsets* $\mathcal{R}$ *and* Θ *of* $\boldsymbol{R}^{d_w}$ *and* $\boldsymbol{R}^{d_\phi}$, *respectively, such that Jacobian matrix* $\frac{\partial}{\partial(r^\top, t^\top)}\Pi(b, r, t, \tau)$ *is continuous and has full column rank, uniformly in* n *and over* $\mathcal{B} \times \mathcal{R} \times \Theta \times \Upsilon$.

(vi) $\sup_{i \in I_j, j \in \mathcal{J}, \tau \in \Upsilon} \mathbb{E}||\Psi_{i,j}(\tau)||^{2+a} < \infty$ *for some* $a > 0$.

Remark 10. First, Assumption 5 allows for the case in which $\beta_n(\tau)$ is partially or weakly identified as we do not require the Jacobian matrix w.r.t. β, γ (i.e., $\frac{\partial}{\partial(b^\top, r^\top)}\Pi(b, r, 0, \tau)$) to be of full rank. Such a condition is assumed later in Assumption 8 when we do need point identification for the subvector inference. Second, under Assumption 5, Chernozhukov and Hansen (2006) show that $(\gamma_n^\top(\tau), 0_{d_\phi \times 1}^\top)^\top$ is the unique solution to the weighted quantile regression of $y_{i,j} - X_{i,j}^\top \beta_n(\tau)$ on $W_{i,j}$ and $\Phi_{i,j}(\tau)$ at the population level. Then, the inference of $\beta_n(\tau)$ can be implemented via a profiled method described below. For a given value b of $\beta_n(\tau)$, we first compute

$$(\hat{\gamma}(b, \tau), \hat{\theta}(b, \tau)) = \arg\inf_{r,t} \sum_{j \in J} \sum_{i \in I_{n,j}} \rho_\tau(y_{i,j} - X_{i,j}^\top b - W_{i,j}^\top r - \hat{\Phi}_{i,j}^\top(\tau) t)\hat{V}_{i,j}(\tau), \tag{19}$$

where $\rho_\tau(u) = u(\tau - 1\{u \leq 0\})$. For the subvector inference, we will assume the identification of $\beta_n(\tau)$ and proceed to estimate it by $\hat{\beta}(\tau)$ defined as

$$\hat{\beta}(\tau) = \arg\inf_{b \in \mathcal{B}} ||\hat{\theta}(b, \tau)||_{\hat{A}_\phi(\tau)}, \tag{20}$$

where $\mathcal{B}$ is a compact subset of $\mathbf{R}^{d_x}$ and $\hat{A}_\phi(\tau)$ is some $d_\phi \times d_\phi$ weighting matrix. Last, we define $\hat{\gamma}(\tau) = \hat{\gamma}(\hat{\beta}(\tau), \tau)$ and $\hat{\theta}(\tau) = \hat{\theta}(\hat{\beta}(\tau), \tau)$. For the full-vector inference, we will not rely on the point identification of $\beta_n(\tau)$. Instead, we use AR-type test statistics in which b is evaluated at the null hypothesis. In the following, we state the regularity conditions for both subvector and full-vector inferences.

Assumption 6. *(i) Let*

$$\hat{Q}_n(b,r,t,\tau) = \mathbb{P}_n \rho_\tau(y_{i,j} - X_{i,j}^\top b - W_{i,j}^\top r - \hat{\Phi}_{i,j}^\top(\tau) t)\hat{V}_{i,j}(\tau),$$
$$Q_n(b,r,t,\tau) = \overline{\mathbb{P}}_n \rho_\tau(y_{i,j} - X_{i,j}^\top b - W_{i,j}^\top r - \Phi_{i,j}^\top(\tau) t) V_{i,j}(\tau),$$

and $Q_\infty(b,r,t,\tau) = \lim_{n\to\infty} Q_n(b,r,t,\tau)$. Suppose $(\gamma_n(b,\tau), \theta_n(b,\tau))$ and $(\gamma_\infty(b,\tau), \theta_\infty(b,\tau))$ are the unique minimizers of $Q_n(b,r,t,\tau)$ and $Q_\infty(b,r,t,\tau)$ with respect to (r,t), respectively. In addition, suppose $(\gamma_n(b,\tau), \theta_n(b,\tau), \gamma_\infty(b,\tau), \theta_\infty(b,\tau))$ are continuous in $b \in \mathcal{B}$ uniformly over $\tau \in \Upsilon$, $(\gamma_n(b,\tau), \theta_n(b,\tau)) \in int(\mathcal{R} \times \Theta)$ for all $(b,\tau) \in \mathcal{B} \times \Upsilon$ where $\mathcal{R}$ and Θ are defined in Assumption 8, $\theta_\infty(b,\tau)$ has a unique root for $\tau \in \Upsilon$,

$$\sup_{(b,\tau)\in\mathcal{B}\times\Upsilon} |Q_\infty(b,r,t,\tau) - Q_n(b,r,t,\tau)| = o_p(1), \text{ and } \sup_{(b,\tau)\in\mathcal{B}\times\Upsilon} |\hat{Q}_n(b,r,t,\tau) - Q_n(b,r,t,\tau)| = o_p(1).$$

(ii) For any $\varepsilon > 0$,

$$\lim_{\delta\to 0} \limsup_{n\to\infty} \mathbb{P}\left(\begin{array}{c} \sup \Big\| \sqrt{n_j}(\mathbb{P}_{n,j} - \overline{\mathbb{P}}_{n,j}) \Big(\hat{f}_\tau(D_{i,j}, \beta_n(\tau) + v_b, \gamma_n(\tau) + v_r, v_t) \\ - f_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0) \Big) \Big\|_2 \geq \varepsilon \end{array} \right) = 0,$$

where the supremum inside the probability is taken over $\{j \in J, ||v||_2 \leq \delta, \tau \in \Upsilon\}$,

$$f_\tau(D_{i,j}, b, r, t) = (\tau - 1\{y_{i,j} - X_{i,j}^\top b - W_{i,j}^\top r - \Phi_{i,j}^\top(\tau) t \leq 0\})\Psi_{i,j}(\tau),$$

$$\hat{f}_\tau(D_{i,j}, b, r, t) = (\tau - 1\{y_{i,j} - X_{i,j}^\top b - W_{i,j}^\top r - \hat{\Phi}_{i,j}^\top(\tau) t \leq 0\})\hat{\Psi}_{i,j}(\tau),$$

$v = (v_b^\top, v_r^\top, v_t^\top)^\top$, $\hat{\Psi}_{i,j}(\tau) = \hat{V}_{i,j}(\tau) \cdot [W_{i,j}, \hat{\Phi}_{i,j}(\tau)]^\top$, and $\mathbb{E}\hat{g}(W_{i,j})$ is interpreted as $\mathbb{E}g(W_{i,j})|_{g=\hat{g}}$ following the convention in the empirical processes literature.

(iii) Denote $\varepsilon_{i,j}(\tau) = y_{i,j} - X_{i,j}^\top \beta_n(\tau) - W_{i,j}^\top \gamma_n(\tau)$, $\pi = (\gamma^\top, \theta^\top)^\top$ for generic (γ, θ) and $\delta_{i,j}(v,\tau) = X_{i,j}^\top v_b + W_{i,j}^\top v_r + \hat{\Phi}_{i,j}^\top(\tau) v_t$. Then, for any $\varepsilon > 0$, we have

$$\lim_{\delta\to 0} \limsup_{n\to\infty} \mathbb{P}\left[\sup \left\| \overline{\mathbb{P}}_{n,j} f_{\varepsilon_{i,j}(\tau)}(\delta_{i,j}(v,\tau)|W_{i,j}, Z_{i,j}) \hat{\Psi}_{i,j}(\tau)\hat{\Psi}_{i,j}^\top(\tau) - \mathcal{J}_{\pi,\pi,j}(\tau) \right\|_{op} \geq \varepsilon \right] = 0, \quad \text{and}$$
$$\lim_{\delta\to 0} \limsup_{n\to\infty} \mathbb{P}\left[\sup \left\| \overline{\mathbb{P}}_{n,j} f_{\varepsilon_{i,j}(\tau)}(\delta_{i,j}(v,\tau)|W_{i,j}, Z_{i,j}) \hat{\Psi}_{i,j}(\tau) X_{i,j}^\top - \mathcal{J}_{\pi,\beta,j}(\tau) \right\|_{op} \geq \varepsilon \right] = 0,$$

where the suprema inside the probability are taken over $\{j \in J, ||v||_2 \leq \delta, \tau \in \Upsilon\}$, $v =$

$(v_b^\top, v_r^\top, v_t^\top)^\top$,

$$\mathcal{J}_{\pi,\beta,j}(\tau) = \lim_{n\to\infty} \overline{\mathbb{P}}_{n,j} f_{\varepsilon_{i,j}(\tau)}(0|W_{i,j}, Z_{i,j})\Psi_{i,j}(\tau)X_{i,j}^\top, \quad \textit{and}$$
$$\mathcal{J}_{\pi,\pi,j}(\tau) = \lim_{n\to\infty} \overline{\mathbb{P}}_{n,j} f_{\varepsilon_{i,j}(\tau)}(0|W_{i,j}, Z_{i,j})\Psi_{i,j}(\tau)\Psi_{i,j}^\top(\tau).$$

(iv) $\sup_{\tau\in\Upsilon} \sqrt{n}||\mathbb{P}_n f_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0)||_2 = O_p(1)$.

(v) Let n_j be the size of the j-th cluster. Then we treat the number of clusters q as fixed and $n_j/n \to \xi_j$ for $\xi_j > 0$ and $j \in J$.

(vi) We further write $\mathcal{J}_{\pi,\pi,j}(\tau)$ as $\begin{pmatrix} \mathcal{J}_{\gamma,\gamma,j}(\tau) & \mathcal{J}_{\gamma,\theta,j}(\tau) \\ \mathcal{J}_{\gamma,\theta,j}^\top(\tau) & \mathcal{J}_{\theta,\theta,j}(\tau) \end{pmatrix}$, where $\mathcal{J}_{\gamma,\gamma,j}(\tau)$, $\mathcal{J}_{\gamma,\theta,j}(\tau)$, and $\mathcal{J}_{\theta,\theta,j}(\tau)$ are $d_w \times d_w$, $d_w \times d_\phi$, and $d_\phi \times d_\phi$ matrices. Then, there exist constants (c, C) such that

$$0 < c < \inf_{\tau\in\Upsilon} \lambda_{\min}\left(\sum_{j\in J} \xi_j \mathcal{J}_{\gamma,\gamma,j}(\tau)\right) < \sup_{\tau\in\Upsilon} \lambda_{\max}\left(\sum_{j\in J} \xi_j \mathcal{J}_{\gamma,\gamma,j}(\tau)\right) < C < \infty.$$

Remark 11. First, Assumption 6(i) ensures $\gamma_n(b,\tau)$ and $\theta_n(b,\tau)$ are uniquely defined in the drifting parameters setting. Second, Assumption 6(ii) is the stochastic equicontinuity of the empirical process

$$\sqrt{n_j}(\mathbb{P}_{n,j} - \overline{\mathbb{P}}_{n,j})\Big(\hat{f}_\tau(D_{i,j}, \beta_n(\tau) + v_b, \gamma_n(\tau) + v_r, v_t) - f_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0)\Big)$$

with respect to v. Such condition is verified by Chernozhukov and Hansen (2006) when the data are i.i.d. and $\hat{V}_{i,j}(\tau)$ and $\hat{\Phi}_{i,j}(\tau)$ uniformly converge to their population counterparts in probability. Their argument can be extended to data with weak dependence. Third, Assumption 6(iii) requires the uniform consistency of the Jacobian matrices, which holds even when observations are dependent. Fourth, Assumption 6(iv) requires the convergence rate of the sample mean of the score function to be parametric. In the case that we have a panel dataset and clusters are defined at the individual level so that observations in each cluster are just an individual-level time series, Assumption 6(iv) excludes non-stationarity and long-memory dependence. Fifth, Assumption 6(v) implies that we focus on the case with a small number of large clusters.

Assumption 7. *(i) For $j \in J$ and $\tau \in \Upsilon$, $\mathcal{J}_{\gamma,\theta,j}(\tau) = 0$.*

(ii) There exist versions of tight Gaussian processes $\{\mathcal{Z}_j(\tau) : \tau \in \Upsilon\}_{j\in J}$ such that $\mathcal{Z}_j(\tau) \in \boldsymbol{R}^{d_\phi}$, $\mathcal{Z}_j(\cdot)$ are independent across $j \in J$, $\mathbb{E}\mathcal{Z}_j(\tau)\mathcal{Z}_j^\top(\tau') = \Sigma_j(\tau,\tau')$,

$$0 < c < \inf_{\tau\in\Upsilon, j\in J} \lambda_{\min}(\Sigma_j(\tau,\tau)) \le \sup_{\tau\in\Upsilon, j\in J} \lambda_{\max}(\Sigma_j(\tau,\tau)) \le C < \infty$$

for some constants (c, C) independent of n, and

$$\sup_{j \in J, \tau \in \Upsilon} ||\sqrt{n_j}\mathbb{P}_{n,j}\tilde{f}_\tau(D, \beta_n(\tau), \gamma_n(\tau), 0) - \mathcal{Z}_j(\tau)||_2 \xrightarrow{p} 0,$$

where $\tilde{f}_\tau(D, \beta_n(\tau), \gamma_n(\tau), 0) = (\tau - 1\{\varepsilon_{i,j}(\tau) \leq 0\})V_{i,j}(\tau)\Phi_{i,j}(\tau)$.

Remark 12. First, we provide below a full-sample projection method to construct IVs that satisfy Assumption 7(i). Such construction forces the Jacobian matrix to be block-diagonal and thus introduces Neyman orthogonality between the estimators of the coefficients of the endogenous and control variables. Second, we discuss a cluster-level projection method in Appendix A. Third, consistently estimating $\Sigma_j(\cdot)$ in Assumption 7(ii) requires further assumptions on the within-cluster dependence structure and potential tuning parameters. Instead, the key benefit of our approach is that it is fully agnostic about the expression of the covariance matrices.

Now we describe the projection method to construct IVs that satisfy Assumption 7(i). Suppose $\underline{\Phi}_{i,j}(\tau)$ and $\hat{\underline{\Phi}}_{i,j}(\tau)$ are the original (infeasible) IVs and their estimators, respectively, such that Assumption 7(i) is violated (i.e., $\limsup_{n\to\infty} \overline{\mathbb{P}}_{n,j} f_{\varepsilon_{i,j}(\tau)}(0|W_{i,j}, Z_{i,j})W_{i,j}(\tau)\underline{\Phi}_{i,j}^\top(\tau) \neq 0$). We construct $\hat{\Phi}_{i,j}(\tau)$ as $\hat{\Phi}_{i,j}(\tau) = \hat{\underline{\Phi}}_{i,j}(\tau) - \hat{\chi}^\top(\tau)W_{i,j}$. To compute $\hat{\chi}(\tau)$, first let $K(\cdot)$ be a symmetric kernel function, h be a bandwidth, and $\hat{\underline{\varepsilon}}_{i,j}(\tau)$ be some approximation of $\varepsilon_{i,j}(\tau)$, which will be specified later,

$$\hat{\chi}(\tau) = \left(\sum_{j \in J} \xi_j \hat{\mathcal{J}}_{\gamma,\gamma,j}(\tau)\right)^{-1} \left(\sum_{j \in J} \xi_j \hat{\underline{\mathcal{J}}}_{\gamma,\theta,j}(\tau)\right), \tag{21}$$

$$\hat{\mathcal{J}}_{\gamma,\gamma,j}(\tau) = \mathbb{P}_{n,j}\left(\frac{1}{h}K\left(\frac{\hat{\underline{\varepsilon}}_{i,j}(\tau)}{h}\right)V_{i,j}^2(\tau)W_{i,j}W_{i,j}^\top\right), \quad \text{and} \tag{22}$$

$$\hat{\underline{\mathcal{J}}}_{\gamma,\theta,j}(\tau) = \mathbb{P}_{n,j}\left(\frac{1}{h}K\left(\frac{\hat{\underline{\varepsilon}}_{i,j}(\tau)}{h}\right)V_{i,j}^2(\tau)W_{i,j}\hat{\underline{\Phi}}_{i,j}^\top(\tau)\right). \tag{23}$$

Next, we discuss how to obtain $\hat{\underline{\varepsilon}}_{i,j}(\tau)$. First, for the subvector inference, we require the original IVs $\underline{\Phi}_{i,j}(\tau)$ to be valid, that is, Assumptions 5, 6, 8(i)–8(iii) hold with $\Phi_{i,j}(\tau)$ and $\hat{\Phi}_{i,j}(\tau)$ replaced by $\underline{\Phi}_{i,j}(\tau)$ and $\hat{\underline{\Phi}}_{i,j}(\tau)$. Then, we construct $\hat{\underline{\varepsilon}}_{i,j}(\tau)$ as

$$\hat{\underline{\varepsilon}}_{i,j}(\tau) = Y_{i,j} - X_{i,j}^\top\hat{\underline{\beta}}(\tau) - W_{i,j}^\top\hat{\underline{\gamma}}(\tau), \tag{24}$$

where $(\hat{\underline{\beta}}(\tau), \hat{\underline{\gamma}}(\tau))$ are the IVQR estimators obtained via (19) and (20) with $\hat{\Phi}_{i,j}(\tau)$ replaced by $\hat{\underline{\Phi}}_{i,j}(\tau)$. Second, for the full-vector inference, IVs may be weak or invalid, so we define $\hat{\underline{\varepsilon}}_{i,j}(\tau)$ as

$$\hat{\underline{\varepsilon}}_{i,j}(\tau) = Y_{i,j} - X_{i,j}^\top\beta_0(\tau) - W_{i,j}^\top\hat{\gamma}(\beta_0(\tau), \tau), \tag{25}$$

where $\hat{\gamma}(\beta_0(\tau), \tau)$ is defined in (19) and $\beta_0(\tau)$ is the null hypothesis. Then, under the null, we have $\sup_{\tau\in\Upsilon} ||\hat{\gamma}(\beta_0(\tau), \tau) - \gamma_n(\tau)||_2 = O_p(n^{-1/2})$. In Appendix A, we provide regularity conditions under which $\hat{\Phi}_{i,j}(\tau)$ constructed via the full sample projection as described above satisfies Assumption 7(i). We also give another procedure based on the cluster-level projection to construct $\hat{\Phi}_{i,j}(\tau)$ and show that it satisfies Assumption 7(i) too.

3.1 Subvector Inference

In this and the next sections, we consider the following testing problem. Let $\lambda_\beta^\top(\tau)\beta_0(\tau) = \lambda_0(\tau)$ and $\mu(\tau) = \lambda_\beta^\top(\tau)\mu_\beta(\tau)$, for $\tau \in \Upsilon$. Then, define the null and local alternative hypotheses as

$$\mathcal{H}_0 : \mu(\tau) = 0,\ \forall\ \tau \in \Upsilon \quad v.s. \quad \mathcal{H}_{1,n} : \mu(\tau) \neq 0,\ \exists\ \tau \in \Upsilon, \tag{26}$$

where $\lambda_\beta(\tau) \in \mathbf{R}^{d_x \times d_r}$, $\lambda_0(\tau) \in \mathbf{R}^{d_r}$, and Υ is a compact subset of $(0,1)$. In Appendix B, we further study the hypothesis in which each restriction involves two quantile indexes. Consider the test statistic with a $d_r \times d_r$ weighting matrix $\hat{A}_r(\tau)$, and let

$$T_n^{QR} = \sup_{\tau \in \Upsilon} ||\sqrt{n}(\lambda_\beta^\top(\tau)\hat{\beta}(\tau) - \lambda_0(\tau))||_{\hat{A}_r(\tau)} \tag{27}$$

be the test statistic. In the following, we describe the gradient wild bootstrap procedure.

Step 1: We define $\hat{\beta}^r(\tau) = \arg\inf_{b \in \mathcal{B}, \lambda_\beta^\top(\tau)b = \lambda_0(\tau)} ||\hat{\theta}(b,\tau)||_{\hat{A}_\phi(\tau)}$, where $\hat{\theta}(b,\tau)$ is defined in (19). Let $\hat{\gamma}^r(\tau) = \hat{\gamma}(\hat{\beta}^r(\tau), \tau)$.

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

$$\begin{aligned}(\hat{\gamma}_g^*(b,\tau), \hat{\theta}_g^*(b,\tau)) = \arg\inf_{r,t} \Bigg[&\sum_{j \in J} \sum_{i \in I_{n,j}} \rho_\tau(y_{i,j} - X_{i,j}^\top b - W_{i,j}^\top r - \hat{\Phi}_{i,j}^\top(\tau)t)\hat{V}_{i,j}(\tau) \\ &- \sum_{j \in J} g_j \sum_{i \in I_{n,j}} \hat{f}_\tau^\top(D_{i,j}, \hat{\beta}^r(\tau), \hat{\gamma}^r(\tau), 0) \begin{pmatrix} r \\ t \end{pmatrix} \Bigg], \\ \hat{\beta}_g^*(\tau) = \arg\inf_{b \in \mathcal{B}} \left[||\hat{\theta}_g^*(b,\tau)||_{\hat{A}_\phi(\tau)} \right], &\quad \text{and} \quad \hat{\gamma}_g^*(\tau) = \hat{\gamma}_g^*(\hat{\beta}_g^*(\tau), \tau),\end{aligned} \tag{28}$$

where the restricted estimators $(\hat{\beta}^r(\tau), \hat{\gamma}^r(\tau))$ are defined in the previous step.

Step 3: Let $T_n^{QR*}(g) = \sup_{\tau \in \Upsilon} ||\sqrt{n}\lambda_\beta^\top(\tau)(\hat{\beta}_g^*(\tau) - \hat{\beta}(\tau))||_{\hat{A}_{r,g}^*(\tau)}$, where $\hat{A}_{r,g}^*(\tau)$ is the bootstrap counterpart of the weighting matrix $\hat{A}_r(\tau)$. Let $\hat{c}_n^{QR}(1-\alpha)$ denote the $1-\alpha$ quantile of $\{T_n^{QR*}(g)\}_{g \in \mathbf{G}}$, and we reject the null hypothesis if $T_n^{QR} > \hat{c}_n^{QR}(1-\alpha)$.

Denote $\mathcal{J}_{\pi,\pi}(\tau) = \sum_{j \in J} \xi_j \mathcal{J}_{\pi,\pi,j}(\tau)$, $\mathcal{J}_{\pi,\beta}(\tau) = \sum_{j \in J} \xi_j \mathcal{J}_{\pi,\beta,j}(\tau)$, $\overline{\mathcal{J}}_{\pi,\pi}(\tau) = \mathcal{J}_{\pi,\pi}^{-1}(\tau)$, $\mathcal{J}_{\theta,\beta,j}(\tau) = \omega\mathcal{J}_{\pi,\beta,j}(\tau)$, $\mathcal{J}_{\theta,\beta}(\tau) = \sum_{j \in J} \xi_j \mathcal{J}_{\theta,\beta,j}(\tau)$, and $\overline{\mathcal{J}}_\theta(\tau) = \omega\overline{\mathcal{J}}_{\pi,\pi}(\tau)$, where $\omega = (0_{d_\phi \times d_w}, \mathbb{I}_{d_\phi})$, $0_{d_\phi \times d_w}$ is a $d_\phi \times d_w$ matrix of zeros, and $\mathbb{I}_{d_\phi}$ is a $d_\phi \times d_\phi$ identity matrix.

We discuss two cases for the weighting matrix $\hat{A}_r(\tau)$: (1) it has a deterministic limit and (2) it is the inverse of the CCE. For case (1), we do not bootstrap the weighting matrix and let

$\hat{A}^*_{r,g}(\tau) = \hat{A}_r(\tau)$. By abuse of notation, the corresponding test and bootstrap statistics and the critical value are still denoted as T_n^{QR}, $T_n^{QR*}(g)$, and $\hat{c}_n^{QR}(1-\alpha)$, respectively.

For case (2), in order to formally define the weighting matrix, we need extra notation. Let $\hat{\varepsilon}_{i,j}(\tau) = y_{i,j} - X_{i,j}^\top \hat{\beta}(\tau) - W_{i,j}^\top \hat{\gamma}(\tau)$,

$$\hat{\mathcal{J}}_{\pi,\beta,j}(\tau) = \mathbb{P}_{n,j}\left(\frac{1}{h}K\left(\frac{\hat{\varepsilon}_{i,j}(\tau)}{h}\right)\hat{\Psi}_{i,j}(\tau)X_{i,j}^\top\right), \quad \hat{\mathcal{J}}_{\pi,\pi,j}(\tau) = \mathbb{P}_{n,j}\left(\frac{1}{h}K\left(\frac{\hat{\varepsilon}_{i,j}(\tau)}{h}\right)\hat{\Psi}_{i,j}(\tau)\hat{\Psi}_{i,j}^\top(\tau)\right),$$

$$\hat{\mathcal{J}}_{\pi,\beta}(\tau) = \sum_{j\in J}\xi_j \hat{\mathcal{J}}_{\pi,\beta,j}(\tau), \quad \hat{\mathcal{J}}_{\pi,\pi}(\tau) = \sum_{j\in J}\xi_j \hat{\mathcal{J}}_{\pi,\pi,j}(\tau), \tag{29}$$

$$\hat{\Omega}(\tau) = \left[\hat{\mathcal{J}}_{\pi,\beta}^\top(\tau)\hat{\mathcal{J}}_{\pi,\pi}^{-1}(\tau)\omega^\top \hat{A}_\phi(\tau)\omega\hat{\mathcal{J}}_{\pi,\pi}^{-1}(\tau)\hat{\mathcal{J}}_{\pi,\beta}(\tau)\right]^{-1}\hat{\mathcal{J}}_{\pi,\beta}^\top(\tau)\hat{\mathcal{J}}_{\pi,\pi}^{-1}(\tau)\omega^\top \hat{A}_\phi(\tau)\omega\hat{\mathcal{J}}_{\pi,\pi}^{-1}(\tau), \quad \text{and}$$

$$\hat{V}(\tau,\tau') = \frac{1}{n}\sum_{j\in J}\left[\sum_{i\in I_{n,j}}((\tau - 1\{\hat{\varepsilon}_{i,j}(\tau)\le 0\})\hat{\Psi}_{i,j}(\tau))\right]\left[\sum_{i\in I_{n,j}}((\tau' - 1\{\hat{\varepsilon}_{i,j}(\tau')\le 0\})\hat{\Psi}_{i,j}(\tau))\right]^\top, \tag{30}$$

where $K(\cdot)$ and h are the kernel function and bandwidth as defined above. Then, we denote $\hat{A}_r(\tau)$ as $\hat{A}_{r,CR}(\tau)$ and define it as

$$\hat{A}_{r,CR}(\tau) = \left[\lambda_\beta^\top(\tau)\hat{\Omega}(\tau)\hat{V}(\tau,\tau)\hat{\Omega}^\top(\tau)\lambda_\beta(\tau)\right]^{-1}. \tag{31}$$

The corresponding CCE-weighted Wald test statistic is defined as

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

As q, the number of clusters, is fixed, we will show that $\hat{A}_{r,CR}(\tau)$ converges to a random limit. In addition, we use the estimators $\hat{\beta}$ and $\hat{\gamma}$ to construct the cluster-robust covariance matrix, and this introduces extra randomness. Therefore, we need to design a bootstrap counterpart of $\hat{A}_{r,CR}(\tau)$ to mimic such randomness. Let

$$\overline{f}^*_{\tau,g}(D_{i,j}) = g_j\hat{f}_\tau(D_{i,j},\hat{\beta}^r(\tau),\hat{\gamma}^r(\tau),0) + \hat{f}_\tau(D_{i,j},\hat{\beta}^*_g(\tau),\hat{\gamma}^*_g(\tau),0) - \hat{f}_\tau(D_{i,j},\hat{\beta}(\tau),\hat{\gamma}(\tau),0),$$

$$\hat{V}^*_g(\tau,\tau') = \frac{1}{n}\sum_{j\in J}\left\{\sum_{i\in I_{n,j}}\overline{f}^*_{\tau,g}(D_{i,j})\right\}\left\{\sum_{i\in I_{n,j}}\overline{f}^*_{\tau,g}(D_{i,j})\right\}^\top. \tag{32}$$

We let $\hat{A}^*_{r,g}(\tau)$ in the previous algorithm equal $\hat{A}^*_{r,CR,g}(\tau)$ which is defined as

$$\hat{A}^*_{r,CR,g}(\tau) = \left[\lambda_\beta^\top(\tau)\hat{\Omega}(\tau)\hat{V}^*_g(\tau,\tau)\hat{\Omega}^\top(\tau)\lambda_\beta(\tau)\right]^{-1},$$

where $\hat{\beta}^*_g(\tau)$ and $\hat{\gamma}^*_g(\tau)$ are defined in (28). The bootstrap counterpart of $T_{CR,n}^{QR}$ and critical value are defined as

$$T^{QR*}_{CR,n}(g) = \sup_{\tau\in\Upsilon} ||\sqrt{n}(\lambda_\beta^\top(\tau)(\hat{\beta}^*_g(\tau) - \hat{\beta}(\tau)))||_{\hat{A}^*_{r,CR,g}(\tau)} \quad \text{and} \quad \hat{c}^{QR}_{CR,n}(1-\alpha), \quad \text{respectively.}$$

Assumption 8. *(i) There are compact subsets $\mathcal{R}$ and Θ of $\boldsymbol{R}^{d_w}$ and $\boldsymbol{R}^{d_\phi}$, respectively, such that Jacobian matrix $\frac{\partial}{\partial(b^\top,r^\top)}\Pi(b,r,0,\tau)$ is continuous and has full column rank, uniformly in n and over $\mathcal{B}\times\mathcal{R}\times\Theta\times\Upsilon$.*[14]

(ii) The image of $\mathcal{B}\times\mathcal{R}$ under the mapping $(b,r)\mapsto\Pi(b,r,0,\tau)$ is simply connected.

(iii) Suppose $\sup_{\tau\in\Upsilon}||\hat{A}_\phi(\tau)-A_\phi(\tau)||_{op} = o_p(1)$, where $A_\phi(\tau)$ is a symmetric $d_\phi\times d_\phi$ deterministic matrix such that $0 < c \le \inf_{\tau\in\Upsilon}\lambda_{\min}(A_\phi(\tau)) \le \sup_{\tau\in\Upsilon}\lambda_{\max}(A_\phi(\tau)) \le C < \infty$, and

$$\begin{aligned} 0 < c \le& \inf_{\tau\in\Upsilon}\lambda_{\min}\left(\mathcal{J}^\top_{\pi,\beta}(\tau)\overline{\mathcal{J}}^\top_\theta(\tau)A_\phi(\tau)\overline{\mathcal{J}}_\theta(\tau)\mathcal{J}_{\pi,\beta}(\tau)\right) \\ \le& \sup_{\tau\in\Upsilon}\lambda_{\max}\left(\mathcal{J}^\top_{\pi,\beta}(\tau)\overline{\mathcal{J}}^\top_\theta(\tau)A_\phi(\tau)\overline{\mathcal{J}}_\theta(\tau)\mathcal{J}_{\pi,\beta}(\tau)\right) \le C < \infty \end{aligned}$$

for some constants c, C.

(iv) Suppose $\sup_{\tau\in\Upsilon}||\hat{A}_r(\tau)-A_r(\tau)||_{op} = o_p(1)$, where $A_r(\tau)$ is a symmetric $d_r\times d_r$ deterministic weighting matrix such that $0 < c \le \inf_{\tau\in\Upsilon}\lambda_{\min}(A_r(\tau)) \le \sup_{\tau\in\Upsilon}\lambda_{\max}(A_r(\tau)) \le C < \infty$, for some constants c, C.

Assumptions 8(i) and 8(ii) are Assumptions R5* and R6* in Chernozhukov and Hansen (2008a). They, along with Assumption 5, imply that $\beta_n(\tau)$ is uniquely defined. Second, by Chernozhukov and Hansen (2006, Theorem 2), under Assumptions 5 and 8(i)–8(iii), $(\beta_n(\tau),\gamma_n(\tau))$ uniquely solves the system of equations $\mathbb{E}(\tau - 1\{y_{i,j}\le X^\top_{i,j}b + W^\top_{i,j}r\})\Psi_{i,j}(\tau) = 0$. Third, Assumption 8(iii) implies $\mathcal{J}_{\pi,\beta}(\tau)$ is of full column rank, and thus, $\beta_n(\tau)$ is strongly identified. It requires $d_\phi \ge d_x$, which means the number of instruments is no less than the number of endogenous regressors.

Assumption 9. *One of the following two conditions holds: (i) $d_x = 1$ and (ii) there exist scalars $\{a_j(\tau)\}_{j\in J}$ such that $\mathcal{J}_{\theta,\beta,j}(\tau) = a_j(\tau)\mathcal{J}_{\theta,\beta}(\tau)$.*

Assumption 9 is the same as Assumption 3 to which we refer readers for more discussion. In particular, it allows for the presence of weak clusters. Define $a_j(\tau) = \Omega(\tau)\mathcal{J}_{\pi,\beta,j}(\tau)$ when $d_x = 1$, where

$$\Omega(\tau) = \left[\mathcal{J}^\top_{\pi,\beta}(\tau)\overline{\mathcal{J}}^\top_\theta(\tau)A_\phi(\tau)\overline{\mathcal{J}}_\theta(\tau)\mathcal{J}_{\pi,\beta}(\tau)\right]^{-1}\mathcal{J}^\top_{\pi,\beta}(\tau)\overline{\mathcal{J}}^\top_\theta(\tau)A_\phi(\tau)\overline{\mathcal{J}}_\theta(\tau). \qquad (33)$$

[14] For a sequence of matrices $A_n(v)$ indexed by $v\in\mathcal{V}$ and n, we say that $A_n(v)$ is of full column rank uniformly over $v\in\mathcal{V}$ and n if $\inf_{v\in\mathcal{V},n\to\infty}\lambda_{\min}(A^\top_n(v)A_n(v)) \ge \underline{c} > 0$, for some constant $\underline{c}$.

Assumption 10. *Suppose there exists a subset J_s of J such that $\inf_{j\in J_s,\tau\in\Upsilon} a_j(\tau) \geq c_0 > 0$ and $a_j(\tau) = 0$ for $(j,\tau) \in J\backslash J_s \times \Upsilon$. Further denote $q_s = |J_s|$ and $c_\mu = \sup_{\tau\in\Upsilon} ||\mu(\tau)||_2 / \inf_{\tau\in\Upsilon} ||\mu(\tau)||_2$. Then, we have*

$$\inf_{\tau\in\Upsilon} \lambda_{\min}(A_r(\tau)) > (1 - 2\min_{j\in J} \xi_j c_0) \sup_{\tau\in\Upsilon} \lambda_{\max}(A_r(\tau))c_\mu \quad \textit{and} \quad \lceil |\boldsymbol{G}|(1-\alpha)\rceil \leq |\boldsymbol{G}| - 2^{q-q_s+1}.$$

Remark 13. As $c_0 > 0$, the inequality in Assumption 10 holds automatically if we consider a constant local alternative such that $\mu(\tau) = \mu$ for $\tau \in \Upsilon$ and choose $A_r(\tau)$ to be the $d_r \times d_r$ identity matrix.

Theorem 3.1. *Suppose Assumptions 5-9 hold and*

$$\sup_{\tau\in\Upsilon} (||\lambda_\beta(\tau)||_2 + ||\lambda_0(\tau)||_2) \leq C < \infty.$$

Then under $\mathcal{H}_0$, that is, $\mu(\tau) = 0$ for $\tau \in \Upsilon$,

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

In addition, if Assumption 10 holds, then under $\mathcal{H}_{1,n}$,

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

Theorem 3.1 parallels Theorem 2.1. It shows that the gradient wild bootstrap controls size asymptotically when at least one of the clusters is strong and has power against $n^{-1/2}$-local alternatives if further Assumption 10 holds. We conjecture that this procedure is also valid with a large number of clusters when $\beta_n(\tau)$ is strongly identified, under regulatory conditions similar to those in Hagemann (2017). In Appendix B, we further consider the wild bootstrap inference for the case when the null hypothesis involves two quantile indexes.

For the subvector inference of IVQR with the CCE, we need the following assumption.

Assumption 11. *Suppose (i) $q > d_r$ and (ii)*

$$\sup_{\tau\in\Upsilon, j\in J} ||\hat{\mathcal{J}}_{\pi,\beta,j}(\tau) - \mathcal{J}_{\pi,\beta,j}(\tau)||_{op} = o_p(1) \quad \textit{and} \quad \sup_{\tau\in\Upsilon, j\in J} ||\hat{\mathcal{J}}_{\pi,\pi,j}(\tau) - \mathcal{J}_{\pi,\pi,j}(\tau)||_{op} = o_p(1).$$

Remark 14. Assumption 11(i) guarantees that the CCE $\hat{A}_{r,CR}(\tau)$ is invertible and the corresponding test statistic $T_{CR,n}^{QR}$ does not degenerate. As $\hat{\mathcal{J}}_{\pi,\beta,j}(\tau)$ and $\hat{\mathcal{J}}_{\pi,\pi,j}(\tau)$ are just sample analogues of $\mathcal{J}_{\pi,\beta,j}(\tau)$ and $\mathcal{J}_{\pi,\pi,j}(\tau)$, respectively, it is expected that Assumption 11(ii) holds when the dependence of the observations within each cluster is not too strong so that some version of the uniform weak law of large numbers still holds.

Theorem 3.2. *Suppose Assumptions 5–7, 8(i)–8(iii), 9, and 11 hold and*

$$\sup_{\tau\in\Upsilon} (||\lambda_\beta(\tau)||_2 + ||\lambda_0(\tau)||_2) \leq C < \infty.$$

Then under $\mathcal{H}_0$*, that is,* $\mu(\tau) = 0$ *for* $\tau \in \Upsilon$*,*

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

In addition, suppose there exists a subset J_s *of* J *such that* $\inf_{j\in J_s,\tau\in\Upsilon} |a_j(\tau)| \geq c_0 > 0$, $a_j(\tau) = 0$ *for* $(j,\tau) \in J\backslash J_s \times \Upsilon$*, and* $\lceil |\boldsymbol{G}|(1-\alpha)\rceil \leq |\boldsymbol{G}| - 2^{q-q_s+1}$*, where* $q_s = |J_s|$*. Then under* $\mathcal{H}_{1,n}$*,*

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

Theorem 3.2 parallels Theorem 2.2 and Remark 7 still applies. In particular, the $T_{CR,n}^{QR}$ test has power advantage over the T_n^{QR} test.

3.2 Full-Vector Inference

In this section, we consider the full-vector inference for $\beta_n(\tau)$ when it may be weakly or partially identified. Recall $\beta_n(\tau) = \beta_0(\tau) + \mu_\beta(\tau)/\sqrt{n}$. Under the null, we have $\mu_\beta(\tau) = 0$, or equivalently, $\beta_n(\tau) = \beta_0(\tau)$. Our test follows the construction by Chernozhukov and Hansen (2008a). Let

$$AR_n^{QR} = \sup_{\tau\in\Upsilon} ||\hat{\theta}(\beta_0(\tau),\tau)||_{\hat{A}_\phi(\tau)},$$

where $\hat{\theta}(b,\tau)$ is defined in (19) and $\hat{A}_\phi(\tau)$ is a $d_\phi \times d_\phi$ weighting matrix, which will be specified later. Next, the bootstrap procedure for the full-vector inference is defined as follows.

Step 1: Recall $\mathbf{G} = \{-1,1\}^q$ and for any $g = (g_1,\cdots,g_q) \in \mathbf{G}$, let

$$(\tilde{\gamma}_g^*(\tau), \tilde{\theta}_g^*(\tau)) = \arg\inf_{r,t} \Bigg[\sum_{j\in J} \sum_{i\in I_{n,j}} \rho_\tau(y_{i,j} - X_{i,j}^\top \beta_0(\tau) - W_{i,j}^\top r - \hat{\Phi}_{i,j}^\top(\tau)t)\hat{V}_{i,j}(\tau) \\ - \sum_{j\in J} g_j \sum_{i\in I_{n,j}} \hat{f}_\tau^\top(D_{i,j}, \beta_0(\tau), \hat{\gamma}(\beta_0(\tau),\tau), 0) \begin{pmatrix} r \\ t \end{pmatrix} \Bigg], \quad (34)$$

where we impose the null when computing the bootstrap estimator.

Step 2: The bootstrap test statistic is then defined as $AR_n^{QR*}(g) = \sup_{\tau\in\Upsilon} ||\tilde{\theta}_g^*(\tau) - \hat{\theta}(\beta_0(\tau),\tau)||_{\hat{A}_\phi(\tau)}$.

Step 3: Let $\hat{c}_{AR,n}^{QR}(1-\alpha)$ denote the $1-\alpha$ quantile of $\{AR_n^{QR*}(g)\}_{g\in\mathbf{G}}$, and we reject the null hypothesis when $AR_n^{QR} > \hat{c}_{AR,n}^{QR}(1-\alpha)$.

It is also possible to studentize $\hat{\theta}(\beta_0(\tau),\tau)$ by the CCE (i.e., let $\hat{A}_\phi(\tau) = \hat{A}_{CR}(\tau)$ defined in Assumption 12 below). Define the corresponding test statistic, its bootstrap counterpart, and critical value as

$$AR_{CR,n}^{QR} = \sup_{\tau\in\Upsilon} ||\hat{\theta}(\beta_0(\tau),\tau)||_{\hat{A}_{CR}(\tau)}, \quad AR_{CR,n}^{QR*}(g) = \sup_{\tau\in\Upsilon} ||\hat{\theta}_g^*(\tau) - \hat{\theta}(\beta_0(\tau),\tau)||_{\hat{A}_{CR}(\tau)},$$

and $\hat{c}_{AR,CR,n}^{QR}(1-\alpha)$, respectively. We reject the null hypothesis when $AR_{CR,n}^{QR} > \hat{c}_{AR,CR,n}^{QR}(1-\alpha)$.

Assumption 12. *Suppose one of the conditions below holds.*

(i) There exists a symmetric $d_\phi \times d_\phi$ *matrix* $A_\phi(\tau)$ *such that* $\sup_{\tau\in\Upsilon}||\hat{A}_\phi(\tau) - A_\phi(\tau)||_{op} = o_p(1)$ *and for some constants* c *and* C,

$$0 < c \leq \inf_{\tau\in\Upsilon} \lambda_{\min}(A_\phi(\tau)) \leq \sup_{\tau\in\Upsilon} \lambda_{\max}(A_\phi(\tau)) \leq C < \infty.$$

(ii) Suppose $\hat{A}_{CR}(\tau) = \left[\omega\hat{\mathcal{J}}_{\pi,\pi}^{-1}(\tau)\tilde{V}(\tau,\tau)\hat{\mathcal{J}}_{\pi,\pi}^{-1}(\tau)\omega^\top\right]^{-1}$, *where* $\omega = (0_{d_\phi\times d_w}, \mathbb{I}_{d_\phi})$, $\hat{\mathcal{J}}_{\pi,\pi}(\tau)$ *is defined in* (29), $\sup_{\tau\in\Upsilon}||\hat{\mathcal{J}}_{\pi,\pi}(\tau) - \mathcal{J}_{\pi,\pi}(\tau)||_{op} = o_p(1)$,

$$\tilde{V}(\tau,\tau) = \frac{1}{n}\sum_{j\in J}\left[\sum_{i\in I_{n,j}}((\tau - 1\{\hat{\underline{\varepsilon}}_{i,j}(\tau) \leq 0\})\hat{\Psi}_{i,j}(\tau))\right]\left[\sum_{i\in I_{n,j}}((\tau - 1\{\hat{\underline{\varepsilon}}_{i,j}(\tau) \leq 0\})\hat{\Psi}_{i,j}(\tau))\right]^\top, \tag{35}$$

where $\hat{\underline{\varepsilon}}_{i,j}(\tau)$ *is defined in* (25) *and* $\hat{\gamma}(b,\tau)$ *is defined in* (19). *We further require* $q > d_\phi$.

Theorem 3.3. *Suppose Assumptions 5–7 and 12 hold and* $\beta_n(\tau) = \beta_0(\tau)$ *for* $\tau \in \Upsilon$. *Then,*

$$\alpha - \frac{1}{2^{q-1}} \leq \liminf_{n\to\infty}\mathbb{P}(AR_n^{QR} > \hat{c}_{AR,n}^{QR}(1-\alpha)) \leq \limsup_{n\to\infty}\mathbb{P}(AR_n^{QR} > \hat{c}_{AR,n}^{QR}(1-\alpha)) \leq \alpha + \frac{1}{2^{q-1}}, \text{ and}$$

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

Remark 15. Theorem 3.3 holds without assuming strong identification (i.e., Assumption 8). The asymptotic size of the AR_n^{QR} and $AR_{CR,n}^{QR}$-based bootstrap inference is therefore controlled up to an error 2^{1-q}, even when $\beta_n(\tau)$ is weakly or partially identified. This is consistent with Theorem 2.3 and also in line with the robust inference approach proposed by Chernozhukov and Hansen (2008a) for i.i.d. data, which is based on chi-squared critical values. Also similar to IV regressions (Remark 9), we need to implement a projection for the weak-IV-robust subvector inference, which may be conservative, especially if d_x is large or the hypothesis involves multiple quantile indexes. Instead, we may apply the Wald-type subvector inference methods in Section 3.1 in the main text and Appendix B if we are confident that $\beta_n(\tau)$ is strongly identified in at least one of the clusters.

4 Cluster-Level Variables

In this section, we discuss whether our inference method allows for cluster-level variables, whose values are invariant within each cluster, in $W_{i,j}$, $X_{i,j}$, and $Z_{i,j}$. First, as discussed after Assumption 2 in the main text and Assumption 13 in the Appendix, for both IV and IVQR, we allow for cluster-level covariates in the control variable $W_{i,j}$. If practitioners are concerned about cross-cluster heterogeneity, which may potentially jeopardize Assumptions 2(iv) and 13(iii), they can implement the cluster-level projection as detailed in Appendix A.

Second, for the wild bootstrap Wald tests (T_n, $T_{CR,n}$, T_n^{QR}, and $T_{CR,n}^{QR}$), we cannot allow for cluster-level endogenous variables. Taking the IV regression as an example, if $X_{i,j}$ contains cluster-level variables, then the within-cluster limiting Jacobian matrix $Q_{\tilde{Z}X,j}$ may be random and potentially correlated with the score as $X_{i,j}$ is endogenous. This may lead to the asymmetry of the limit distributions of the IV and IVQR estimators, jeopardizing the validity of our approach. However, the wild bootstrap AR tests (AR_n, $AR_{CR,n}$, $AR_{R,n}$, AR_n^{QR}, and $AR_{CR,n}^{QR}$) remain valid in this case.

Third, if $Z_{i,j}$ contains cluster-level variables, then in general we cannot allow control variables $W_{i,j}$ such as the intercept term.[15] In this case, we have to let both the endogenous variable $X_{i,j}$ and the IV $Z_{i,j}$ contain the intercept term. Then, Assumptions 3 and 9 would be required as $d_x > 1$, but they are not likely to be satisfied in such a setup, implying that our Wald tests are invalid. Instead, our AR tests do not require Assumptions 3 and 9 and still remain valid. Alternatively, we may consider merging clusters following CRS (Section 4.2) and Canay et al. (2021, Section B) so that in the merged clusters, the IV is not invariant.

Fourth, for a linear IV regression, we allow for an unobserved cluster-level effect in the outcome $y_{i,j}$, or equivalently in $\varepsilon_{i,j}$, in the form of $\varepsilon_{i,j} = \eta_j + v_{i,j}$. If $\tilde{Z}_{i,j}$ is obtained via the full-sample projection with the full set of cluster fixed effects as $W_{i,j}$, or the cluster-level projection as detailed in Appendix A, we have $\sum_{i \in I_{n,j}} \tilde{Z}_{i,j}\eta_j = 0$ and

$$\frac{1}{\sqrt{n_j}} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j}\varepsilon_{i,j} = \frac{1}{\sqrt{n_j}} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} v_{i,j}.$$

Therefore, Assumption 2(i) is expected to hold. If $\tilde{Z}_{i,j}$ is obtained via the full-sample projection without controlling for the cluster fixed effects, we have

$$\frac{1}{\sqrt{n_j}} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j}\varepsilon_{i,j} = \frac{\eta_j}{\sqrt{n_j}} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} + \frac{1}{\sqrt{n_j}} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} v_{i,j}.$$

Assumption 2 is violated unconditionally as η_j is random. However, all of the assumptions and proofs in the paper can be modified to condition on $\{\eta_j\}_{j \in J}$, and the validity of our wild bootstrap inference for the IV regression still holds. Specifically, here we need $\frac{1}{\sqrt{n_j}} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j}$ and $\frac{1}{\sqrt{n_j}} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} v_{i,j}$ to jointly converge to some normal distribution, conditionally on η_j.

[15]Similar to Canay et al. (2021, Remark 2.3), if $Z_{i,j}$ contains cluster-level variables and the control variables include the intercept term, then Assumption 2(iv) in the main text and Assumption 13(iii) in the Appendix are violated. For example, consider the IV regression when $Z_{i,j} = Z_j$ is a scalar. Then, $\hat{\Gamma}_n = (\sum_{j \in J} \xi_j B_j)^{-1}(\sum_{j \in J} \xi_j C_j Z_j)$, $\hat{\Gamma}^c_{n,j} = B_j^{-1} C_j Z_j$, where $B_j = \frac{1}{n_j} \sum_{i \in I_{n,j}} W_{i,j} W_{i,j}^\top$ and $C_j = \frac{1}{n_j} \sum_{i \in I_{n,j}} W_{i,j}$. Suppose that $W_{i,j}$ is the intercept term, then $B_j = C_j = 1$, and Assumption 2(iv) implies $Z_j - \sum_{\tilde{j} \in J} \xi_{\tilde{j}} Z_{\tilde{j}} \xrightarrow{p} 0$ for $j \in J$, or equivalently, $\{Z_j\}_{j \in J}$ are the same asymptotically. In general, we will have such issue if the marginal distribution of $\{W_{i,j}\}_{i \in I_{n,j}}$ is the same across clusters.

Last, unlike linear IV regression, the cluster-level projection cannot cancel out the cluster-level fixed effect in $\varepsilon_{i,j}(\tau)$ in the form of $\varepsilon_{i,j}(\tau) = \eta_j(\tau) + v_{i,j}(\tau)$ in IVQR. In such a nonlinear model, to account for the cluster-level fixed effect, in general, one needs both the cluster size and the number of clusters to be large. See, for example, Kato, Galvao, and Montes-Rojas (2012), Galvao and Kato (2016), Chetverikov, Larsen, and Palmer (2016), and Galvao, Gu, and Volgushev (2020). In contrast, we focus on the case with few clusters. However, it is still possible to use our bootstrap inference for IVQR with cluster-level fixed effects for some specifications, as illustrated in the following example. Suppose

$$y_{i,j} = X_{i,j}^\top \beta + W_{i,j}^\top \gamma + \eta_j + v_{i,j},$$

and define $\dot{u}_{i,j} = u_{i,j} - \frac{1}{n_j}\sum_{i \in I_{n,j}} u_{i,j}$ for $u \in \{y, X, W, v\}$. Then, we have

$$\dot{y}_{i,j} = \dot{X}_{i,j}^\top \beta + \dot{W}_{i,j}^\top \gamma + \dot{v}_{i,j}.$$

Our bootstrap inference is valid if we assume $\mathbb{P}(\dot{v}_{i,j} \leq 0 | W_{i,j}, Z_{i,j}) = \tau$.[16]

5 Monte Carlo Simulation

5.1 Linear IV Regression

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 Canay et al. (2021) and extend theirs to the IV model. The data are generated as

$$y_{i,j} = \gamma + X_{i,j}\beta + \sigma(Z_{i,j})\left(a_{\varepsilon,j} + \varepsilon_{i,j}\right), \quad X_{i,j} = \gamma + Z_{i,j}^\top \Pi + \sigma(Z_{i,j})\left(a_{v,j} + v_{i,j}\right), \tag{36}$$

for $i = 1, ..., n$ and $j = 1, ..., q$. The total sample size n equals 500, the number of clusters q equals 10, and the cluster size n_j is set to be the same.[17] The disturbances $(\varepsilon_{i,j}, v_{i,j})$ and cluster effects $(a_{\varepsilon,j}, a_{v,j})$ are specified as follows: $(\varepsilon_{i,j}, u_{i,j})^\top \sim N(0, I_2)$, $v_{i,j} = \rho\varepsilon_{i,j} + (1-\rho^2)^{1/2}u_{i,j}$, $(a_{\varepsilon,j}, a_{u,j})^\top \sim N(0, I_2)$, $a_{v,j} = \rho a_{\varepsilon,j} + (1-\rho^2)^{1/2}a_{u,j}$. $\rho \in \{0.2, 0.5, 0.8\}$ correspond to the degree of endogeneity. The IVs are generated by $Z_{i,j} \sim N(0, I_{d_z})$ and $\sigma(Z_{i,j}) = (\sum_{k=1}^{d_z} Z_{i,j,k})^2$, where $Z_{i,j,k}$ denotes the k-th element of $Z_{i,j}$. The IV strength is characterized by $\Pi = (\Pi_0/\sqrt{d_z}, ..., \Pi_0/\sqrt{d_z})^\top$, with $\Pi_0 \in \{2, 1, 1/2, 1/4, 1/8\}$, and the number of IVs equals $d_z \in \{1, 3, 5\}$. 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, and we estimate β using TSLS or LIML with cluster fixed effects included. For T_n and AR_n,

[16] Such condition holds when, for example, $\tau = 0.5$ and $\{v_{i,j}\}_{i \in I_{n,j}}$ are jointly normally distributed with mean zero conditionally on $\{Z_{i,j}, W_{i,j}\}_{i \in I_{n,j}}$ so that $\mathbb{P}(\dot{v}_{i,j} \leq 0|W_{i,j}, Z_{i,j}) = \mathbb{E}(\mathbb{P}(\dot{v}_{i,j} \leq 0|\{W_{i,j}, Z_{i,j}\}_{i \in I_{n,j}})|W_{i,j}, Z_{i,j}) = 0.5$.

[17] We also did simulations with heterogeneity in cluster size and cluster-level IV strength, and the patterns are similar to the simulation results reported here. Results are omitted for brevity but are available upon request.

we set the weighting matrices $\hat{A}_r = 1$ and $\hat{A}_z = I_{d_z}$, where I_{d_z} is the $d_z \times d_z$ identity matrix.

		One IV					Three IVs					Five IVs				
	$\rho \backslash \Pi_0$	2	1	1/2	1/4	1/8	2	1	1/2	1/4	1/8	2	1	1/2	1/4	1/8
T_n^{TSLS}	0.2	10.0	9.8	10.2	9.9	9.6	10.3	10.5	10.8	10.0	10.6	10.6	10.8	11.3	11.2	11.0
	0.5	10.1	9.5	10.7	9.8	10.4	9.4	11.7	12.5	12.7	14.2	11.0	14.1	17.0	17.4	17.9
	0.8	10.5	10.0	9.7	9.5	10.2	9.3	11.5	17.0	18.6	19.3	12.5	22.8	29.0	32.4	31.7
T_n^{LIML}	0.2	10.0	9.8	10.2	9.9	9.6	10.5	10.7	10.1	10.2	10.2	9.9	10.2	10.1	10.4	10.9
	0.5	10.1	9.5	10.7	9.8	10.4	9.6	10.7	10.6	10.0	11.0	9.8	9.9	10.1	11.1	11.5
	0.8	10.5	10.0	9.7	9.5	10.2	9.8	9.7	11.7	11.9	12.0	8.7	11.2	12.4	12.8	12.4
$T_{CR,n}^{TSLS}$	0.2	10.0	9.9	10.4	11.3	11.9	10.6	11.4	12.5	11.7	11.9	11.2	11.4	12.7	12.5	12.6
	0.5	9.8	9.4	10.9	15.5	20.4	9.5	16.1	22.5	23.5	26.1	14.2	21.5	27.5	28.7	29.2
	0.8	10.5	9.9	9.9	21.4	37.7	10.8	21.7	43.0	54.0	57.4	20.7	45.2	59.4	65.4	66.5
$T_{CR,n}^{LIML}$	0.2	10.0	9.9	10.4	11.3	11.9	10.7	11.5	11.8	12.4	11.1	10.6	10.8	11.5	12.3	12.3
	0.5	9.8	9.4	10.9	15.5	20.4	9.3	14.1	20.3	21.2	24.1	11.4	17.8	23.1	24.3	24.6
	0.8	10.5	9.9	9.9	21.4	37.7	9.7	15.4	33.0	43.8	47.2	12.4	28.6	42.2	47.7	49.6
CRS^{TSLS}	0.2	8.2	6.6	6.2	6.4	6.2	11.9	14.2	14.1	15.9	14.7	16.8	17.8	18.9	19.3	19.6
	0.5	7.5	5.5	9.3	12.7	13.4	23.5	39.7	45.5	46.0	47.2	53.9	61.6	63.0	64.0	64.0
	0.8	9.3	5.7	13.6	25.9	33.1	46.4	79.8	88.1	88.6	89.0	92.5	96.9	97.8	97.6	97.9
CRS^{LIML}	0.2	8.2	6.6	6.2	6.4	6.2	5.7	6.2	6.9	7.1	6.8	6.5	7.0	7.2	6.3	6.9
	0.5	7.5	5.5	9.3	12.7	13.4	7.0	11.6	13.1	13.6	13.8	10.7	12.9	13.3	14.2	13.9
	0.8	9.3	5.7	13.6	25.9	33.1	7.9	22.3	31.8	32.9	35.0	17.6	29.4	32.6	33.9	34.9
IM^{TSLS}	0.2	10.2	10.4	11.2	11.3	11.0	12.7	15.4	15.7	17.7	15.9	17.9	18.8	19.7	19.5	20.4
	0.5	10.1	10.1	14.7	19.6	20.2	25.4	41.6	47.3	48.2	48.9	54.2	62.2	64.0	64.7	64.7
	0.8	12.0	10.2	20.5	34.9	41.7	49.0	81.0	88.7	88.9	89.4	92.5	96.8	97.8	97.6	97.9
IM^{LIML}	0.2	10.2	10.4	11.2	11.3	11.0	9.6	11.2	11.6	12.0	11.6	11.4	11.5	11.8	11.2	11.7
	0.5	10.1	10.1	14.7	19.6	20.2	11.8	17.6	19.9	20.7	20.6	17.0	19.5	20.2	21.6	21.1
	0.8	12.0	10.2	20.5	34.9	41.7	13.8	31.5	40.8	42.3	43.9	25.0	38.4	41.9	42.9	43.5
BCH^{TSLS}	0.2	9.7	8.1	5.5	2.9	2.2	7.5	5.5	4.9	4.2	4.4	7.2	6.6	6.7	6.9	7.1
	0.5	9.0	8.6	8.4	6.2	6.1	8.9	11.4	12.3	12.4	13.2	13.1	17.0	21.4	21.8	21.6
	0.8	10.2	9.9	10.4	14.1	18.7	12.2	22.6	35.0	40.6	43.4	23.9	46.5	56.2	61.4	62.0
BCH^{LIML}	0.2	9.7	8.1	5.5	2.9	2.2	8.3	7.1	7.0	6.4	6.5	11.3	11.8	13.0	13.6	13.6
	0.5	9.0	8.6	8.4	6.2	6.1	9.1	11.8	12.7	13.2	14.1	13.7	19.3	23.7	24.4	23.9
	0.8	10.2	9.9	10.4	14.1	18.7	11.0	18.3	28.6	34.3	36.9	17.1	33.5	44.0	48.4	50.0

Table 1: Null Rejection Probabilities in Percentage of Wald Tests for IV Regressions

Note: T_n^{TSLS}, $T_{CR,n}^{TSLS}$, CRS^{TSLS}, IM^{TSLS}, and BCH^{TSLS} denote the T_n and $T_{CR,n}$-based wild bootstrap tests, CRS tests, IM tests, and BCH tests with TSLS, respectively. T_n^{LIML}, $T_{CR,n}^{LIML}$, CRS^{LIML}, IM^{LIML}, and BCH^{LIML} denote their LIML counterparts.

Tables 1 and 2 study the size properties. Specifically, Table 1 reports the null empirical rejection frequencies of the Wald tests that are based on TSLS and LIML, respectively, including the T_n and $T_{CR,n}$-based wild bootstrap procedures in Section 2.1, the randomization tests of CRS, the group-based t-tests of IM, and the CCE-based t-tests with Bester et al. (2011, hereafter

BCH)'s critical values. The TSLS and LIML-based tests are numerically the same with one IV. We highlight several observations below. First, in line with Remark 3, the T_n-based wild bootstrap tests have null rejection frequencies very close to the nominal level with one IV. Except for this case, size distortions increase for all tests in Table 1 when the IVs become weak, the degree of endogeneity becomes high, or the number of IVs becomes large. Second, CRS and IM's tests with LIML show a substantial improvement over their counterparts with TSLS, perhaps because these tests are based on cluster-level estimates, which could produce serious finite-sample bias when TSLS is employed, especially in the over-identified case. All the other LIML-based procedures, including the bootstrap tests, also show size improvement over their TSLS-based counterparts. Third, overall the wild bootstrap procedures compare favorably with alternatives in terms of size control, and the T_n-based procedure is found to have the smallest size distortions among these procedures no matter when TSLS or LIML is used. In particular, the T_n-based bootstrap tests with LIML have null rejection frequencies close to the nominal level across different settings of IV strength, degree of endogeneity, and number of IVs.

		One IV					Three IVs					Five IVs				
	$\rho\backslash\Pi_0$	2	1	1/2	1/4	1/8	2	1	1/2	1/4	1/8	2	1	1/2	1/4	1/8
AR_n	0.2	10.1	9.8	10.2	10.0	9.7	10.8	10.4	10.3	9.9	10.3	10.7	10.2	9.4	10.0	10.2
	0.5	10.1	9.6	10.7	9.8	10.4	9.9	11.4	10.4	9.9	10.4	10.5	9.2	9.5	9.9	10.4
	0.8	10.5	10.0	9.8	9.6	10.3	11.0	10.7	11.1	9.8	10.2	9.9	10.4	10.2	10.5	10.4
$AR_{CR,n}$	0.2	10.1	9.8	10.2	10.0	9.7	10.5	10.6	10.5	9.4	9.9	10.6	10.7	9.5	9.6	10.0
	0.5	10.1	9.6	10.7	9.8	10.4	10.2	10.2	10.3	9.5	9.2	10.2	10.8	9.8	10.1	10.8
	0.8	10.5	10.0	9.8	9.6	10.3	10.1	10.2	10.2	10.3	9.9	9.9	9.9	10.5	10.8	10.5
$AR_{R,n}$	0.2	9.9	9.7	10.2	9.8	9.6	10.7	10.7	10.9	9.3	10.0	10.7	10.3	9.4	9.7	9.5
	0.5	9.9	9.6	10.8	9.8	10.4	10.0	10.2	10.6	9.1	9.1	10.4	10.6	9.2	10.0	10.9
	0.8	10.4	10.0	9.5	9.7	10.3	10.5	10.5	10.6	10.2	10.1	9.9	10.4	10.7	10.6	10.3
$AR_{CR,n}^{ASY}$	0.2	9.2	9.3	9.7	9.6	9.5	5.1	5.1	5.4	4.2	4.7	0.8	0.6	0.5	0.6	0.6
	0.5	9.1	9.1	10.4	9.4	10.0	5.3	4.9	5.3	4.7	4.3	0.6	0.8	0.6	0.5	0.6
	0.8	10.3	9.5	9.3	9.3	9.7	5.0	5.0	5.1	5.5	5.0	0.5	0.6	0.5	0.7	0.6
$AR_{R,n}^{ASY}$	0.2	15.8	15.1	15.5	15.5	15.1	32.0	31.8	32.1	30.7	31.2	56.0	55.3	54.5	54.7	55.5
	0.5	15.0	15.1	16.6	15.6	16.5	32.2	32.3	30.7	30.6	30.4	54.6	55.1	53.9	54.6	56.5
	0.8	16.2	16.2	15.2	15.4	15.9	32.5	32.4	31.5	30.9	31.0	55.9	55.1	54.6	56.6	55.0

Table 2: Null Rejection Probabilities in Percentage of AR Tests for IV Regressions

Note: AR_n, $AR_{CR,n}$, and $AR_{R,n}$ denote the three wild bootstrap AR tests, and $AR_{CR,n}^{ASY}$ and $AR_{R,n}^{ASY}$ denote the $AR_{CR,n}$ and $AR_{R,n}$-based asymptotic AR tests with chi-squared critical values.

Table 2 reports the null rejection frequencies of AR tests, including the $AR_{CR,n}$ and $AR_{R,n}$-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 three wild bootstrap AR tests in Section

2.2 that are based on AR_n, $AR_{CR,n}$, and $AR_{R,n}$, respectively. We notice that $AR_{CR,n}$-based asymptotic tests control the size but under-reject in the over-identified case.[18] On the other hand, the $AR_{R,n}$-based asymptotic tests tend to have substantial over-rejections. By contrast, the three bootstrap AR tests always have rejection frequencies very close to the nominal level.

Figure 1 compares the power properties of the tests that have good size control across different settings, namely, the T_n-based bootstrap tests, the three bootstrap AR tests, and the $AR_{CR,n}$-based asymptotic AR tests. We also include $T_{CR,n}$-based bootstrap tests for a power comparison with their T_n-based counterparts. We let the number of IVs be 3 and Π_0 be 2 or 1.[19] For the Wald tests, we use both LIML and its modified version proposed by Fuller (1977, hereafter FULL), which has finite moments and is thus less dispersed than LIML.[20] First, we notice that among the AR tests, the AR_n-based bootstrap tests have the highest power, the asymptotic AR tests have the lowest, and the $AR_{CR,n}$ and $AR_{R,n}$-based bootstrap tests have almost the same power. Second, the T_n-based bootstrap tests with LIML have power properties that are very similar to the AR_n-based bootstrap tests. Third, the bootstrap tests with FULL are always more powerful than their LIML counterparts. Fourth, in line with our theory, the $T_{CR,n}$-based bootstrap tests are more powerful than their T_n-based counterparts across different settings and therefore may be preferred when identification is strong.

5.2 IVQR

To examine the performance of the wild bootstrap inference for the IVQR, we consider the following DGP. Let $a_{l,j} = (a_{l,1,j}, \cdots, a_{l,n/q,j})^\top$ for $l = 1, \cdots, K$, where $K \geq 1$ is the number of instruments, and $v_{l,j} = (v_{l,1,j}, \cdots, v_{l,n/q,j})^\top$ for $l = 1, 2$. Then, $\{a_{1,j}, \cdots, a_{K,j}, v_{1,j}, v_{2,j}\}$ are independent across $j \in J$, and each of them follow an $n/q \times 1$ multivariate normal distribution with mean zero and covariance $\Sigma(\rho_j)$, where $\Sigma(\rho_j)$ is an $n/q \times n/q$ toeplitz matrix with coefficient ρ_j and $\rho_j = 0.2 + 0.05j$. Define $u_{1,i,j} = v_{1,i,j}$, $u_{2,i,j} = \rho v_{1,i,j} + \sqrt{1-\rho^2} v_{2,i,j}$,

$$Y_{i,j} = X_{i,j}\beta + \underline{W}_{i,j}\gamma + (0.25 + 0.5\underline{W}_{i,j} + 0.1X_{i,j})u_{1,i,j}, \quad X_{i,j} = 0.6\sqrt{j} + \sum_{k=1}^{K} \Pi_k G(Z_{k,i,j}) + G(u_{2,i,j}),$$

$$Z_{1,i,j} = a_{1,i,j}, Z_{k,i,j} = 0.6a_{1,i,j} + \sqrt{0.64}a_{k,i,j}, \forall k \geq 2,$$

[18]The null rejection probabilities of this AR 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 $\widehat{f} = \left(\widehat{f}_1, ..., \widehat{f}_q\right)^\top$ and $\widehat{f}_j = n^{-1}\sum_{i \in I_{n,j}} f_{i,j}$, $AR_{CR,n} = \iota_q^\top \widehat{f}\left(\widehat{f}^\top \widehat{f}\right)^{-1} \widehat{f}^\top \iota_q = \iota_q^\top \iota_q = d_z$ as long as $\widehat{f}$ 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.

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

[20]Following the recommendation in the literature, we set the tuning parameter of FULL equal to 1, in which case FULL is best unbiased to a second order among k-class estimators when the errors are normal (Rothenberg, 1984).

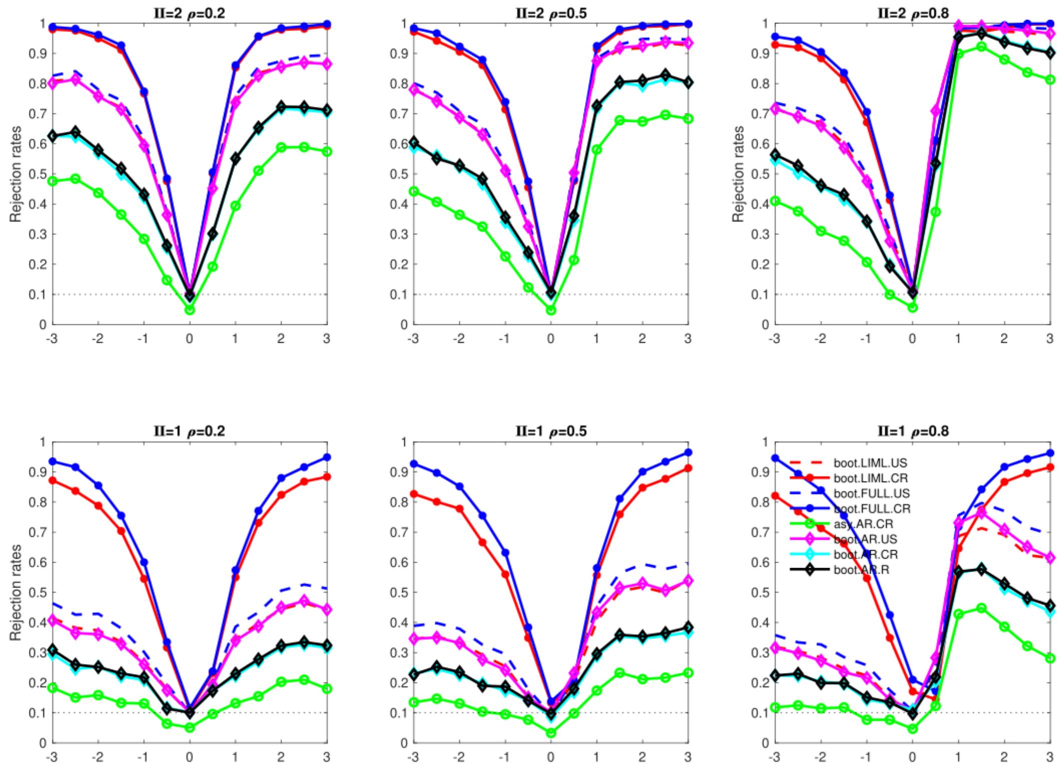

Figure 1: Power of Wild Bootstrap Wald and AR Tests for IV Regressions

Note: "wb.LIML.US," "wb.LIML.CR," "wb.FULL.US," and "wb.FULL.CR" denote the T_n and $T_{CR,n}$-based wild bootstrap tests with LIML and FULL, respectively. "asy.AR.CR" denotes the $AR_{CR,n}$-based AR tests with asymptotic CVs, and "wb.AR.US," "wb.AR.CR," and "wb.AR.R" denote the AR_n, $AR_{CR,n}$ and $AR_{R,n}$-based wild bootstrap AR tests.

where $W_{i,j} = (1, \underline{W}_{i,j})^\top$, $\underline{W}_{i,j}$ is distributed following $\frac{1}{2}\chi_1^2$, and $G(\cdot)$ is the standard normal CDF. We set $\beta = 0.5$, $\gamma = 0$, $\rho = 0.5$, $\Pi_1 = \cdots = \Pi_K = \Pi_0$, $\Pi_0 \in (1, 1/2, 1/4, 1/8)$, $n = 500$, $q = 10$, set the number of bootstrap and simulation replications to be 500, and test the following hypothesis for $K \in \{1, 5\}$ and $\tau \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$:

$$H_0 : \beta(\tau) = 0.5 + 0.1G^{-1}(\tau) \quad v.s. \quad H_1 : \beta(\tau) = 1 + 0.1G^{-1}(\tau).$$

To implement the IVQR, we let the original IV be $\underline{\Phi}_{i,j}(\tau) = \hat{\underline{\Phi}}_{i,j}(\tau) = Z_{i,j}$ and obtain $\hat{\Phi}_{i,j}(\tau)$ as $\hat{\Phi}_{i,j}(\tau) = \hat{\underline{\Phi}}_{i,j}(\tau) - \hat{\chi}^\top(\tau) W_{i,j}$, where $W_{i,j} = (1, \underline{W}_{i,j})^\top$ and $\hat{\chi}(\tau)$ is computed following (21). Specifically, when computing T_n^{QR} and $T_{CR,n}^{QR}$, we first use $Z_{i,j}$ as IV, implement the IVQR, and obtain the preliminary estimator $(\underline{\beta}(\tau), \underline{\gamma}(\tau))$. We then define the residual as

$$\hat{\underline{\varepsilon}}_{i,j} = y_{i,j} - X_{i,j}^\top \underline{\beta}(\tau) - W_{i,j}^\top \underline{\gamma}(\tau). \tag{37}$$

When computing AR_n^{QR} and $AR_{CR,n}^{QR}$, we let $\underline{\beta}(\tau)$ be the value imposed under the null and alternative and compute $\underline{\gamma}(\tau)$ via a linear quantile regression of $Y_{i,j} - X_{i,j}\underline{\beta}(\tau)$ on $(W_{i,j}, Z_{i,j})$. Given the estimator $\underline{\gamma}(\tau)$, we define the residual $\hat{\underline{\varepsilon}}_{i,j}$ as (37). To compute $\hat{\chi}(\tau)$, we use the fourth order Epanechnikov kernel and the rule of thumb bandwidth

$$h = 3.536\hat{s}(q_\tau^4 - 6q_\tau^2 + 3)^{-2/9}n^{-1/5},$$

where q_τ is the τ-th quantile of the standard normal distribution and $\hat{s}$ is the standard error of $\{\hat{\underline{\varepsilon}}_{i,j}\}_{i\in I_j, j\in J}$.[21]

To compute $T_{CR,n}^{QR}$ and AR_n^{QR}, we set the weighting matrix $\hat{A}_\phi(\tau) = I_{d_\phi}$, where I_{d_ϕ} is the $d_\phi \times d_\phi$ identity matrix. To compute T_n^{QR}, we set $\hat{A}_\phi(\tau) = I_{d_\phi}$ and $\hat{A}_r(\tau) = 1$. In addition, to compute the test statistics $T_{CR,n}^{QR}$ and $AR_{CR,n}^{QR}$ with the CCE, we use the same kernel and bandwidth as described above to obtain $\hat{A}_{r,CR}(\tau)$ and $\hat{A}_{CR}(\tau)$ defined in (31) and Assumption 12(ii), respectively. For comparison, we also compute the rejection probabilities for the randomization test of CRS, which is based on cluster-level IVQR estimators.

Table 3 collects the simulation results when the nominal rate of rejection is 10%.[22] We can make five observations. First, both AR_n^{QR} and $AR_{CR,n}^{QR}$ control size regardless of IV strength, while T_n^{QR}, $T_{CR,n}^{QR}$, and CRS have size distortions under weak identification, especially in the case with 5 IVs. Second, $T_{CR,n}^{QR}$ is more powerful than T_n^{QR} under strong identification, which is consistent with our theory. Third, similar to the results for IV regression in Table 2, the power of $AR_{CR,n}^{QR}$ is weaker than that of AR_n^{QR} as the number of IVs increases. Fourth, when there is one endogenous variable, full- and subvector inferences coincide. In this case and under strong identification, AR_n^{QR} is more powerful than $T_{CR,n}^{QR}$ when there is only one IV and less powerful when there are five IVs. Fifth, the power of CRS is similar to AR_n^{QR} and $T_{CR,n}^{QR}$, but it has a larger size distortion under weak identification, multiple IVs, or both.

5.3 Practical Recommendations

For both linear IV and IV quantile regressions, we have the following practical recommendations. First, when conducting the full-vector inference, we suggest using the wild bootstrap AR tests with deterministic studentization (i.e., AR_n and AR_n^{QR}-based tests), which control size irrespective of the strength of the IVs and have a power advantage over the wild bootstrap AR tests with the CCE in the over-identified case. Second, to conduct the subvector inference that is fully robust to weak IVs, one may apply the projection approach after the AR-based full-vector inference. However, projection may result in a power loss with multiple endogenous variables or quantile indexes. Third, for conducting the subvector inference with at least one strong cluster, we may implement the wild bootstrap Wald tests. In particular, those with

[21] This is the optimal bandwidth that minimizes the MSE of the kernel density estimator at q_τ when the underlying density is Gaussian and the data are i.i.d.

[22] The rejection probabilities using AR_n^{QR} and $AR_{CR,n}^{QR}$ are numerically the same with one IV.

		One IV								Five IVs							
		H0				H1				H0				H1			
	$\tau\backslash\Pi_0$	1	1/2	1/4	1/8	1	1/2	1/4	1/8	1	1/2	1/4	1/8	1	1/2	1/4	1/8
AR_n^{QR}	0.1	7.2	5.4	5.8	8.6	53.8	20.2	12.0	6.8	4.2	5.4	6.0	4.8	54.8	17.8	5.6	3.2
	0.25	8.0	5.6	8.4	8.6	76.0	29.0	12.4	7.0	5.2	4.0	5.6	5.4	80.0	22.8	7.6	4.0
	0.5	8.8	9.2	7.0	7.2	84.2	41.4	15.4	9.8	4.8	6.4	5.8	5.6	85.8	29.6	8.6	6.0
	0.75	8.6	8.2	7.8	6.6	80.6	34.8	14.4	9.6	5.0	3.6	4.6	4.2	72.8	19.8	8.6	4.4
	0.9	9.0	9.0	8.4	7.0	59.8	21.4	9.2	4.4	5.4	4.8	6.0	2.6	50.6	15.8	4.0	2.0
$AR_{CR,n}^{QR}$	0.1	7.2	5.4	5.8	8.6	53.8	20.2	12.0	6.8	1.8	1.4	0.6	0.6	21.2	6.6	1.8	1.8
	0.25	8.0	5.6	8.4	8.6	76.0	29.0	12.4	7.0	2.8	2.2	2.8	2.8	11.8	6.8	6.8	2.8
	0.5	8.8	9.2	7.0	7.2	84.2	41.4	15.4	9.8	2.2	3.0	2.6	1.4	16.4	11.8	6.0	2.2
	0.75	8.6	8.2	7.8	6.6	80.6	34.8	14.4	9.6	3.6	2.2	4.0	4.2	9.6	7.6	5.0	3.2
	0.9	9.0	9.0	8.4	7.0	59.8	21.4	9.2	4.4	1.2	1.4	1.8	0.8	13.8	6.2	1.2	1.0
T_n^{QR}	0.1	6.4	4.8	7.8	9.6	46.2	19.4	12.0	11.4	1.6	3.8	5.6	7.4	77.8	38.8	18.0	14.8
	0.25	5.0	7.0	7.2	9.0	53.2	25.4	14.6	14.2	0.4	2.2	5.6	8.4	84.0	45.8	19.6	15.6
	0.5	7.2	8.0	10.4	10.2	51.2	25.2	15.4	12.0	1.2	4.2	9.2	9.8	85.0	46.8	21.6	16.2
	0.75	7.4	8.2	9.2	10.4	50.4	24.0	9.8	13.2	2.2	4.0	9.8	8.0	85.8	47.0	20.4	13.8
	0.9	9.6	9.8	9.2	11.2	43.8	18.2	11.4	8.0	1.2	5.0	8.8	10.4	77.0	31.0	12.4	10.0
$T_{CR,n}^{QR}$	0.1	9.4	8.0	11.0	13.2	52.6	22.2	15.2	16.0	7.0	7.6	10.2	12.8	82.6	45.6	23.2	19.2
	0.25	7.4	8.6	9.0	11.2	56.2	28.0	17.8	16.6	6.6	7.2	9.4	11.6	90.2	51.2	24.0	18.2
	0.5	9.2	11.0	13.2	11.6	56.2	27.4	18.8	14.2	5.8	8.6	11.2	12.6	91.0	52.4	25.0	18.6
	0.75	10.0	9.6	11.6	13.6	57.6	27.6	12.2	15.8	7.8	9.8	14.8	12.6	92.0	53.0	24.4	16.6
	0.9	12.8	13.2	13.2	13.8	52.0	22.4	14.6	11.8	6.4	10.0	13.0	14.8	82.4	37.8	20.0	16.0
CRS	0.1	8.8	11.6	12.6	14.2	64.4	16.8	8.6	10.6	27.2	37.2	34.2	38.2	60.4	14.4	8.0	15.8
	0.25	10.8	10.0	12.6	15.0	82.4	24.2	10.4	12.4	23.2	28.4	35.2	39.4	76.8	16.2	11.2	10.6
	0.5	9.2	11.4	13.2	17.0	84.4	26.4	10.8	8.8	23.6	30.8	33.4	35.8	85.0	21.4	10.4	13.8
	0.75	10.4	10.4	13.4	14.0	79.8	26.0	7.6	10.4	19.6	29.0	33.6	38.6	81.6	16.8	9.6	14.2
	0.9	10.0	11.2	13.6	14.4	66.0	18.4	8.0	11.2	22.2	30.2	39.8	37.8	68.0	12.8	12.8	16.2

Table 3: Rejection Probabilities in Percentage for IVQR

deterministic studentization (i.e., T_n and T_n^{QR}-based tests) have relatively good size control. On the other hand, those with the CCE (i.e., $T_{CR,n}$ and $T_{CR,n}^{QR}$-based tests) have better power properties, and we therefore recommend them when we are confident that the identification is strong. Fourth, in the special case with only one endogenous regressor, the sub- and full-vector inferences coincide. Additionally, AR_n and T_n-based wild bootstrap tests are numerically the same for the linear IV regression with one IV. However, for IVQR, AR_n^{QR} and T_n^{QR} are not numerically the same. We recommend using AR_n^{QR} with one IV, as it has better size and power properties in this case, and using $T_{CR,n}^{QR}$ with multiple IVs when identification is strong.

6 Empirical Application

In an influential study, Autor et al. (2013) analyze the effect of rising Chinese import competition on wages and employment in 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 average and distributional effects of such import exposure by applying IV and IVQR 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.[23]

For both IV and IVQR models, 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.[24] 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 4, with the number of observations (n) and clusters (q) for each region, the TSLS estimates, and the 90% bootstrap confidence sets (CSs) constructed by inverting the corresponding T_n-based (equivalently, 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. The results in Table 4 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). Furthermore, according to the wild bootstrap procedure, the effect on CZs in the South is significantly different from zero at the 10% level, while the effect on CZs in the other two regions is not. Compared with that for the South, the wider CSs for the Midwest and West may be due to relatively weak identification, which our procedure is able to guard against. Table

[23] 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).

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

4 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. Moreover, the effects for both female and male samples in the South are significantly different from zero.

			All		Female		Male	
Region	n	q	Estimate	Bootstrap CS	Estimate	Bootstrap CS	Estimate	Bootstrap CS
South	578	16	-0.97	[-1.73, -0.58]	-0.81	[-1.49, -0.40]	-1.08	[-1.95, -0.66]
Midwest	504	12	-0.025	[-0.68,0.78]	0.024	[-0.63,0.74]	-0.17	[-1.04,0.77]
West	276	11	-1.05	[-1.47,0.26]	-0.6	[-1.49,0.82]	-1.26	[-1.99,0.73]

Table 4: IV regressions of Autor et al. (2013) with all, female, and male samples for three US regions

Next, we study the region-wise distributional effects of Chinese import competition by running IVQR for South, Midwest, and West regions separately. The results are reported in Table 5. The IVQR point estimates are computed using the original IV (Chinese import growth in other high-income countries) without any projection on the exogenous variables. The weak-IV-robust 90% bootstrap CSs are computed by inverting the AR_n^{QR} test when the IV is computed via the projection described in Remark 12 with $\hat{\underline{\varepsilon}}_{i,j}(\tau)$ constructed via (25) so that the null hypothesis is imposed. For the results of Midwest and West regions, some CSs are unbounded or contain several segments as the identification can be rather weak. We follow Chernozhukov and Hansen (2008a) and use the maximum and minimum values admitted by the AR_n^{QR} test as the two endpoints of our CSs. The resulting interval covers all the segments and thus controls coverage asymptotically. We provide the bootstrap p-values of the AR_n^{QR} tests for all our IVQR results in Section S.A in the Online Supplement. In line with the results in IV regressions, the distributional effects of Chinese imports competition are significantly negative for the South region at all quantiles except 90%, while those for the Midwest and West regions are not significant at any of the quantiles. Furthermore, similar to IV regressions, Table 5 shows that the negative effects of import exposure are more substantial for the male samples at most quantiles. In addition, the bootstrap CSs report that the effects are significant for males in the South at the 10%, 50%, and 75% quantiles and significant for females at the 75% and 90% quantiles.

Last, for the South region, whose coefficients are strongly identified, we further test the distributional homogeneity of the effect of Chinese imports exposure on wages by considering the following three null hypotheses:

$$H_0 : \beta(0.1) = \beta(0.5), \quad H_0 : \beta(0.1) = \beta(0.9), \quad \text{and} \quad H_0 : \beta(0.5) = \beta(0.9),$$

where $\beta(\tau)$ is the coefficient of interest and $\tau \in (0, 1)$ is the quantile index. We conduct the Wald-test-based subvector inference for IVQR that involves two quantile indexes.[25] The boot-

[25] We refer readers to Appendix B for its implementation detail and theoretical guarantee.

strap p-values for the three hypotheses are reported in Table 6. We can make three observations. First, the p-values computed based on the Wald test with the CCE (i.e., $T^{QR}_{CR,n}$) are all smaller than those without. This is consistent with our theory that $T^{QR}_{CR,n}$ is more powerful than T^{QR} in detecting more distant alternatives when there is only one restriction. Second, we reject the null hypothesis of $\beta(0.1) = \beta(0.9)$ for female and male samples at 1% and 5% significance levels, respectively. Third, although we cannot reject $\beta(0.1) = \beta(0.9)$ for the full sample, this does not contradict the previous significant results because it is not a combination of the female and male samples. In fact, the samples in Autor et al. (2013) are obtained by averaging data in a given CZ for all, female, and male individuals, respectively, and therefore they share the same sample size.

	Region	South		Midwest		West	
Gender	τ	Estimate	Bootstrap CS	Estimate	Bootstrap CS	Estimate	Bootstrap CS
All	0.1	-0.88	[-2.74,-0.01]	-0.19	[-1.87,0.97]	0.55	[-6.54,9.76]
	0.25	-0.84	[-1.65,-0.16]	-0.44	$(-\infty, 8.00]$	-0.36	[-8.92,2.72]
	0.5	-0.74	[-2.34,-0.42]	-0.24	[-1.61,0.39]	-1.02	$(-\infty, 9.39]$
	0.75	-1.22	[-2.60,-0.39]	-0.20	[-0.95,2.04]	-0.89	[-7.63,0.26]
	0.9	-0.52	[-2.47,0.07]	-0.08	[-1.61,2.57]	0.08	[-9.43,3.07]
Female	0.1	-0.22	[-7.06,0.19]	0.01	[-0.54,9.74]	0.43	$(-\infty, 8.37]$
	0.25	-0.20	[-1.01,0.27]	0.16	$(-\infty, 7.49]$	0.34	[-7.65,2.99]
	0.5	-0.29	[-2.53,0.06]	-0.04	[-2.12,1.46]	-0.44	[-5.43,7.15]
	0.75	-0.69	[-1.76,-0.07]	-0.09	[-1.14,1.22]	-0.36	[-6.20,1.88]
	0.9	-0.54	[-1.48,-0.22]	0.12	[-1.64,1.92]	-1.03	$(-\infty, 1.26]$
Male	0.1	-0.44	[-3.52,-0.13]	-0.24	[-0.87,3.01]	1.61	$[-6.65, \infty)$
	0.25	-0.95	[-2.16,0.08]	-0.48	$(-\infty, 8.10]$	-0.12	[-9.96,3.36]
	0.5	-1.05	[-3.92,-0.73]	-0.28	[-4.10,0.57]	0.14	$(-\infty, 2.54]$
	0.75	-1.26	[-2.15,-0.53]	-0.28	[-0.97,3.30]	-0.63	$(-\infty, 0.85]$
	0.9	-1.07	[-1.87,0.07]	-0.15	[-1.80,2.06]	-2.93	[-9.13,2.57]

Table 5: IVQRs of Autor et al. (2013) with all, female, and male samples for three US regions

	T^{QR}_n			$T^{QR}_{CR,n}$		
(τ_1, τ_2)	$(0.1, 0.5)$	$(0.1, 0.9)$	$(0.5, 0.9)$	$(0.1, 0.5)$	$(0.1, 0.9)$	$(0.5, 0.9)$
All	0.757	0.880	0.778	0.556	0.416	0.334
Female	0.789	0.257	0.554	0.675	0.003	0.272
Male	0.260	0.506	0.945	0.111	0.036	0.894

Table 6: Bootstrap p-values for testing for distributional homogeneity for South region

Appendices

Appendix A Constructing the IVs

In this section, we discuss how to implement projections at both the full-sample and the cluster levels for IV regression and IVQR even when the covariate W contains cluster-level variables.

A.1 IV Regression

Suppose $Z_{i,j}$ is the original IV. We can construct $\tilde{Z}_{i,j}$ as $\tilde{Z}_{i,j} = Z_{i,j} - \hat{\chi}_j^\top W_{i,j}$, where $\hat{\chi}_j = \hat{Q}_{WW,j}\hat{Q}_{WW,j}^{-}\hat{Q}_{WW,j}^{-}\hat{Q}_{WZ,j}$, $\hat{Q}_{WW,j} = \frac{1}{n_j}\sum_{i \in I_{n,j}} W_{i,j}W_{i,j}^\top$, $\hat{Q}_{WW,j} = \frac{1}{n_j}\sum_{i \in I_{n,j}} W_{i,j}Z_{i,j}^\top$, and A^- denotes the pseudo inverse of the positive semidefinite matrix A. We can show that $\frac{1}{n_j}\sum_{i \in I_{n,j}} \tilde{Z}_{i,j}W_{i,j}^\top = 0$ following the same argument in Section A.2.2.

A.2 IVQR

A.2.1 Full-Sample Projection

We first consider the full-sample projection mentioned in the main text.

Assumption 13. *Recall* $\hat{\mathcal{J}}_{\gamma,\gamma,j}$ *and* $\underline{\hat{\mathcal{J}}}_{\gamma,\theta,j}$ *defined in* (22) *and* (23). *Define*

$$\mathcal{J}_{\gamma,\gamma,j}(\tau) = \lim_{n\to\infty} \overline{\mathbb{P}}_{n,j} f_{\varepsilon_{i,j}(\tau)}(0|W_{i,j}, Z_{i,j})V_{i,j}^2(\tau)W_{i,j}W_{i,j}^\top,$$
$$\underline{\mathcal{J}}_{\gamma,\theta,j}(\tau) = \lim_{n\to\infty} \overline{\mathbb{P}}_{n,j} f_{\varepsilon_{i,j}(\tau)}(0|W_{i,j}, Z_{i,j})V_{i,j}^2(\tau)W_{i,j}\underline{\Phi}_{i,j}^\top(\tau),$$
$$\mathcal{J}_{\gamma,\gamma}(\tau) = \sum_{j\in J}\xi_j\mathcal{J}_{\gamma,\gamma,j}(\tau), \quad \underline{\mathcal{J}}_{\gamma,\theta}(\tau) = \sum_{j\in J}\xi_j\underline{\mathcal{J}}_{\gamma,\theta,j}(\tau), \quad \textit{and} \quad \chi(\tau) = \mathcal{J}_{\gamma,\gamma}^{-1}(\tau)\underline{\mathcal{J}}_{\gamma,\theta}(\tau).$$

(i) Suppose

$$\sup_{\tau\in\Upsilon}\left[||\hat{\mathcal{J}}_{\gamma,\gamma,j}(\tau) - \mathcal{J}_{\gamma,\gamma,j}(\tau)||_{op} + ||\underline{\hat{\mathcal{J}}}_{\gamma,\theta,j}(\tau) - \underline{\mathcal{J}}_{\gamma,\theta,j}(\tau)||_{op}\right] = o_p(1).$$

(ii) Recall $\delta_{i,j}(v,\tau) = X_{i,j}^\top v_b + W_{i,j}^\top v_r + \hat{\Phi}_{i,j}^\top(\tau)v_t$. *Then,*

$$\sup\left\|\overline{\mathbb{P}}_{n,j} f_{\varepsilon_{i,j}(\tau)}(\delta_{i,j}(v,\tau)|W_{i,j}, Z_{i,j})V_{i,j}^2(\tau)W_{i,j}\underline{\hat{\Phi}}_{i,j}^\top(\tau) - \underline{\mathcal{J}}_{\gamma,\theta,j}(\tau)\right\|_{op} \xrightarrow{p} 0,$$

where the supremum is taken over $\{j \in J, ||v||_2 \leq \delta, \tau \in \Upsilon\}$, $v = (v_b^\top, v_r^\top, v_t^\top)^\top$.

(iii) For $j \in J$, *there exists* $\chi_{n,j}(\tau)$ *such that*

$$\underline{\mathcal{J}}_{\gamma,\theta,j}(\tau) = \underline{\mathcal{J}}_{\gamma,\gamma,j}(\tau)\chi_{n,j}(\tau) \quad \textit{and} \quad \overline{\mathbb{P}}_{n,j}||W_{i,j}^\top(\chi(\tau) - \chi_{n,j}(\tau))||_{op}^2 = o(1).$$

(iv) There exist constants c, C *such that*

$$0 < c < \inf_{\tau\in\Upsilon}\lambda_{\min}(\mathcal{J}_{\gamma,\gamma}(\tau)) \leq \sup_{\tau\in\Upsilon}\lambda_{\max}(\mathcal{J}_{\gamma,\gamma}(\tau)) \leq C < \infty.$$

First, for the subvector inference, we require the original IVs $\underline{\Phi}_{i,j}(\tau)$ to be valid, i.e., Assumptions 5, 6, 8(i)–8(iii) hold with $\Phi_{i,j}(\tau)$ and $\hat{\Phi}_{i,j}(\tau)$ replaced by $\underline{\Phi}_{i,j}(\tau)$ and $\hat{\underline{\Phi}}_{i,j}(\tau)$. Then, we construct $\hat{\underline{\varepsilon}}_{i,j}(\tau)$ as

$$\hat{\underline{\varepsilon}}_{i,j}(\tau) = Y_{i,j} - X_{i,j}^\top \hat{\underline{\beta}}(\tau) - W_{i,j}^\top \hat{\underline{\gamma}}(\tau),$$

where $(\hat{\underline{\beta}}(\tau), \hat{\underline{\gamma}}(\tau))$ are the IVQR estimators obtained via (19) and (20) with $\hat{\Phi}_{i,j}(\tau)$ replaced by $\hat{\underline{\Phi}}_{i,j}(\tau)$. Note that following the same argument as in Lemma S.K.2, we can show

$$\sup_{\tau \in \Upsilon} \left(||\hat{\underline{\beta}}(\tau) - \beta_n(\tau)||_2 + ||\hat{\underline{\gamma}}(\tau) - \gamma_n(\tau)||_2 \right) = O_p(n^{-1/2}).$$

For the full-vector inference, the IVs may be weak or invalid. Instead, we define $\hat{\underline{\varepsilon}}_{i,j}(\tau)$ as

$$\hat{\underline{\varepsilon}}_{i,j}(\tau) = Y_{i,j} - X_{i,j}^\top \beta_0(\tau) - W_{i,j}^\top \hat{\gamma}(\beta_0(\tau), \tau),$$

where $\hat{\gamma}(\beta_0(\tau), \tau)$ is defined in (19) and $\beta_0(\tau)$ is the null hypothesis. Then, under the null, we have $\sup_{\tau \in \Upsilon} ||\hat{\gamma}(\beta_0(\tau), \tau) - \gamma_n(\tau)||_2 = O_p(n^{-1/2})$. In both cases, $\hat{\underline{\varepsilon}}_{i,j}(\tau)$ is a valid approximation of $\varepsilon_{i,j}(\tau)$ and Assumption 13(i) holds under mild regularity conditions. Second, Assumption 13 requires the nonparametric estimation of $\mathcal{J}_{\gamma,\gamma,j}(\tau)$ and $\underline{\mathcal{J}}_{\gamma,\theta,j}(\tau)$ via kernel smoothing. We discuss the choice of kernel function and bandwidth in Section 5. Third, Assumption 13(ii) holds under mild smoothness conditions. Fourth, Assumption 13(iii) parallels Assumption 2(iv) so that the same comments still apply.

Proposition A.1. *Suppose Assumption 13 holds and $\hat{\Phi}_{i,j}(\tau) = \hat{\underline{\Phi}}_{i,j}(\tau) - \hat{\chi}^\top(\tau) W_{i,j}$ where $\hat{\chi}(\tau)$ is defined in* (21)*. Then, $\mathcal{J}_{\gamma,\theta,j}(\tau)$ defined in Assumption 7 is zero for $j \in J$ and $\tau \in \Upsilon$, i.e., by letting $n \to \infty$ followed by $\delta \to 0$, we have*

$$\sup \left\| \overline{\mathbb{P}}_{n,j} f_{\varepsilon_{i,j}(\tau)}(\delta_{i,j}(v,\tau) | W_{i,j}, Z_{i,j}) V_{i,j}^2(\tau) W_{i,j} \hat{\Phi}_{i,j}^\top(\tau) \right\|_{op} \xrightarrow{p} 0,$$

where the supremum is taken over $\{j \in J, ||v||_2 \leq \delta, \tau \in \Upsilon\}$ for $v = (v_b^\top, v_r^\top, v_t^\top)^\top$.

A.2.2 Cluster-Level Projection

Suppose $\underline{\Phi}_{i,j}(\tau)$ and $\hat{\underline{\Phi}}_{i,j}(\tau)$ are the original (infeasible) IV and its estimator, respectively, such that Assumption 7(i) is violated, i.e.,

$$\limsup_{n \to \infty} \overline{\mathbb{P}}_{n,j} f_{\varepsilon_{i,j}(\tau)}(0 | W_{i,j}, Z_{i,j}) W_{i,j}(\tau) \underline{\Phi}_{i,j}^\top(\tau) \neq 0.$$

We construct $\hat{\Phi}_{i,j}(\tau)$ as $\hat{\Phi}_{i,j}(\tau) = \hat{\underline{\Phi}}_{i,j}(\tau) - \hat{\chi}_j^\top(\tau) W_{i,j}$ such that Assumption 7(i) holds with $\hat{\Phi}_{i,j}(\tau)$ as the instrument. Let

$$\hat{\chi}_j(\tau) = \hat{\mathcal{J}}_{\gamma,\gamma,j}(\tau) \hat{\mathcal{J}}_{\gamma,\gamma,j}^{-}(\tau) \hat{\mathcal{J}}_{\gamma,\gamma,j}^{-}(\tau) \hat{\underline{\mathcal{J}}}_{\gamma,\theta,j}(\tau) \tag{38}$$

where $\hat{\mathcal{J}}_{\gamma,\gamma,j}(\tau)$ and $\hat{\underline{\mathcal{J}}}_{\gamma,\theta,j}(\tau)$ are defined in Section 3.[26]

Assumption 14. *(i) Suppose*

$$\sup_{\tau\in\Upsilon}\left[||\hat{\mathcal{J}}_{\gamma,\gamma,j}(\tau)-\mathcal{J}_{\gamma,\gamma,j}(\tau)||_{op}+||\hat{\mathcal{J}}_{\gamma,\theta,j}(\tau)-\underline{\mathcal{J}}_{\gamma,\theta,j}(\tau)||_{op}\right]=o_p(1).$$

(ii) Recall $\delta_{i,j}(v,\tau)=X_{i,j}^\top v_b+W_{i,j}^\top v_r+\hat{\Phi}_{i,j}^\top(\tau)v_t$. *Then,*

$$\sup\left\|\overline{\mathbb{P}}_{n,j}f_{\varepsilon_{i,j}(\tau)}(\delta_{i,j}(v,\tau)|W_{i,j},Z_{i,j})V_{i,j}^2(\tau)W_{i,j}\hat{\underline{\Phi}}_{i,j}^\top(\tau)-\underline{\mathcal{J}}_{\gamma,\theta,j}(\tau)\right\|_{op}\xrightarrow{p}0,$$

where the supremum is taken over $\{j\in J,||v||_2\le\delta,\tau\in\Upsilon\}$, $v=(v_b^\top,v_r^\top,v_t^\top)^\top$.

(iii) Define the kth largest singular value of $\mathcal{J}_{\gamma,\gamma,j}(\tau)$ *as* $\sigma_k(\mathcal{J}_{\gamma,\gamma,j}(\tau))$ *for* $k\in[d_w]$. *Then there exist constants* c,C *and integer* $R\in[1,d_w]$ *such that*

$$0<c\le\inf_{\tau\in\Upsilon}\sigma_R(\mathcal{J}_{\gamma,\gamma,j}(\tau))\le\sup_{\tau\in\Upsilon}\sigma_1(\mathcal{J}_{\gamma,\gamma,j}(\tau))\le C<\infty.$$

The use of generalized inverse in (21) accommodates the case that $\mathcal{J}_{\gamma,\gamma,j}(\tau)$ is not invertible. Assumption 13(iii) only requires that the minimum nonzero eigenvalue of $\mathcal{J}_{\gamma,\gamma,j}(\tau)$ is bounded away from zero uniformly over $\tau\in\Upsilon$.

Proposition A.2. *Suppose Assumption 14 holds and* $\hat{\Phi}_{i,j}(\tau)=\hat{\underline{\Phi}}_{i,j}(\tau)-\hat{\chi}_j^\top(\tau)W_{i,j}$ *where* $\hat{\chi}_j^\top(\tau)$ *is defined in* (21). *Then,* $\mathcal{J}_{\gamma,\theta,j}(\tau)$ *defined in Assumption 7 is zero for* $j\in J$ *and* $\tau\in\Upsilon$, *i.e., by letting* $n\to\infty$ *followed by* $\delta\to 0$, *we have*

$$\sup\left\|\overline{\mathbb{P}}_{n,j}f_{\varepsilon_{i,j}(\tau)}(\delta_{i,j}(v,\tau)|W_{i,j},Z_{i,j})V_{i,j}^2(\tau)W_{i,j}\hat{\Phi}_{i,j}^\top(\tau)\right\|_{op}\xrightarrow{p}0,$$

where the suprema are taken over $\{j\in J,||v||_2\le\delta,\tau\in\Upsilon\}$ *for* $v=(v_b^\top,v_r^\top,v_t^\top)^\top$.

Appendix B Wald Test with Two Quantile Indexes

In this section, we consider the following hypothesis. Let $\lambda_1^\top\beta_0(\tau_1)+\lambda_2^\top\beta_0(\tau_2)=\lambda_0$. Then, the null and local alternative hypotheses can be written as

$$\mathcal{H}_0:\lambda_1^\top\mu_\beta(\tau_1)+\lambda_2^\top\mu_\beta(\tau_2)=\lambda_0\quad v.s.\quad\mathcal{H}_{1,n}:\lambda_1^\top\mu_\beta(\tau_1)+\lambda_2^\top\mu_\beta(\tau_2)\neq\lambda_0,$$

where $\lambda_1,\lambda_2\in\mathbf{R}^{d_x\times d_m}$ and $\lambda_0\in\mathbf{R}^{d_m}$. We define $\Upsilon=\{\tau_1,\tau_2\}$. Let $\hat{A}_r$ be some weighting matrix and

$$T_{2,n}^{QR}=||\sqrt{n}(\lambda_1^\top\hat{\beta}(\tau_1)+\lambda_2^\top\hat{\beta}(\tau_2)-\lambda_0(\tau))||_{\hat{A}_r}.$$

[26] For a symmetric and positive semidefinite matrix A, we define A^- as its generalized inverse. The expression for $\hat{\chi}_j(\tau)$ is inspired by the minimum norm least squares solution of the normal equation that $\frac{1}{n}\sum_{i=1}^n X_i(Y_i-X_i^\top b)=0$ when $\frac{1}{n}\sum_{i=1}^n X_iX_i^\top$ is not invertible.

We consider the gradient wild bootstrap procedure below.

Step 1. Compute

$$(\hat{\beta}^r(\tau_1), \hat{\beta}^r(\tau_2)) = \arg \inf_{(b_1,b_2)\in\mathcal{B}(\tau_1,\tau_2)} \left[||\hat{\theta}(b_1,\tau_1)||_{\hat{A}_\phi(\tau_1)} + ||\hat{\theta}(b_2,\tau_2)||_{\hat{A}_\phi(\tau_2)}\right]$$

where $\mathcal{B}(\tau_1,\tau_2) = \{b_1, b_2 \in \mathcal{B} : \lambda_1^\top b_1 + \lambda_2^\top b_2 = \lambda_0\}$. Last, let $\hat{\gamma}^r(\tau) = \hat{\gamma}(\hat{\beta}^r(\tau),\tau)$ for $\tau = \tau_1, \tau_2$.

Step 2. For $\mathbf{G} = \{-1,1\}^q$ and any $g = (g_1, \cdots, g_q) \in \mathbf{G}$, compute $\hat{\beta}_g^*(\tau)$ and $\hat{\gamma}_g^*(\tau)$ via (28).

Step 3. Compute the $1-\alpha$ quantile of $T_{2,n}^{QR*} = ||\sqrt{n}(\lambda_1^\top(\hat{\beta}_g^*(\tau_1) - \hat{\beta}(\tau_1)) + \lambda_2^\top(\hat{\beta}_g^*(\tau_2) - \hat{\beta}(\tau_2)))||_{\hat{A}_r}$ as

$$\hat{c}_{2,n}^{QR}(1-\alpha) = \inf\left\{u \in \mathbf{R} : \frac{1}{|\mathbf{G}|}\sum_{g\in\mathbf{G}} 1\left\{T_{2,n}^{QR*} \le u\right\} \ge 1-\alpha\right\}.$$

We reject the null hypothesis if $T_{2,n}^{QR} > \hat{c}_{2,n}^{QR}(1-\alpha)$.

Theorem B.1. *Suppose Assumptions 5–7, 8, 9 hold. We further define* $\mu = \lambda_1^\top \mu_\beta(\tau_1) + \lambda_2^\top \mu_\beta(\tau_2)$ *and further impose that* $a_j(\tau_1) = a_j(\tau_2)$ *for* $j \in J$*. Then under* $\mathcal{H}_0$*, i.e.,* $\mu = 0$*,*

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

In addition, suppose $\min_{j\in J,\tau\in\{\tau_1,\tau_2\}} a_j(\tau) > 0$*. Then, under* $\mathcal{H}_{1,n}$*,*

$$\lim_{||\mu||_2\to\infty} \liminf_{n\to\infty} \mathbb{P}(T_{2,n}^{QR} > \hat{c}_{2,n}^{QR}(1-\alpha)) \to 1.$$

Next, we consider the Wald test with the CCE. We compute the Wald statistic as

$$T_{2,CR,n}^{QR} = ||\sqrt{n}(\lambda_1^\top \hat{\beta}(\tau_1) + \lambda_2^\top \hat{\beta}(\tau_2) - \lambda_0(\tau))||_{\hat{A}_{r,CR}},$$

where $\hat{A}_{r,CR}$ is the cluster-robust covariance matrix defined as

$$\hat{A}_{r,CR} = \Big[\lambda_1^\top \hat{\Omega}(\tau_1)\hat{V}(\tau_1,\tau_1)\hat{\Omega}^\top(\tau_1)\lambda_1 + \lambda_2^\top \hat{\Omega}(\tau_2)\hat{V}(\tau_2,\tau_2)\hat{\Omega}^\top(\tau_2)\lambda_2$$
$$+ \lambda_1^\top \hat{\Omega}(\tau_1)\hat{V}(\tau_1,\tau_2)\hat{\Omega}^\top(\tau_2)\lambda_2 + \lambda_2^\top \hat{\Omega}(\tau_2)\hat{V}(\tau_2,\tau_1)\hat{\Omega}^\top(\tau_1)\lambda_1\Big]^{-1}, \quad (39)$$

where $\hat{V}(\tau_1,\tau_2)$ is defined in (30) in Section 3.1.

The corresponding wild bootstrap statistic is

$$T_{2,CR,n}^{QR*} = ||\sqrt{n}(\lambda_1^\top(\hat{\beta}_g^*(\tau_1) - \hat{\beta}(\tau_1)) + \lambda_2^\top(\hat{\beta}_g^*(\tau_2) - \hat{\beta}(\tau_2)))||_{\hat{A}^*_{r,CR,g}},$$

where $\hat{A}^*_{r,CR,g}$ is the bootstrap counterpart of $\hat{A}_{r,CR}$ defined as

$$\hat{A}^*_{r,CR,g} = \Big[\lambda_1^\top \hat{\Omega}(\tau_1)\hat{V}_g^*(\tau_1,\tau_1)\hat{\Omega}^\top(\tau_1)\lambda_1 + \lambda_2^\top \hat{\Omega}(\tau_2)\hat{V}_g^*(\tau_2,\tau_2)\hat{\Omega}^\top(\tau_2)\lambda_2$$

$$+ \lambda_1^\top \hat{\Omega}(\tau_1)\hat{V}_g^*(\tau_1,\tau_2)\hat{\Omega}^\top(\tau_2)\lambda_2 + \lambda_2^\top \hat{\Omega}(\tau_2)\hat{V}_g^*(\tau_2,\tau_1)\hat{\Omega}^\top(\tau_1)\lambda_1\Big]^{-1}$$

and $\hat{V}_g^*(\tau_1,\tau_2)$ is defined in (32). We compute the critical value as the $1-\alpha$ quantile of $T_{2,CR,n}^{QR*}$, i.e.,

$$\hat{c}_{2,CR,n}^{QR}(1-\alpha) = \inf\left\{u \in \mathbf{R} : \frac{1}{|\mathbf{G}|}\sum_{g\in\mathbf{G}} 1\left\{T_{2,CR,n}^{QR*} \le u\right\} \ge 1-\alpha\right\}.$$

Theorem B.2. *Suppose Assumptions 5–7, 8(i)–8(iii), 9, and 11 hold. If Assumption 9(ii) holds, we further assume $a_j(\tau_1) = a_j(\tau_2)$ for $j \in J$. Define $\mu = \lambda_1^\top \mu_\beta(\tau_1) + \lambda_2^\top \mu_\beta(\tau_2)$. Then under $\mathcal{H}_0$, i.e., $\mu = 0$, we have*

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

Appendix C 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

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

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 $\widehat{\Gamma}_l = n^{-1}\sum_{j\in J}\sum_{i\in I_{n,j}}\sum_{k\in I_{n,j}}\left(\tilde{Z}_{i,j}X_{i,j,l}\right)f_{k,j}^\top$, 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

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

where $P_A = A(A'A)^- A'$ for any matrix A. 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 $rk_n = n\widehat{D}^\top \widehat{\Omega}^{-1}\widehat{D}$.[27]

[27] This choice follows Newey and Windmeijer (2009). Kleibergen (2005) uses alternative formula for rk_n, and Andrews

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

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

for any $g = (g_1, ..., g_q) \in \mathbf{G}$. Then, 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 $l = 1, ..., d_x$,

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

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 C.1. *If Assumptions 1-2 hold, $\beta_n = \beta_0$, and $q > d_z$, then*

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

and Guggenberger (2019) introduce alternative CQLR test statistic.

Supplementary Material for

"Wild Bootstrap for Instrumental Variables Regressions with Weak and Few Clusters "

Wenjie Wang[a] and Yichong Zhang[b]

[a]Division of Economics, School of Social Sciences, Nanyang Technological University

[b]School of Economics, Singapore Management University

Abstract

This document gathers together the supplementary material to the main paper. Section S.A provides additional empirical results. Sections S.B–S.D provide the proofs of results in Section 2. Sections S.E–S.G provide the proofs of results in Section 3. Section S.H provide the proofs of results in Appendix A. Section S.I provides the proofs of results in Appendix B. Section S.J provides the proofs of results in Appendix C. Sections S.K collects the Lemmas used in the proofs.

S.A Additional Empirical Results

In this section, we report the p-values of the AR_n^{QR} tests with all, female, and male samples for midwest, west, and south regions. In Figures 2–10, the X-axis represents the value of effect of Chinese imports on wages under the null while the Y-axis represents the corresponding p-values.

S.B Proof of Theorem 2.1

The arguments follow those in Canay et al. (2021). 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

$$T(s) = \left\| \lambda_\beta^\top \left(s_1^\top s_2^{-1} s_1\right)^{-1} s_1^\top s_2^{-1} \left(\sum_{j\in J} s_{3,j} \right) \right\|_{s_4} \tag{40}$$

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 $T^{(1)}(s|\mathbf{G'}) \leq \ldots \leq T^{(|\mathbf{G'}|)}(s|\mathbf{G'})$. Given this notation we can define the statistics $S_n, \widehat{S}_n \in \mathbb{S}$ as

$$S_n = \left(\widehat{Q}_{\tilde{Z}X}, \widehat{Q}_{\tilde{Z}\tilde{Z}}, \left\{ \frac{\sqrt{n_j}}{\sqrt{n}} \frac{1}{\sqrt{n_j}} \sum_{i\in I_{n,j}} \tilde{Z}_{i,j}\varepsilon_{i,j} : j \in J \right\}, \hat{A}_r \right),$$

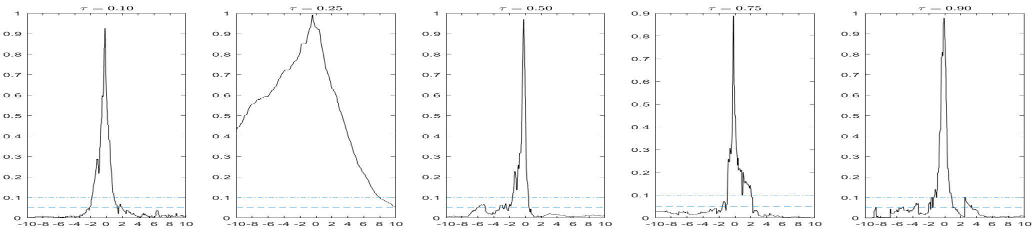

Figure 2: AR_n^{QR}-based Bootstrap p-values for Midwest Region

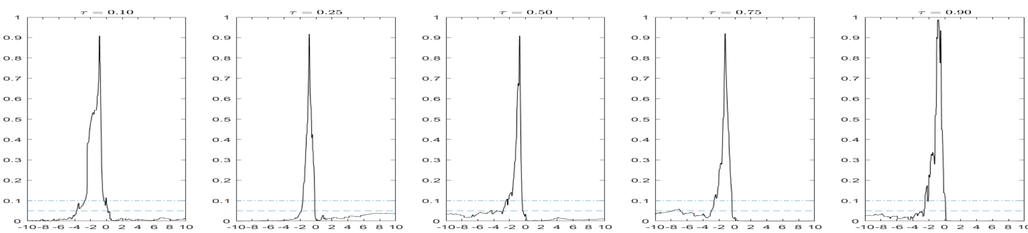

Figure 3: AR_n^{QR}-based Bootstrap p-values for South Region

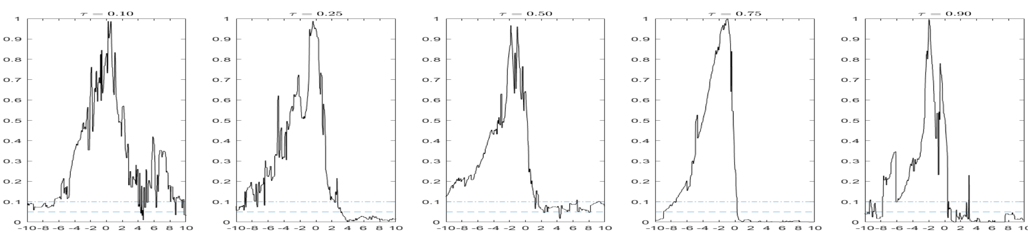

Figure 4: AR_n^{QR}-based Bootstrap p-values for West Region

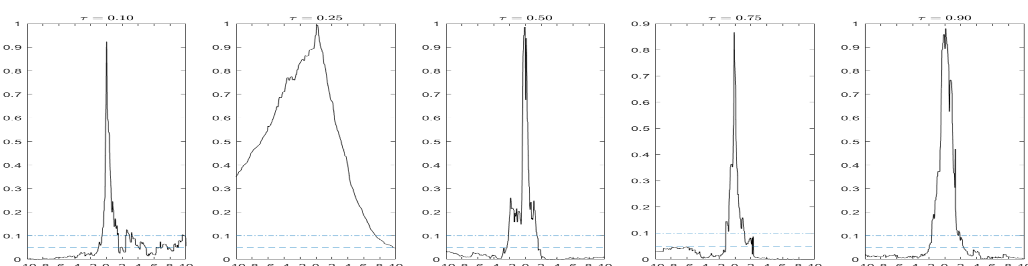

Figure 5: AR_n^{QR}-based Bootstrap p-values for the Female Sample in Midwest Region

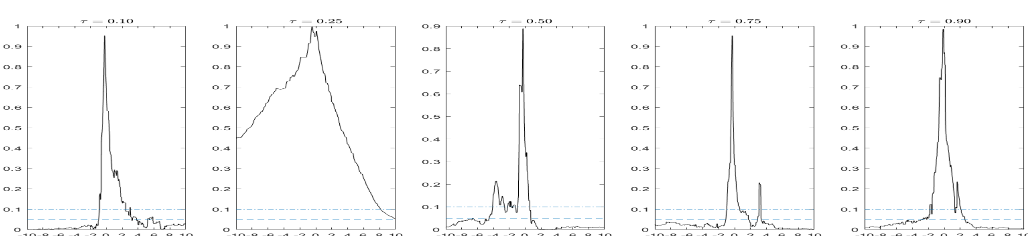

Figure 6: AR_n^{QR}-based Bootstrap p-values for the Male Sample in Midwest Region

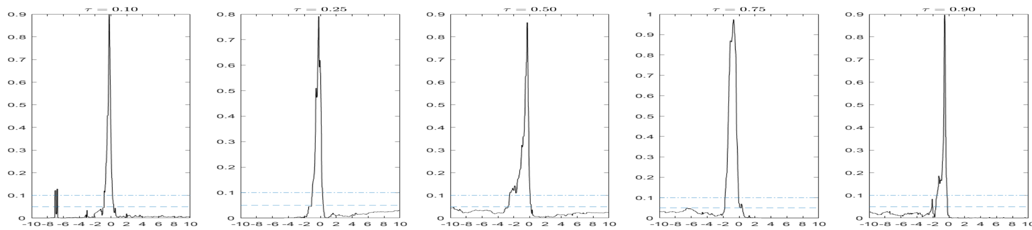

Figure 7: AR_n^{QR}-based Bootstrap p-values for the Female Sample in South Region

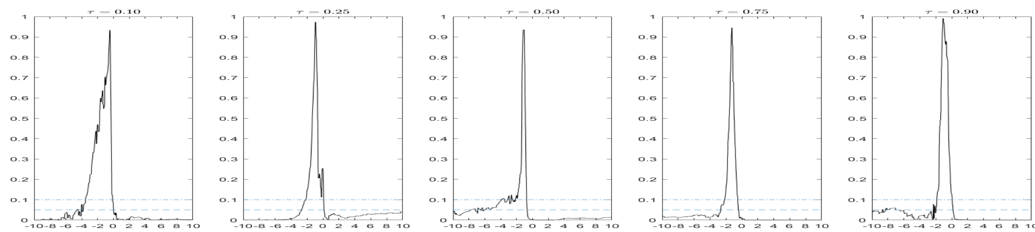

Figure 8: AR_n^{QR}-based Bootstrap p-values for the Male Sample in South Region

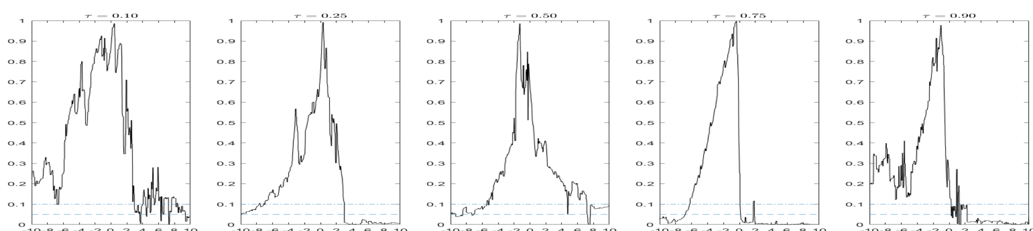

Figure 9: AR_n^{QR}-based Bootstrap p-values for the Female Sample in West Region

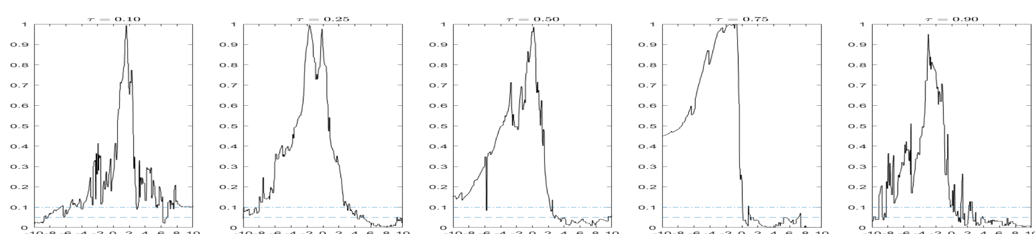

Figure 10: AR_n^{QR}-based Bootstrap p-values for the Female Sample in West Region

$$\widehat{S}_n = \left(\widehat{Q}_{\tilde{Z}X}, \widehat{Q}_{\tilde{Z}\tilde{Z}}, \left\{ \frac{\sqrt{n_j}}{\sqrt{n}} \frac{1}{\sqrt{n_j}} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} \hat{\varepsilon}^r_{i,j} : j \in J \right\}, \hat{A}_r \right). \tag{41}$$

Let E_n denote the event $E_n = I \left\{ \widehat{Q}_{\tilde{Z}X} \text{ is of full rank value and } \widehat{Q}_{\tilde{Z}\tilde{Z}} \text{ is invertible} \right\}$, and Assumption 1 implies that $\liminf_{n\to\infty} \mathbb{P}\{E_n = 1\} = 1$.

Note that whenever $E_n = 1$ and $\mathcal{H}_0$ is true, the Frisch-Waugh-Lovell theorem implies that

$$T_n = \left\| \sqrt{n} \left(\lambda_\beta^\top \hat{\beta} - \lambda_0 \right) \right\|_{\hat{A}_r} = \left\| \sqrt{n} \lambda_\beta^\top \left(\hat{\beta} - \beta_n \right) \right\|_{\hat{A}_r}$$
$$= \left\| \lambda_\beta^\top \widehat{Q}^{-1} \widehat{Q}_{\tilde{Z}X}^\top \widehat{Q}_{\tilde{Z}\tilde{Z}}^{-1} \sum_{j \in J} \frac{1}{\sqrt{n}} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} \varepsilon_{i,j} \right\|_{\hat{A}_r} = T(S_n). \tag{42}$$

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

$$T_n^*(g) = \left\| \sqrt{n} \lambda_\beta^\top \left(\hat{\beta}_g^* - \hat{\beta}^r \right) \right\|_{\hat{A}_r} = \left\| \lambda_\beta^\top \widehat{Q}^{-1} \widehat{Q}_{\tilde{Z}X}^\top \widehat{Q}_{\tilde{Z}\tilde{Z}}^{-1} \sum_{j \in J} \frac{1}{\sqrt{n}} \sum_{i \in I_{n,j}} g_j \tilde{Z}_{i,j} \hat{\varepsilon}_{i,j}^r \right\|_{\hat{A}_r} = T(g\widehat{S}_n). \tag{43}$$

Therefore, for any $x \in \mathbf{R}$ letting $\lceil x \rceil$ denote the smallest integer larger than x and $k^* \equiv \lceil |\mathbf{G}|(1-\alpha) \rceil$, we obtain from (42)-(43) that

$$I\left\{ T_n > \hat{c}_n(1-\alpha) \right\} = I\left\{ T(S_n) > T^{(k^*)}(\widehat{S}_n | \mathbf{G}) \right\}. \tag{44}$$

Furthermore, let $\iota_q \in \mathbf{G}$ correspond to the identity action, i.e., $\iota_q = (1, ..., 1) \in \mathbf{R}^q$, and similarly define $-\iota_q = (-1, ..., -1) \in \mathbf{R}^q$. Since $T(-\iota_q \widehat{S}_n) = T(\iota_q \widehat{S}_n)$, we obtain from (43) that

$$T\left(-\iota_q \widehat{S}_n \right) = T\left(\iota_q \widehat{S}_n \right) = \left\| \lambda_\beta^\top \widehat{Q}^{-1} \widehat{Q}_{\tilde{Z}X}^\top \widehat{Q}_{\tilde{Z}\tilde{Z}}^{-1} \sum_{j \in J} \frac{1}{\sqrt{n}} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} \left(y_{i,j} - X_{i,j}^\top \hat{\beta}^r - W_{i,j}^\top \hat{\gamma}^r \right) \right\|_{\hat{A}_r}$$
$$= \left\| \lambda_\beta^\top \widehat{Q}^{-1} \widehat{Q}_{\tilde{Z}X}^\top \widehat{Q}_{\tilde{Z}\tilde{Z}}^{-1} \sum_{j \in J} \frac{1}{\sqrt{n}} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} \left(y_{i,j} - X_{i,j}^\top \hat{\beta}^r \right) \right\|_{\hat{A}_r} = \left\| \sqrt{n} \lambda_\beta^\top (\hat{\beta} - \hat{\beta}^r) \right\|_{\hat{A}_r} = T\left(S_n \right), \tag{45}$$

where the third equality follows from $\sum_{j \in J} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} W_{i,j}^\top = 0$. (45) implies that if $k^* > |\mathbf{G}| - 2$, then $I\{T(S_n) > T^{(k^*)}(\widehat{S}_n | \mathbf{G})\} = 0$, and this gives the upper bound in Theorem 2.1. We therefore assume that $k^* \leq |\mathbf{G}| - 2$, in which case

$$\limsup_{n \to \infty} \mathbb{P}\{T(S_n) > T^{(k^*)}(\widehat{S}_n | \mathbf{G}); E_n = 1\}$$
$$= \limsup_{n \to \infty} \mathbb{P}\{T(S_n) > T^{(k^*)}(\widehat{S}_n | \mathbf{G} \setminus \{\pm \iota_q\}); E_n = 1\}$$
$$\leq \limsup_{n \to \infty} \mathbb{P}\{T(S_n) \geq T^{(k^*)}(\widehat{S}_n | \mathbf{G} \setminus \{\pm \iota_q\}); E_n = 1\}, \tag{46}$$

where the equality follows from (45) and $k^* \leq |\mathbf{G}| - 2$, and the inequality follows by set inclusion.

Then, to examine the right hand side of (46), first note that by Assumptions 1-2, 4, and the continuous mapping theorem we have

$$\left(\widehat{Q}_{\tilde{Z}X}, \widehat{Q}_{\tilde{Z}\tilde{Z}}, \left\{ \frac{\sqrt{n_j}}{\sqrt{n}} \frac{1}{\sqrt{n_j}} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} \varepsilon_{i,j} : j \in J \right\}, \hat{A}_r \right) \xrightarrow{d} \left(Q_{\tilde{Z}X}, Q_{\tilde{Z}\tilde{Z}}, \left\{ \sqrt{\xi_j} \mathcal{Z}_j : j \in J \right\}, A \right) \equiv S, \tag{47}$$

where $\xi_j > 0$ for all $j \in J$ by Assumption 2(ii). We further note that whenever $E_n = 1$, for every $g \in \mathbf{G}$,

$$\left|T(gS_n) - T(g\widehat{S}_n)\right| \leq \left\|\lambda_\beta^\top \widehat{Q}^{-1}\widehat{Q}_{\tilde{Z}X}^\top \widehat{Q}_{\tilde{Z}\tilde{Z}}^{-1} \sum_{j\in J} \frac{n_j}{n}\frac{1}{n_j}\sum_{i\in I_{n,j}} g_j \tilde{Z}_{i,j} X_{i,j}^\top \sqrt{n}(\beta_n - \hat{\beta}^r)\right\|_{\hat{A}_r} + \left\|\lambda_\beta^\top \widehat{Q}^{-1}\widehat{Q}_{\tilde{Z}X}^\top \widehat{Q}_{\tilde{Z}\tilde{Z}}^{-1} \sum_{j\in J} \frac{n_j}{n}\frac{1}{n_j}\sum_{i\in I_{n,j}} g_j \tilde{Z}_{i,j} W_{i,j}^\top \sqrt{n}(\gamma - \hat{\gamma}^r)\right\|_{\hat{A}_r}. \quad (48)$$

Note that whenever $\lambda_\beta^\top \beta_n = \lambda_0$ it follows from Assumption 1 and Amemiya (1985, Eq.(1.4.5)) that $\sqrt{n}(\hat{\beta}^r - \beta_n)$ and $\sqrt{n}(\hat{\gamma}^r - \gamma)$ are bounded in probability. Also, we have $\sum_{i\in I_{n,j}} \tilde{Z}_{i,j} W_{i,j}^\top / n_j = o_p(1)$ by using Assumptions 1(i), 2(iv), and the same argument as in Lemma A.2 of Canay et al. (2021). Therefore,

$$\limsup_{n\to\infty} \mathbb{P}\left\{\left\|\lambda_\beta^\top \widehat{Q}^{-1}\widehat{Q}_{\tilde{Z}X}^\top \widehat{Q}_{\tilde{Z}\tilde{Z}}^{-1} \sum_{j\in J} \frac{n_j}{n}\frac{1}{n_j}\sum_{i\in I_{n,j}} g_j \tilde{Z}_{i,j} W_{i,j}^\top \sqrt{n}(\gamma - \hat{\gamma}^r)\right\|_{\hat{A}_r} > \varepsilon; E_n = 1\right\} = 0. \quad (49)$$

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

$$\begin{aligned}
&\limsup_{n\to\infty} \mathbb{P}\left\{\left\|\lambda_\beta^\top \widehat{Q}^{-1}\widehat{Q}_{\tilde{Z}X}^\top \widehat{Q}_{\tilde{Z}\tilde{Z}}^{-1} \sum_{j\in J} \frac{n_j}{n}\frac{1}{n_j}\sum_{i\in I_{n,j}} g_j \tilde{Z}_{i,j} X_{i,j}^\top \sqrt{n}(\beta_n - \hat{\beta}^r)\right\|_{\hat{A}_r} > \varepsilon; E_n = 1\right\} \\
&= \limsup_{n\to\infty} \mathbb{P}\left\{\left\|\lambda_\beta^\top \left(Q_{\tilde{Z}X}^\top Q_{\tilde{Z}\tilde{Z}}^{-1} Q_{\tilde{Z}X}\right)^{-1} Q_{\tilde{Z}X}^\top Q_{\tilde{Z}\tilde{Z}}^{-1} \sum_{j\in J} \xi_j g_j a_j Q_{\tilde{Z}X} \sqrt{n}(\beta_n - \hat{\beta}^r)\right\|_{\hat{A}_r} > \varepsilon; E_n = 1\right\} \\
&= \limsup_{n\to\infty} \mathbb{P}\left\{\left\|\sum_{j\in J} \xi_j g_j a_j \sqrt{n}(\lambda_\beta^\top \beta - \lambda_\beta^\top \hat{\beta}^r)\right\|_{\hat{A}_r} > \varepsilon; E_n = 1\right\} = 0, \quad (50)
\end{aligned}$$

where the last equality holds because $\lambda_\beta^\top \hat{\beta}^r = \lambda_0$ under $\mathcal{H}_0$. Furthermore, under Assumption 3(i), $\lambda_\beta = 1$ and $\beta_n = \hat{\beta}^r$, so that (50) holds immediately.

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 (48), (49) and (50) imply that $T(g\widehat{S}_n) = T(gS_n) + o_P(1)$ for any $g \in \mathbf{G}$, and we obtain from (47) that

$$\left(T(S_n), \left\{T(g\widehat{S}_n) : g \in \mathbf{G}\right\}\right) \xrightarrow{d} \left(T(S), \{T(gS) : g \in \mathbf{G}\}\right). \quad (51)$$

Moreover, since $\liminf_{n\to\infty} P\{E_n = 1\} = 1$, it follows that $\left(T(S_n), E_n, \{T(g\widehat{S}_n) : g \in \mathbf{G}\}\right)$ converge jointly as well. Hence, Portmanteau's theorem implies that

$$\begin{aligned}
&\limsup_{n\to\infty} \mathbb{P}\{T(S_n) \geq T^{(k^*)}(\widehat{S}_n | \mathbf{G} \setminus \{\pm\iota_q\}); E_n = 1\} \\
&\leq \mathbb{P}\{T(S) \geq T^{(k^*)}(S | \mathbf{G} \setminus \{\pm\iota_q\})\} = \mathbb{P}\{T(S) > T^{(k^*)}(S | \mathbf{G} \setminus \{\pm\iota_q\})\}, \quad (52)
\end{aligned}$$

where the equality follows from $P\{T(S)=T(gS)\}=0$ for all $g\in\mathbf{G}\backslash\{\pm\iota_q\}$ since the covariance matrix of $\mathcal{Z}_j$ is full rank for all $j\in J$, and the limit of $\widehat{Q}^{-1}\widehat{Q}_{\tilde{Z}X}^{\top}\widehat{Q}_{\tilde{Z}\tilde{Z}}^{-1}$ is of full rank by Assumption 1. Finally, since $T(\iota_q S)=T(-\iota_q S)$, we obtain that $T(S)>T^{(k^*)}(S|\mathbf{G}\setminus\{\pm\iota_q\})$ if and only if $T(S)>T^{(k^*)}(S|\mathbf{G})$, which yields

$$\mathbb{P}\{T(S)>T^{(k^*)}(S|\mathbf{G}\backslash\{\pm\iota_q\})\}=\mathbb{P}\{T(S)>T^{(k^*)}(S|\mathbf{G})\}\leq\alpha, \tag{53}$$

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

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

$$\begin{aligned}\limsup_{n\to\infty}\mathbb{P}\{T_n>\hat{c}_n(1-\alpha)\}&\geq\liminf_{n\to\infty}\mathbb{P}\{T(S_n)>T^{(k^*)}(S_n|\mathbf{G})\}\geq\mathbb{P}\{T(S)>T^{(k^*)}(S|\mathbf{G})\}\\&\geq\mathbb{P}\{T(S)>T^{(k^*+2)}(S|\mathbf{G})\}+\mathbb{P}\{T(S)=T^{(k^*+2)}(S|\mathbf{G})\}\geq\alpha-\frac{1}{2^{q-1}},\end{aligned} \tag{54}$$

where the first inequality follows from (44), the second inequality follows from Portmanteau's theorem, the third 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 (40) and Assumption 2(i)-(ii), and the last equality follows from noticing that $k^*+2=\lceil|\mathbf{G}|((1-\alpha)+2/|\mathbf{G}|)\rceil=\lceil|\mathbf{G}|(1-\alpha')\rceil$ with $\alpha'=\alpha-\frac{1}{2^{q-1}}$ and the properties of randomization tests. This completes the proof of the lower bound of part (i) in the theorem.

For part (ii) in the theorem, the power of the T_n-based wild bootstrap test, notice that

$$\begin{aligned}||\sqrt{n}(\lambda_\beta^\top\hat{\beta}-\lambda_0)||_{\hat{A}_r}&=||\sqrt{n}\lambda_\beta^\top(\hat{\beta}-\beta_n)+\sqrt{n}\lambda_\beta^\top(\beta_n-\hat{\beta}^r)||_{\hat{A}_r}\\&=\Big\|\lambda_\beta^\top\widehat{Q}^{-1}\widehat{Q}_{\tilde{Z}X}^{\top}\widehat{Q}_{\tilde{Z}\tilde{Z}}^{-1}\sum_{j\in J}\sum_{i\in I_{n,j}}\frac{\tilde{Z}_{i,j}\varepsilon_{i,j}}{\sqrt{n}}+\sqrt{n}\lambda_\beta^\top(\beta_n-\hat{\beta}^r)\Big\|_{\hat{A}_r}\\&=\Big\|\lambda_\beta^\top\widehat{Q}^{-1}\widehat{Q}_{\tilde{Z}X}^{\top}\widehat{Q}_{\tilde{Z}\tilde{Z}}^{-1}\sum_{j\in J}\sum_{i\in I_{n,j}}\left(\frac{\tilde{Z}_{i,j}\varepsilon_{i,j}}{\sqrt{n}}+\frac{\tilde{Z}_{i,j}X_{i,j}^\top}{n}\sqrt{n}(\beta_n-\hat{\beta}^r)\right)\Big\|_{\hat{A}_r}.\end{aligned} \tag{55}$$

Notice that Assumption 1, Amemiya (1985, Eq.(1.4.5)), and $\sqrt{n}(\lambda_\beta^\top\beta_n-\lambda_0)=\lambda_\beta^\top\mu_\beta=\mu$ imply that $\sqrt{n}(\hat{\beta}^r-\beta_n)$ and $\sqrt{n}(\hat{\gamma}^r-\gamma)$ are bounded in probability. Therefore, we have for the bootstrap analogue,

$$\begin{aligned}&||\sqrt{n}\lambda_\beta^\top(\hat{\beta}_g^*-\hat{\beta}^r)||_{\hat{A}_r}\\&=\Big\|\lambda_\beta^\top\widehat{Q}^{-1}\widehat{Q}_{\tilde{Z}X}^{\top}\widehat{Q}_{\tilde{Z}\tilde{Z}}^{-1}\sum_{j\in J}\sum_{i\in I_{n,j}}g_j\left(\frac{\tilde{Z}_{i,j}\varepsilon_{i,j}}{\sqrt{n}}+\frac{\tilde{Z}_{i,j}X_{i,j}^\top}{n}\sqrt{n}(\beta_n-\hat{\beta}^r)+\frac{\tilde{Z}_{i,j}W_{i,j}^\top}{n}\sqrt{n}(\gamma-\hat{\gamma}^r)\right)\Big\|_{\hat{A}_r}\\&=\Big\|\lambda_\beta^\top\widehat{Q}^{-1}\widehat{Q}_{\tilde{Z}X}^{\top}\widehat{Q}_{\tilde{Z}\tilde{Z}}^{-1}\sum_{j\in J}\sum_{i\in I_{n,j}}g_j\left(\frac{\tilde{Z}_{i,j}\varepsilon_{i,j}}{\sqrt{n}}+\frac{\tilde{Z}_{i,j}X_{i,j}^\top}{n}\sqrt{n}(\beta_n-\hat{\beta}^r)\right)\Big\|_{\hat{A}_r}+o_P(1),\end{aligned} \tag{56}$$

where the last equality follows from $\sum_{i\in I_{n,j}}\tilde{Z}_{i,j}W_{i,j}^\top/n_j=o_p(1)$ by Assumptions 1(i) and 2(iv).

Furthermore, we notice that

$$\hat{\beta}^r = \hat{\beta} - \widehat{Q}^{-1}\lambda_\beta\left(\lambda_\beta^\top\widehat{Q}^{-1}\lambda_\beta\right)^{-1}\left(\lambda_\beta^\top\hat{\beta}-\lambda_0\right)$$
$$= \hat{\beta} - \widehat{Q}^{-1}\lambda_\beta\left\{\left(\lambda_\beta^\top\widehat{Q}^{-1}\lambda_\beta\right)^{-1}\lambda_\beta^\top(\hat{\beta}-\beta_n) + \left(\lambda_\beta^\top\widehat{Q}^{-1}\lambda_\beta\right)^{-1}(\lambda_\beta^\top\beta_n-\lambda_0)\right\}. \quad (57)$$

Therefore, employing (57) with $\sqrt{n}(\lambda_\beta^\top\beta_n-\lambda_0)=\lambda_\beta^\top\mu_\beta$, we conclude that whenever $E_n=1$,

$$\sum_{i\in I_{n,j}}\frac{\tilde{Z}_{i,j}X_{i,j}^\top}{n}\sqrt{n}(\beta_n-\hat{\beta}^r) = \sum_{i\in I_{n,j}}\frac{\tilde{Z}_{i,j}X_{i,j}^\top}{n}\left\{\left(I_{d_x}-\widehat{Q}^{-1}\lambda_\beta\left(\lambda_\beta^\top\widehat{Q}^{-1}\lambda_\beta\right)^{-1}\lambda_\beta^\top\right)\sqrt{n}(\beta_n-\hat{\beta})\right.$$
$$\left.+\widehat{Q}^{-1}\lambda_\beta\left(\lambda_\beta^\top\widehat{Q}^{-1}\lambda_\beta\right)^{-1}\lambda_\beta^\top\mu_\beta\right\}. \quad (58)$$

Together with (56), this implies that

$$\limsup_{n\to\infty}\mathbb{P}\left(\begin{array}{c}\left\|\sqrt{n}\lambda_\beta^\top(\hat{\beta}_g^*-\hat{\beta}^r)-\lambda_\beta^\top\widehat{Q}^{-1}\widehat{Q}_{\tilde{Z}X}^\top\widehat{Q}_{\tilde{Z}\tilde{Z}}^{-1}\right.\\ \times\sum_{j\in J}\sum_{i\in I_{n,j}}g_j\left(\frac{\tilde{Z}_{i,j}\varepsilon_{i,j}}{\sqrt{n}}+\xi_j\widehat{Q}_{\tilde{Z}X,j}\widehat{Q}^{-1}\lambda_\beta\left(\lambda_\beta^\top\widehat{Q}^{-1}\lambda_\beta\right)^{-1}\lambda_\beta^\top\mu_\beta\right)\Big\|_{\hat{A}_r}>\varepsilon;E_n=1\end{array}\right)$$
$$=\limsup_{n\to\infty}\mathbb{P}\left(\begin{array}{c}\left\|\sqrt{n}\lambda_\beta^\top(\hat{\beta}_g^*-\hat{\beta}^r)-\lambda_\beta^\top Q^{-1}Q_{\tilde{Z}X}^\top Q_{\tilde{Z}\tilde{Z}}^{-1}\right.\\ \times\sum_{j\in J}g_j\left[\sqrt{\xi_j}\mathcal{Z}_j+\xi_ja_jQ_{\tilde{Z}X}Q^{-1}\lambda_\beta\left(\lambda_\beta^\top Q^{-1}\lambda_\beta\right)^{-1}\mu\right]\Big\|_{\hat{A}_r}>\varepsilon;E_n=1\end{array}\right)$$
$$=\limsup_{n\to\infty}\mathbb{P}\left(\left\|\sqrt{n}\lambda_\beta^\top(\hat{\beta}_g^*-\hat{\beta}^r)-\sum_{j\in J}g_j\left[\sqrt{\xi_j}\lambda_\beta^\top Q^{-1}Q_{\tilde{Z}X}^\top Q_{\tilde{Z}\tilde{Z}}^{-1}\mathcal{Z}_j+\xi_ja_j\mu\right]\right\|_{\hat{A}_r}>\varepsilon;E_n=1\right)=0.$$

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, to establish the result in part (ii), it suffices to show that as $||\mu||_2\to\infty$,

$$\liminf_{n\to\infty}\mathbb{P}\{T_n>\max_{g\in\mathbf{G}_s}T_n^*(g)\}\to 1, \quad (59)$$

which follows under similar arguments as those employed in the proof of Theorem 3.2 in Canay et al. (2021). ■

S.C Proof of Theorem 2.2

The proof for the asymptotic size of the $T_{CR,n}$-based wild bootstrap test follows similar arguments as those for the T_n-based wild bootstrap test in Theorem 2.1 and the arguments in the proof of Theorem 3.3 in Canay et al. (2021), and is thus omitted.

For the power analysis, we define

$$T_{CR,\infty}(g)=\left\|\lambda_\beta^\top\tilde{Q}\left[\sum_{j\in J}g_j\sqrt{\xi}_j\mathcal{Z}_j\right]+c_{0,g}\mu\right\|_{A_{r,CR,g}}, \quad (60)$$

where $\tilde{Q} = Q^{-1}Q_{\tilde{Z}X}^{\top}Q_{\tilde{Z}\tilde{Z}}^{-1}$, $c_{0,g} = \sum_{j\in J}\xi_j g_j a_j$, 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: $(T_{CR,\infty})^{(1)} \le \cdots \le (T_{CR,\infty})^{|\mathbf{G}|}$.

By the Portmanteau theorem, we have

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

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

$$\mathbb{P}\left\{T_{CR,\infty}(\iota_q) > \max_{g\in\mathbf{G}_s} T_{CR,\infty}(g)\right\} \to 1. \tag{61}$$

Then, given $|\mathbf{G}_s| = |\mathbf{G}| - 2^{q-q_s+1}$ and $\lceil|\mathbf{G}|(1-\alpha)\rceil \le |\mathbf{G}| - 2^{q-q_s+1}$, (61) implies that as $||\mu(\tau)||_2 \to \infty$,

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

Therefore, it suffices to establish (61).

By (60), we see that

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

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

$$T_{CR,\infty}(\iota_q) \ge c||\mu||_2^2 - O_p(1). \tag{62}$$

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

$$T_{CR,\infty}(g) = \left\{(N_{0,g} + c_{0,g}\mu)^\top\left[\sum_{j\in J}\xi_j(N_{j,g} + c_{j,g}\mu)(N_{j,g} + c_{j,g}\mu)^\top\right]^{-1}(N_{0,g} + c_{0,g}\mu)\right\}, \tag{63}$$

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 contradict the definition of $\mathbf{G}_s$), which, together with the assumption that $\min_{j\in J_s}|a_j| > 0$, implies that

$$\min_{g\in\mathbf{G}_s}\sum_{j\in J}\xi_j c_{j,g}^2 > 0.$$

In addition, we note that

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

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

$$
M_1+M_2\mu^\top+\mu M_2^\top+\overline{c}^2\mu\mu^\top=M_1-\frac{M_2M_2^\top}{\overline{c}^2}+\left(\frac{M_2}{\overline{c}}+\overline{c}\mu\right)\left(\frac{M_2}{\overline{c}}+\overline{c}\mu\right)^\top.
$$

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

$$
u^\top\left(M_1-\frac{M_2M_2^\top}{\overline{c}^2}\right)u=\sum_{j\in J}\xi_j(u^\top N_{j,g})^2-\frac{\left(\sum_{j\in J}\xi_j u^\top N_{j,g}c_{j,g}\right)^2}{\sum_{j\in J}\xi_j c_{j,g}^2}\geq 0,
$$

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 $q>d_r$. Therefore, the matrix $\mathbb{M}\equiv M_1-\frac{M_2M_2^\top}{\overline{c}^2}$ is invertible with probability one. Specifically, denote $\mathbb{M}$ as $\mathbb{M}(g)$ to highlight its dependence on g. We have $\sup_{g\in\mathbf{G}_s}(\lambda_{\min}(\mathbb{M}(g)))^{-1}=O_p(1)$. In addition, denote $\frac{M_2}{\overline{c}}+\overline{c}\mu$ 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

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

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

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

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

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 nonnegative scalar, and the last inequality holds by substituting in the expressions for $\mathbb{M}_0$ and $\bar{c}$ and denoting $\mathbb{M}$ as $\mathbb{M}(g)$.

Then, we have

$$
\max_{g\in\mathbf{G}_s} T_{CR,\infty}(g) \leq \max_{g\in\mathbf{G}_s} C(g). \tag{64}
$$

Combining (62) and (64), we have, as $||\mu||_2 \to \infty$,

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

where the last convergence holds because $\max_{g\in\mathbf{G}_s} C(g) = O_p(1)$ and does not depend on μ. ■

S.D Proof of Theorem 2.3

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

$$
T_{AR}(s) = \left\|\sum_{j\in J} s_{1j}\right\|_{s_2}. \tag{65}
$$

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

$$
S_n = \left\{\frac{\sqrt{n_j}}{\sqrt{n}}\frac{1}{\sqrt{n_j}}\sum_{i\in I_{n,j}}\tilde{Z}_{i,j}\varepsilon_{i,j} : j \in J, \hat{A}_z\right\}, \widehat{S}_n = \left\{\frac{\sqrt{n_j}}{\sqrt{n}}\frac{1}{\sqrt{n_j}}\sum_{i\in I_{n,j}}\tilde{Z}_{i,j}\hat{\varepsilon}_{i,j}^r : j \in J, \hat{A}_z\right\}, \tag{66}
$$

where $\hat{\varepsilon}_{i,j}^r = y_{i,j} - X_{i,j}^\top\beta_0 - W_{i,j}^\top\bar{\gamma}^r$. Note that by the Frisch-Waugh-Lovell theorem,

$$
AR_n = \left\|\sum_{j\in J}\frac{1}{\sqrt{n}}\sum_{i\in I_{n,j}}\tilde{Z}_{i,j}\varepsilon_{i,j}\right\|_{\hat{A}_z} = T_{AR}(S_n). \tag{67}
$$

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

$$AR_n^*(g) = \left\| \sum_{j \in J} \frac{1}{\sqrt{n}} \sum_{i \in I_{n,j}} g_j \tilde{Z}_{i,j} \hat{\varepsilon}_{i,j}^r \right\|_{\hat{A}_z} = T_{AR}(g\widehat{S}_n). \tag{68}$$

Therefore, letting $k^* \equiv \lceil |\mathbf{G}|(1-\alpha) \rceil$, we obtain from (67)-(68) that

$$I\left\{AR_n > \hat{c}_{AR,n}(1-\alpha)\right\} = I\left\{T_{AR}(S_n) > T_{AR}^{(k^*)}(\widehat{S}_n|\mathbf{G})\right\}. \tag{69}$$

Furthermore, similar to the arguments in the proof of Theorem 2.1, we have

$$\begin{aligned} T_{AR}\left(-\iota_q \widehat{S}_n\right) = T_{AR}\left(\iota_q \widehat{S}_n\right) &= \left\| \sum_{j \in J} \frac{1}{\sqrt{n}} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j}\left(y_{i,j} - X_{i,j}^\top \beta_0 - W_{i,j}^\top \bar{\gamma}^r\right) \right\|_{\hat{A}_z} \\ &= \left\| \sum_{j \in J} \frac{1}{\sqrt{n}} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j}\left(\varepsilon_{i,j} + W_{i,j}^\top(\gamma - \bar{\gamma}^r)\right) \right\|_{\hat{A}_z} = T_{AR}\left(S_n\right). \end{aligned} \tag{70}$$

(70) implies that if $k^* > |\mathbf{G}| - 2$, then $I\{T_{AR}(S_n) > T_{AR}^{(k^*)}(\widehat{S}_n|\mathbf{G})\} = 0$, and this gives the upper bound in Theorem 2.3. We therefore assume that $k^* \leq |\mathbf{G}| - 2$, in which case

$$\begin{aligned} \limsup_{n\to\infty} \mathbb{P}\{T_{AR}(S_n) > T_{AR}^{(k^*)}(\widehat{S}_n|\mathbf{G})\} &= \limsup_{n\to\infty} \mathbb{P}\{T_{AR}(S_n) > T_{AR}^{(k^*)}(\widehat{S}_n|\mathbf{G} \setminus \{\pm\iota_q\})\} \\ &\leq \limsup_{n\to\infty} \mathbb{P}\{T_{AR}(S_n) \geq T_{AR}^{(k^*)}(\widehat{S}_n|\mathbf{G} \setminus \{\pm\iota_q\})\}. \end{aligned} \tag{71}$$

Then, to examine the right hand side of (71), first note that by Assumption 2(i) and the continuous mapping theorem we have

$$\left\{ \frac{\sqrt{n_j}}{\sqrt{n}} \frac{1}{\sqrt{n_j}} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} \varepsilon_{i,j} : j \in J, \hat{A}_z \right\} \xrightarrow{d} \left\{ \sqrt{\xi_j} \mathcal{Z}_j : j \in J, A_z \right\} \equiv S, \tag{72}$$

where $\xi_j > 0$ for all $j \in J$ by Assumption 2(ii). Furthermore, by Assumptions 1(i), 2(iii)-(iv) and $\beta_n = \beta_0$, we have for every $g \in \mathbf{G}$,

$$T_{AR}(g\widehat{S}_n) = T_{AR}(gS_n) + o_P(1). \tag{73}$$

We thus obtain from results in (72)-(73) and the continuous mapping theorem that

$$\left(T_{AR}(S_n), \left\{T_{AR}(g\widehat{S}_n) : g \in \mathbf{G}\right\}\right) \xrightarrow{d} \left(T_{AR}(S), \{T_{AR}(gS) : g \in \mathbf{G}\}\right). \tag{74}$$

Then, the upper and lower bounds follow by applying similar arguments as those for Theorem 2.1. The proof for the $AR_{CR,n}$ and $AR_{R,n}$ bootstrap tests follows by using similar arguments as those for AR_n and the proof of Theorem 3.3 in Canay et al. (2021), and is thus omitted. ■

S.E Proof of Theorem 3.1

Let

$$\tilde{\Omega}(\tau)=\left[\mathcal{J}_{\theta,\beta}^{\top}(\tau)\mathcal{J}_{\theta,\theta}^{-1}(\tau)A_{\phi}(\tau)\mathcal{J}_{\theta,\theta}^{-1}(\tau)\mathcal{J}_{\theta,\beta}(\tau)\right]^{-1}\mathcal{J}_{\theta,\beta}^{\top}(\tau)\mathcal{J}_{\theta,\theta}^{-1}(\tau)A_{\phi}(\tau)\mathcal{J}_{\theta,\theta}^{-1}(\tau) \tag{75}$$

Because $\mathcal{J}_{\pi,\pi}(\tau)$ is a block matrix, we have

$$\mathcal{J}_{\pi,\beta}^{\top}(\tau)\overline{\mathcal{J}}_{\theta}^{\top}(\tau)A_{\phi}(\tau)\overline{\mathcal{J}}_{\theta}(\tau)\mathcal{J}_{\pi,\beta}(\tau)=\mathcal{J}_{\theta,\beta}^{\top}(\tau)\mathcal{J}_{\theta,\theta}^{-1}(\tau)A_{\phi}(\tau)\mathcal{J}_{\theta,\theta}^{-1}(\tau)\mathcal{J}_{\theta,\beta}(\tau)$$

and

$$\overline{\mathcal{J}}_{\theta}(\tau)f_{\tau}(D_{i,j},\beta_n(\tau),\gamma_n(\tau),0)=\mathcal{J}_{\theta,\theta}^{-1}(\tau)\tilde{f}_{\tau}(D_{i,j},\beta_n(\tau),\gamma_n(\tau),0).$$

Therefore, by Lemma S.K.2, we have

$$\sqrt{n}\left(\hat{\beta}(\tau)-\beta_n(\tau)\right)=\tilde{\Omega}(\tau)\left[\sum_{j\in J}\sqrt{\xi_j}\sqrt{n_j}(\mathbb{P}_{n,j}-\overline{\mathbb{P}}_{n,j})\tilde{f}_{\tau}(D_{i,j},\beta_n(\tau),\gamma_n(\tau),0)\right]+o_p(1),$$

where the $o_p(1)$ term holds uniformly over $\tau\in\Upsilon$.

Then we have

$$\begin{aligned}\sqrt{n}(\lambda_{\beta}^{\top}(\tau)\hat{\beta}(\tau)-\lambda_0(\tau))&=\lambda_{\beta}^{\top}(\tau)\sqrt{n}\left(\hat{\beta}(\tau)-\beta_n(\tau)\right)+\mu(\tau)\\&=\lambda_{\beta}^{\top}(\tau)\tilde{\Omega}(\tau)\left[\sum_{j\in J}\sqrt{\xi_j}\mathcal{Z}_j(\tau)\right]+\mu(\tau)+o_p(1),\end{aligned}$$

where the $o_p(1)$ term holds uniformly over $\tau\in\Upsilon$.

By Lemma S.K.4, we have

$$\begin{aligned}&\sqrt{n}\lambda_{\beta}^{\top}(\tau)(\hat{\beta}_g^{*}(\tau)-\hat{\beta}(\tau))\\&=\lambda_{\beta}^{\top}(\tau)\tilde{\Omega}(\tau)\sqrt{n}\left[\sum_{j\in J}\frac{g_j\xi_j}{n_j}\sum_{i\in I_{n,j}}\tilde{f}_{\tau}(D_{i,j},\beta_n(\tau),\gamma_n(\tau),0)\right]\\&-\lambda_{\beta}^{\top}(\tau)\sum_{j\in J}\Omega(\tau)\xi_j g_j\mathcal{J}_{\pi,\beta,j}(\tau)\sqrt{n}(\hat{\beta}^{r}(\tau)-\beta_n(\tau))+o_p(1)\\&=\lambda_{\beta}^{\top}(\tau)\tilde{\Omega}(\tau)\sqrt{n}\left[\sum_{j\in J}\frac{g_j\xi_j}{n_j}\sum_{i\in I_{n,j}}\tilde{f}_{\tau}(D_{i,j},\beta_n(\tau),\gamma_n(\tau),0)\right]+\overline{a}_g^{*}(\tau)\mu(\tau)+o_p(1)\end{aligned}$$

where $o_p(1)$ term holds uniformly over $\tau\in\Upsilon$, $\overline{a}_g^{*}(\tau)=\sum_{j\in J}\xi_j g_j a_j(\tau)$, and the last equality holds because

$$\begin{aligned}&\lambda_{\beta}^{\top}(\tau)\sum_{j\in J}\Omega(\tau)\xi_j g_j\mathcal{J}_{\pi,\beta,j}(\tau)\sqrt{n}(\hat{\beta}^{r}(\tau)-\beta_n(\tau))\\&=\lambda_{\beta}^{\top}(\tau)\sum_{j\in J}\tilde{\Omega}(\tau)\xi_j g_j\mathcal{J}_{\theta,\beta,j}(\tau)\sqrt{n}(\hat{\beta}^{r}(\tau)-\beta_n(\tau))\end{aligned}$$

$$= \overline{a}_g^*(\tau)\sqrt{n}\lambda_\beta^\top(\tau)(\hat{\beta}^r(\tau) - \beta_n(\tau)) = -\overline{a}_g^*(\tau)\mu(\tau). \tag{76}$$

Let $T_\infty^{QR}(g) = \sup_{\tau\in\Upsilon}\left[||\lambda_\beta^\top(\tau)\tilde{\Omega}(\tau)\left[\sum_{j\in J} g_j\sqrt{\xi_j}\mathcal{Z}_j(\tau)\right] + \sum_{j\in J} g_j\xi_j a_j(\tau)\mu(\tau)||_{A_r(\tau)}\right]$. Then, we have, uniformly over $\tau \in \Upsilon$

$$\{T_n^{QR}, \{T_n^{QR*}(g)\}_{g\in\mathbf{G}}\} \rightsquigarrow (T_\infty^{QR}(\iota_q), \{T_\infty^{QR}(g)\}_{g\in\mathbf{G}}),$$

where ι_q is a $q\times 1$ vector of ones and we use the fact that $1 = \sum_{j\in J}\xi_j a_j(\tau)$. In addition, under $\mathcal{H}_0$, $T_\infty^{QR}(g) \overset{d}{=} T_\infty^{QR}(g')$. Let $k^* = \lceil|\mathbf{G}|(1-\alpha)\rceil$. We order $\{T_n^{QR*}(g)\}_{g\in\mathbf{G}}$ and $\{T_\infty^{QR}(g)\}_{g\in\mathbf{G}}$ in ascending order:

$$(T_n^{QR*})^{(1)} \leq \cdots \leq (T_n^{QR*})^{(|\mathbf{G}|)} \quad \text{and} \quad (T_\infty^{QR})^{(1)} \leq \cdots \leq (T_\infty^{QR})^{|\mathbf{G}|}.$$

Then, we have

$$\hat{c}_n^{QR}(1-\alpha) \xrightarrow{p} (T_\infty^{QR})^{(k^*)}$$

and

$$\limsup_{n\to\infty}\mathbb{P}\{T_n^{QR} > \hat{c}_n^{QR}(1-\alpha)\} \leq \limsup_{n\to\infty}\mathbb{P}\{T_n^{QR} \geq \hat{c}_n^{QR}(1-\alpha)\}$$
$$\leq \mathbb{P}\left\{T_\infty^{QR}(\iota_q) \geq (T_\infty^{QR})^{(k^*)}\right\} \leq \alpha + \frac{1}{2^{q-1}},$$

where the second inequality is due to Portmanteau's theorem (see, e.g., (van der Vaart and Wellner, 1996), Theorem 1.3.4(iii)), and the third inequality is due to the properties of randomization tests (see, e.g., (Lehmann and Romano, 2006), Theorem 15.2.1) and the facts that the distribution of $T_\infty^{QR}(g)$ is invariant w.r.t. g, $T_\infty^{QR}(g) = T_\infty^{QR}(-g)$, and $T_\infty^{QR}(g) \neq T_\infty^{QR}(g')$ if $g \notin \{g', -g'\}$. Similarly, we have

$$\liminf_{n\to\infty}\mathbb{P}\{T_n^{QR} > \hat{c}_n^{QR}(1-\alpha)\} = \liminf_{n\to\infty}\mathbb{P}\{T_n^{QR} > (T_n^{QR*})^{(k^*)}\} \geq \mathbb{P}\left\{T_\infty^{QR}(\iota_q) > (T_\infty^{QR})^{(k^*)}\right\} \geq \alpha - \frac{1}{2^{q-1}}.$$

To see the last inequality, we note that $T_\infty^{QR}(g) = T_\infty^{QR}(-g)$, and $T_\infty^{QR}(g) \neq T_\infty^{QR}(g')$ if $g \notin \{g', -g'\}$. Therefore,

$$\sum_{g\in\mathbf{G}} 1\{T_\infty^{QR}(g) \leq (T_\infty^{QR})^{(k^*)}\} \leq k^* + 1.$$

Then, with $|\mathbf{G}| = 2^q$, we have

$$|\mathbf{G}|\mathbb{E}1\{T_\infty^{QR}(\iota_q) > (T_\infty^{QR})^{(k^*)}\} = \mathbb{E}\sum_{g\in\mathbf{G}} 1\{T_\infty^{QR}(g) > (T_\infty^{QR})^{(k^*)}\}$$
$$= |\mathbf{G}| - \mathbb{E}\sum_{g\in\mathbf{G}} 1\{T_\infty^{QR}(g) \leq (T_\infty^{QR})^{(k^*)}\} \geq |\mathbf{G}| - (k^* + 1)$$
$$\geq \lfloor|\mathbf{G}|\alpha\rfloor - 1 \geq |\mathbf{G}|\alpha - 2,$$

where the first equality holds because $T^{QR}_{\infty}(\iota_q) \stackrel{d}{=} T^{QR}_{\infty}(g)$ for $g \in \mathbf{G}$.

Under $\mathcal{H}_{1,n}$, we still have

$$\liminf_{n\to\infty} \mathbb{P}\{T^{QR}_n > \hat{c}^{QR}_n(1-\alpha)\} \geq \mathbb{P}\left\{T^{QR}_{\infty}(\iota_q) > (T^{QR}_{\infty})^{(k^*)}\right\}.$$

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 aim to show that, as $\inf_{\tau\in\Upsilon} ||\mu(\tau)|| \to \infty$,

$$\mathbb{P}\{T^{QR}_{\infty}(\iota_q) > \max_{g\in\mathbf{G}_s} T^{QR}_{\infty}(g)\} \to 1. \tag{77}$$

In addition, we note that $|\mathbf{G}_s| = |\mathbf{G}| - 2^{q-q_s+1} \geq k^*$. This implies as $\inf_{\tau\in\Upsilon} ||\mu(\tau)|| \to \infty$,

$$\mathbb{P}\{T^{QR}_{\infty}(\iota_q) > (T^{QR}_{\infty})^{(k^*)}\} \geq \mathbb{P}\{T^{QR}_{\infty}(\iota_q) > \max_{g\in\mathbf{G}_s} T^{QR}_{\infty}(g)\} \to 1.$$

Therefore, it suffices to establish (77). Note $T^{QR}_{\infty}(\iota_q) \geq \inf_{\tau\in\Upsilon} ||\mu(\tau)||_{A_r(\tau)} - O_p(1)$ and

$$\max_{g\in\mathbf{G}_s} T^{QR}_{\infty}(g) \leq \sup_{\tau\in\Upsilon, g\in\mathbf{G}_s} |\sum_{j\in J} g_j\xi_j a_j(\tau)| \sup_{\tau\in\Upsilon} ||\mu(\tau)||_{A_r(\tau)} + O_p(1).$$

Because when $g \in \mathbf{G}_s$, the signs of $\{g_j\}_{j\in J_s}$ cannot be the same. In addition, we have $\sum_{j\in J} \xi_j a_j(\tau) = \sum_{j\in J_s} \xi_j a_j(\tau) = 1$, and $\inf_{\tau\in\Upsilon, j\in J_s} a_j(\tau) \geq c > 0$. These imply

$$\max_{g\in\mathbf{G}_s, \tau\in\Upsilon} |\sum_{j\in J} g_j\xi_j a_j(\tau)| \leq 1 - 2\min_{j\in J} \xi_j c < 1.$$

Then, as $\inf_{\tau\in\Upsilon} ||\mu(\tau)||_2 \to \infty$, we have

$$\begin{aligned}
&\inf_{\tau\in\Upsilon} ||\mu(\tau)||_{A_r(\tau)} - (1 - 2\min_{j\in J}\xi_j) \sup_{\tau\in\Upsilon} ||\mu(\tau)||_{A_r(\tau)} \\
&\geq (\inf_{\tau\in\Upsilon} \lambda_{\min}(A_r(\tau)) - (1 - 2\min_{j\in J}\xi_j c) \inf_{\tau\in\Upsilon} \lambda_{\min}(A_r(\tau))c_\mu) \inf_{\tau\in\Upsilon} ||\mu(\tau)||_2 \to \infty
\end{aligned}$$

This concludes the proof. ■

S.F Proof of Theorem 3.2

Recall $\tilde{\Omega}(\tau)$ defined in (75),

$$T^{QR}_{CR,n} = \sup_{\tau\in\Upsilon} ||\sqrt{n}(\lambda^\top_\beta(\tau)\hat{\beta}(\tau) - \lambda_0(\tau))||_{\hat{A}_r(\tau)}$$

and

$$T^{QR*}_{CR,n}(g) = \sup_{\tau\in\Upsilon} ||\sqrt{n}(\lambda^\top_\beta(\tau)(\hat{\beta}^*_g(\tau) - \hat{\beta}(\tau)))||_{\hat{A}_{r,CR,g}(\tau)}.$$

Following the proof of Theorem 3.1, we have

$$\sqrt{n}(\lambda^\top_\beta(\tau)\hat{\beta}(\tau) - \lambda_0(\tau)) = \lambda^\top_\beta(\tau)\tilde{\Omega}(\tau)\left[\sum_{j\in J}\sqrt{\xi_j}\mathcal{Z}_j\right] + \mu(\tau) + o_p(1)$$

and

$$
\begin{aligned}
&\sqrt{n}(\lambda_\beta^\top(\tau)(\hat{\beta}_g^*(\tau) - \hat{\beta}(\tau))) \\
&= \lambda_\beta^\top(\tau)\tilde{\Omega}(\tau)\left[\sum_{j\in J} g_j \sqrt{\xi_j}\mathcal{Z}_j(\tau)\right] - \lambda_\beta^\top(\tau)\sum_{j\in J}\Omega(\tau)\xi_j g_j \mathcal{J}_{\pi,\beta,j}(\tau)\sqrt{n}(\hat{\beta}^r(\tau) - \beta_n(\tau)) + o_p(1) \\
&= \lambda_\beta^\top(\tau)\tilde{\Omega}(\tau)\left[\sum_{j\in J} g_j \sqrt{\xi_j}\mathcal{Z}_j(\tau)\right] + \overline{a}_g^*(\tau)\mu(\tau) + o_p(1),
\end{aligned}
$$

where the $o_p(1)$ terms in these two displays hold uniformly over $\tau \in \Upsilon$.

Next, we derive the limits of $\hat{A}_{r,CR}^{-1}(\tau)$ and $(\hat{A}_{r,CR,g}^*)^{-1}(\tau)$. Note

$$
\begin{aligned}
&\hat{A}_{r,CR}^{-1}(\tau) \\
&= \lambda_\beta^\top(\tau)\hat{\Omega}(\tau)\hat{V}(\tau,\tau')\hat{\Omega}^\top(\tau)\lambda_\beta(\tau) \\
&= \sum_{j\in J}\xi_j n_j \lambda_\beta^\top(\tau)\hat{\Omega}(\tau)\left[\mathbb{P}_{n,j}\hat{f}_\tau(D_{i,j},\hat{\beta}(\tau),\hat{\gamma}(\tau),0)\right]\left[\mathbb{P}_{n,j}\hat{f}_{\tau'}(D_{i,j},\hat{\beta}(\tau'),\hat{\gamma}(\tau'),0)\right]\hat{\Omega}^\top(\tau)\lambda_\beta(\tau).
\end{aligned}
$$

By Lemma S.K.6, we have

$$
\begin{aligned}
&\sqrt{n_j}\lambda_\beta^\top(\tau)\hat{\Omega}(\tau)\left[\mathbb{P}_{n,j}\hat{f}_\tau(D_{i,j},\hat{\beta}(\tau),\hat{\gamma}(\tau),0)\right] \\
&= \lambda_\beta^\top(\tau)\left[\tilde{\Omega}(\tau)\sqrt{n_j}\mathbb{P}_{n,j}\tilde{f}_\tau(D_{i,j},\beta_n(\tau),\gamma_n(\tau),0) - a_j(\tau)\tilde{\Omega}(\tau)\sqrt{n_j}\mathbb{P}_n\tilde{f}_\tau(D_{i,j},\beta_n(\tau),\gamma_n(\tau),0)\right] + o_p(1) \\
&= \lambda_\beta^\top(\tau)\tilde{\Omega}(\tau)\left[\mathcal{Z}_j(\tau) - a_j(\tau)\sqrt{\xi_j}\sum_{\tilde{j}\in J}\xi_{\tilde{j}}\mathcal{Z}_{\tilde{j}}(\tau)\right] + o_p(1),
\end{aligned}
$$

where the $o_p(1)$ term holds uniformly over $\tau \in \Upsilon$. Because $q > d_r$,

$$
\inf_{\tau\in\Upsilon}\lambda_{\min}\begin{pmatrix}\lambda_\beta^\top(\tau)\tilde{\Omega}(\tau)\sum_{j\in J}\xi_j\left[\mathcal{Z}_j(\tau) - a_j(\tau)\sqrt{\xi}_j\sum_{\tilde{j}\in J}\sqrt{\xi}_{\tilde{j}}\mathcal{Z}_{\tilde{j}}(\tau)\right] \\ \times\left[\mathcal{Z}_j(\tau) - a_j(\tau)\sqrt{\xi}_j\sum_{\tilde{j}\in J}\sqrt{\xi}_{\tilde{j}}\mathcal{Z}_{\tilde{j}}(\tau)\right]^\top\tilde{\Omega}^\top(\tau)\lambda_\beta(\tau)\end{pmatrix} > 0
$$

with probability one, we have

$$
\begin{aligned}
\hat{A}_{r,CR}(\tau) = &\begin{pmatrix}\lambda_\beta^\top(\tau)\tilde{\Omega}(\tau)\sum_{j\in J}\xi_j\left[\mathcal{Z}_j(\tau) - a_j(\tau)\sqrt{\xi}_j\sum_{\tilde{j}\in J}\sqrt{\xi}_{\tilde{j}}\mathcal{Z}_{\tilde{j}}(\tau)\right] \\ \times\left[\mathcal{Z}_j(\tau) - a_j(\tau)\sqrt{\xi}_j\sum_{\tilde{j}\in J}\sqrt{\xi}_{\tilde{j}}\mathcal{Z}_{\tilde{j}}(\tau)\right]^\top\tilde{\Omega}^\top(\tau)\lambda_\beta(\tau)\end{pmatrix}^{-1} + o_p(1) \\
\equiv &A_{r,CR}(\tau) + o_p(1).
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
(\hat{A}_{r,CR,g}^*)^{-1}(\tau) &= \lambda_\beta^\top(\tau)\hat{\Omega}(\tau)\hat{V}_g^*(\tau,\tau')\hat{\Omega}^\top(\tau)\lambda_\beta(\tau) \\
&= \sum_{j\in J}\xi_j n_j \lambda_\beta^\top(\tau)\hat{\Omega}(\tau)\left[\mathbb{P}_{n,j}\overline{f}_{\tau,g}^*(D_{i,j})\right]\left[\mathbb{P}_{n,j}\overline{f}_{\tau,g}^*(D_{i,j})\right]^\top\hat{\Omega}^\top(\tau)\lambda_\beta(\tau).
\end{aligned}
$$

Furthermore, Lemma S.K.6 shows

$$
\begin{aligned}
&\sqrt{n_j}\lambda_\beta^\top(\tau)\hat{\Omega}(\tau)\left[\mathbb{P}_{n,j}\overline{f}^*_{\tau,g}(D_{i,j})\right] \\
&= \sqrt{n_j}\lambda_\beta^\top(\tau)\Bigg[g_j\tilde{\Omega}(\tau)\mathbb{P}_{n,j}\tilde{f}_\tau(D_{i,j},\beta_n(\tau),\gamma_n(\tau),0) - g_j a_j(\tau)(\hat{\beta}^r(\tau)-\beta_n(\tau)) \\
&- \tilde{\Omega}(\tau)a_j(\tau)\sum_{\tilde{j}\in J} g_{\tilde{j}}\xi_{\tilde{j}}\mathbb{P}_{n,j}\tilde{f}_\tau(D_{i,j},\beta_n(\tau),\gamma_n(\tau),0) + a_j(\tau)\overline{a}_g^*(\tau)(\hat{\beta}^r(\tau)-\beta_n(\tau))\Bigg] + o_p(1) \\
&= \lambda_\beta^\top(\tau)\tilde{\Omega}(\tau)\left[g_j\mathcal{Z}_j(\tau) - a_j(\tau)\sqrt{\xi_j}\sum_{\tilde{j}\in J} g_{\tilde{j}}\sqrt{\xi_{\tilde{j}}}\mathcal{Z}_{\tilde{j}}(\tau)\right] + \sqrt{\xi}_j(g_j-\overline{a}_g^*(\tau))a_j(\tau)\mu(\tau) + o_p(1),
\end{aligned}
$$

where the $o_p(1)$ term holds uniformly over $\tau\in\Upsilon$, $\overline{a}_g^*(\tau) = \sum_{j\in J}\xi_j g_j a_j(\tau)$, and the second equality holds because

$$
-\sqrt{n_j}\lambda_\beta^\top(\tau)\left(g_j - \overline{a}_g^*(\tau)\right)a_j(\tau)(\hat{\beta}^r(\tau)-\beta_n(\tau)) = \sqrt{\xi}_j(g_j-\overline{a}_g^*(\tau))a_j(\tau)\mu(\tau).
$$

Then, as $q > d_r$,

$$
\hat{A}^*_{r,CR,g}(\tau) = A_{r,CR,g}(\tau) + o_p(1),
$$

where the $o_p(1)$ term holds uniformly over $\tau\in\Upsilon$ and

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

Let

$$
T^{QR}_{CR,\infty}(g) = \sup_{\tau\in\Upsilon}\left[\left\|\lambda_\beta^\top(\tau)\tilde{\Omega}(\tau)\left[\sum_{j\in J} g_j\sqrt{\xi}_j\mathcal{Z}_j(\tau) + \sum_{j\in J} g_j\xi_j a_j(\tau)\mu(\tau)\right]\right\|_{A_{r,CR,g}(\tau)}\right].
$$

Because $\overline{a}^*_{\iota_q}(\tau) = 1$ and $\overline{a}^*_{-\iota_q}(\tau) = -1$, we have,

$$
\{T^{QR}_{CR,n}, \{T^{QR*}_{CR,n}(g)\}_{g\in\mathbf{G}} \rightsquigarrow \{T^{QR}_{CR,\infty}(\iota_q), \{T^{QR}_{CR,\infty}(g)\}_{g\in\mathbf{G}}\}.
$$

In addition, because $q > d_r$, under the null, $T^{QR}_{CR,\infty}(g) \overset{d}{=} T^{QR}_{CR,\infty}(g')$ for any $g, g' \in G$ and $T^{QR}_{CR,\infty}(g) = T^{QR}_{CR,\infty}(g')$ if and only if $g\in\{g', -g'\}$. Then following the exact same argument in the proof of Theorem 3.1, we have

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

For the power analysis, we still have

$$
\liminf_{n\to\infty}\mathbb{P}\{T^{QR}_{CR,n} > \hat{c}^{QR}_{n}\} \ge \mathbb{P}\left\{T^{QR}_{CR,\infty}(\iota_q) > (T^{QR}_{CR,\infty})^{(k^*)}\right\}.
$$

In addition, we aim to show that, as $\inf_{\tau\in\Upsilon}||\mu(\tau)||\to\infty$, we have

$$\mathbb{P}\left\{T_{CR,\infty}^{QR}(\iota_q)>\max_{g\in\mathbf{G}_s}T_{CR,\infty}^{QR}(g)\right\}\to 1. \tag{78}$$

Then, given $|\mathbf{G}_s|=|\mathbf{G}|-2^{q-q_s+1}$ and $\lceil|\mathbf{G}|(1-\alpha)\rceil\leq|\mathbf{G}|-2^{q-q_s+1}$, (78) implies, $\inf_{\tau\in\Upsilon}||\mu(\tau)||\to\infty$,

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

Therefore, it suffices to establish (78). Note that

$$T_{CR,\infty}^{QR}(\iota_q)=\sup_{\tau\in\Upsilon}\left[\left\|\lambda_\beta^\top(\tau)\tilde{\Omega}(\tau)\left[\sum_{j\in J}\sqrt{\xi}_j\mathcal{Z}_j(\tau)\right]+\mu(\tau)\right\|_{A_{r,CR,\iota_q}(\tau)}\right],$$

where

$$A_{r,CR,\iota_q}(\tau)=\begin{pmatrix}\sum_{j\in J}\xi_j\left\{\lambda_\beta^\top(\tau)\tilde{\Omega}(\tau)\left[\mathcal{Z}_j(\tau)-a_j(\tau)\sqrt{\xi}_j\sum_{\tilde{j}\in J}\sqrt{\xi}_{\tilde{j}}\mathcal{Z}_{\tilde{j}}(\tau)\right]\right\}\\ \times\left\{\lambda_\beta^\top(\tau)\tilde{\Omega}(\tau)\left[\mathcal{Z}_j(\tau)-a_j(\tau)\sqrt{\xi}_j\sum_{\tilde{j}\in J}\sqrt{\xi}_{\tilde{j}}\mathcal{Z}_{\tilde{j}}(\tau)\right]\right\}^\top\end{pmatrix}^{-1}.$$

We see that $A_{r,CR,\iota_q}(\tau)$ is independent of μ. In addition, with probability one, we have

$$\inf_{\tau\in\Upsilon}\lambda_{\min}(\tilde{\Omega}^\top(\tau)\lambda_\beta^\top(\tau)A_{r,CR,\iota_q}(\tau)\lambda_\beta(\tau)\tilde{\Omega}(\tau))>0.$$

Therefore, for any $e>0$, we can find a constant $c>0$ such that with probability greater than $1-e$,

$$T_{CR,\infty}^{QR}(\iota_q)\geq c\inf_{\tau\in\Upsilon}||\mu(\tau)||^2-O_p(1). \tag{79}$$

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

$$\begin{aligned}T_{CR,\infty}^{QR}(g)=\sup_{\tau\in\Upsilon}\Bigg\{&(N_{0,g}(\tau)+c_{0,g}(\tau)\mu(\tau))^\top\left[\sum_{j\in J}\xi_j(N_{j,g}(\tau)+c_{j,g}(\tau)\mu(\tau))(N_{j,g}(\tau)+c_{j,g}(\tau)\mu(\tau))^\top\right]^{-1}\\ &\times(N_{0,g}(\tau)+c_{0,g}(\tau)\mu(\tau))\Bigg\},\end{aligned}$$

where

$$\begin{aligned}&N_{0,g}(\tau)=\lambda_\beta^\top(\tau)\tilde{\Omega}(\tau)\left[\sum_{j\in J}g_j\sqrt{\xi}_j\mathcal{Z}_j(\tau)\right],\\ &N_{j,g}(\tau)=\lambda_\beta^\top(\tau)\tilde{\Omega}(\tau)\left[g_j\mathcal{Z}_j(\tau)-a_j(\tau)\sqrt{\xi}_j\sum_{\tilde{j}\in J}\sqrt{\xi}_{\tilde{j}}g_{\tilde{j}}\mathcal{Z}_{\tilde{j}}(\tau)\right],\quad j\in J,\\ &c_{0,g}(\tau)=\overline{a}_g^*(\tau),\quad\text{and}\quad c_{j,g}(\tau)=\sqrt{\xi}_j(g_j-\overline{a}_g^*(\tau))a_j(\tau),\quad j\in J.\end{aligned}$$

We claim that for $g\in\mathbf{G}_s$, $c_{j,g}(\tau)\neq 0$ for some $j\in J_s$. To see this claim, suppose it does not hold. Then, it implies $g_j = \overline{a}_g^*(\tau)$ for all $j \in J_s$, i.e., for all $j \in J_s$, g_j shares the same sign. This contradicts with the definition of $\mathbf{G}_s$. This claim and the fact that $\inf_{\tau \in \Upsilon, j \in J_s} |a_j(\tau)| \geq c_0 > 0$ further imply that

$$
\begin{aligned}
&\inf_{\tau \in \Upsilon, g \in \mathbf{G}_s} \sum_{j \in J} \xi_j c_{j,g}^2(\tau) \\
&\geq (c_0^2 \min_{j \in J} \xi_j^2) \max((1-\overline{a}_g^*(\tau))^2, (1+\overline{a}_g^*(\tau))^2) \\
&\geq \frac{c_0^2 \min_{j \in J} \xi_j^2}{2} \left[(1-\overline{a}_g^*(\tau))^2 + (1+\overline{a}_g^*(\tau))^2\right] \geq c_0^2 \min_{j \in J} \xi_j^2 > 0.
\end{aligned}
$$

In addition, we have

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

where we denote

$$
M_1 = \sum_{j \in J} \xi_j N_{j,g}(\tau) N_{j,g}^\top(\tau), \quad M_2 = \sum_{j \in J} \xi_j c_{j,g}(\tau) N_{j,g}(\tau), \quad \text{and} \quad \overline{c}^2 = \sum_{j \in J} \xi_j c_{j,g}^2(\tau) > 0.
$$

For notation ease, we suppress the dependence of $(M_1, M_2, \overline{c})$ on (g, τ). Then, we have

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

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

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

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

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

which has probability zero if $q > d_r$. Therefore, the matrix $\mathbb{M} \equiv M_1 - \frac{M_2 M_2^\top}{\overline{c}^2}$ is invertible with probability one. Specifically, denote $\mathbb{M}$ as $\mathbb{M}_g(\tau)$ to highlight its dependence on (g, τ). We have $\sup_{(g,\tau) \in \mathbf{G}_s \times \Upsilon} (\lambda_{\min}(\mathbb{M}(g, \tau)))^{-1} = O_p(1)$. In addition, denote $\frac{M_2}{\overline{c}} + \overline{c}\mu(\tau)$ as $\mathbb{V}$, which is a $d_r \times 1$

vector. Then, we have

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

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

Next, we note that

$$
\begin{aligned}
N_{0,g}(\tau)+c_{0,g}(\tau)\mu(\tau) &= N_{0,g}(\tau)+c_{0,g}(\tau)\left(\frac{\mathbb{V}}{\overline{c}}-\frac{M_2}{\overline{c}^2}\right)\\
&\equiv \mathbb{M}_0+\frac{c_{0,g}(\tau)}{\overline{c}}\mathbb{V},
\end{aligned}
$$

where

$$
\mathbb{M}_0 = N_{0,g}(\tau)-\frac{c_{0,g}(\tau)M_2}{\overline{c}^2} = N_{0,g}(\tau)-\frac{c_{0,g}(\tau)(\sum_{j\in J}\xi_j c_{j,q}(\tau)N_{j,g}(\tau))}{\sum_{j\in J}\xi_j c_{j,g}^2(\tau)}.
$$

With these notations, we have

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

where the first inequality is due to the fact that $(u+v)^\top A(u+v)\le 2(u^\top Au+v^\top Av)$ 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 $\overline{c}$ and denoting $\mathbb{M}$ as $\mathbb{M}(g,\tau)$. By taking the supremum over $\tau \in \Upsilon$ first followed by $g \in \mathbf{G}_s$, we have

$$\max_{g \in \mathbf{G}_s} T^{QR}_{CR,\infty}(g) \leq \sup_{(g,\tau)\in \mathbf{G}_s \times \Upsilon} C(g,\tau) = O_p(1). \tag{80}$$

Combining (79) and (80), we have, as $\inf_{\tau\in\Upsilon} ||\mu(\tau)|| \to \infty$,

$$\begin{aligned}
&\mathbb{P}\left\{ T^{QR}_{CR,\infty}(\iota_q) > \max_{g\in\mathbf{G}_s} T^{QR}_{CR,\infty}(g) \right\} \\
&\geq \mathbb{P}\left\{ c \inf_{\tau\in\Upsilon} ||\mu(\tau)||^2 - O_p(1) > \sup_{(g,\tau)\in\mathbf{G}_s\times\Upsilon} C(g,\tau) \right\} - e \to 1 - e.
\end{aligned}$$

As e is arbitrary, we have established (78). This concludes the proof. ■

S.G Proof of Theorem 3.3

We focus on the case when Assumption 12(ii) holds. The proof for the case with Assumption 12(i) is similar but simpler, and thus, is omitted for brevity. We divide the proof into three steps. In the first step, we derive the limit distribution of $AR^{QR}_{CR,n}$. In the second step, we derive the limit distribution of $AR^{QR*}_{CR,n}(g)$. In the third step, we prove the desired result. Throughout the proof, we impose the null that $\beta_n(\tau) = \beta_0(\tau)$ for $\tau \in \Upsilon$.

Step 1. By Lemma S.K.1 with $b_n(\tau) = \beta_n(\tau) = \beta_0(\tau)$, we have

$$\begin{aligned}
\sqrt{n}\hat{\theta}(\beta_0(\tau),\tau) &= \omega \left[\Gamma_1(\beta_0(\tau),\tau)\right]^{-1} \sqrt{n}\left[I_n(\tau) + II_n(\beta_0(\tau),\tau) + o_p(1) - \Gamma_2(\beta_0(\tau),\tau)(\beta_0(\tau) - \beta_n(\tau))\right] \\
&= \omega \left[\mathcal{J}_{\pi,\pi}(\tau)\right]^{-1} \sqrt{n} I_n(\tau) + o_p(1), \qquad (81)
\end{aligned}$$

where the $o_p(1)$ terms hold uniformly over $\tau \in \Upsilon$ and

$$I_n(\tau) = (\mathbb{P}_n - \overline{\mathbb{P}}_n) f_\tau(D, \beta_0(\tau), \gamma_n(\tau), 0).$$

In addition, note

$$\begin{aligned}
\hat{A}^{-1}_{CR}(\tau) = \sum_{j\in J} \xi_j n_j \omega \hat{\mathcal{J}}^{-1}_{\pi,\pi}(\tau) &\left[\mathbb{P}_{n,j} \hat{f}_\tau(D_{i,j}, \beta_0(\tau), \hat{\gamma}(\beta_0(\tau),\tau))\right] \\
\times &\left[\mathbb{P}_{n,j} \hat{f}_\tau(D_{i,j}, \beta_0(\tau), \hat{\gamma}(\beta_0(\tau),\tau))\right]^\top \hat{\mathcal{J}}^{-1}_{\pi,\pi}(\tau)\omega^\top
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
&\mathbb{P}_{n,j} \hat{f}_\tau(D_{i,j}, \beta_0(\tau), \hat{\gamma}(\beta_0(\tau),\tau), 0) \\
&= (\mathbb{P}_{n,j} - \overline{\mathbb{P}}_{n,j}) \hat{f}_\tau(D_{i,j}, \beta_0(\tau), \hat{\gamma}(\beta_0(\tau),\tau), 0) + \overline{\mathbb{P}}_{n,j} \hat{f}_\tau(D_{i,j}, \beta_0(\tau), \hat{\gamma}(\beta_0(\tau),\tau), 0) \\
&= (\mathbb{P}_{n,j} - \overline{\mathbb{P}}_{n,j}) f_\tau(D_{i,j}, \beta_0(\tau), \gamma_n(\tau),\tau), 0) + \overline{\mathbb{P}}_{n,j} \hat{f}_\tau(D_{i,j}, \beta_0(\tau), \hat{\gamma}(\beta_0(\tau),\tau), 0) + o_p(n^{-1/2})
\end{aligned}$$

$$= \mathbb{P}_{n,j} f_\tau(D_{i,j}, \beta_0(\tau), \gamma_n(\tau), 0) - \mathcal{J}_{\pi,\pi,j}(\tau) \begin{pmatrix} \hat{\gamma}(\beta_0(\tau), \tau) - \gamma_n(\tau) \\ 0 \end{pmatrix} + o_p(n^{-1/2})$$
$$= O_p(n^{-1/2}),$$

where both the $O_p(n^{-1/2})$ and $o_p(n^{-1/2})$ terms hold uniformly over $\tau \in \Upsilon$, the second equality is due to Assumption 6(ii), the third equality is due to Assumption 6(iii) and the fact that

$$\sup_{\tau \in \Upsilon} ||\hat{\gamma}(\beta_0(\tau), \tau) - \gamma_n(\tau)||_2 = O_p(n^{-1/2})$$

as shown in Lemma S.K.3, and the last equality is due to Lemma S.K.3.

Recall $\omega = (0_{d_\phi \times d_w}, \mathbb{I}_{d_\phi \times d_\phi})$. Then, by Lemma S.K.1, we have

$$\omega \hat{\mathcal{J}}_{\pi,\pi}^{-1}(\tau) \left[\mathbb{P}_{n,j} \hat{f}_\tau(D_{i,j}, \beta_0(\tau), \hat{\gamma}(\beta_0(\tau), \tau))\right]$$
$$= \omega \mathcal{J}_{\pi,\pi}^{-1}(\tau) \left[\mathbb{P}_{n,j} f_\tau(D_{i,j}, \beta_0(\tau), \gamma_n(\tau), 0) - \mathcal{J}_{\pi,\pi,j}(\tau) \begin{pmatrix} \hat{\gamma}(\beta_0(\tau), \tau) - \gamma_n(\tau) \\ 0 \end{pmatrix}\right] + o_p(n^{-1/2})$$
$$= \mathcal{J}_{\theta,\theta}^{-1}(\tau) \left[\mathbb{P}_{n,j} \tilde{f}_\tau(D_{i,j}, \beta_0(\tau), \gamma_n(\tau), 0)\right] + o_p(n^{-1/2}),$$

where the $o_p(n^{-1/2})$ terms hold uniformly over $\tau \in \Upsilon$ and the second inequality holds because under Assumption 7(i),

$$\omega \hat{\mathcal{J}}_{\pi,\pi}^{-1}(\tau) \mathcal{J}_{\pi,\pi,j}(\tau) \begin{pmatrix} \hat{\gamma}(\beta_n(\tau), \tau) - \gamma_n(\tau) \\ 0 \end{pmatrix} = 0.$$

Then we have

$$\sqrt{n} \hat{\theta}(\beta_0(\tau), \tau) = \mathcal{J}_{\theta,\theta}^{-1}(\tau) \sum_{j \in J} \sqrt{\xi}_j \mathcal{Z}_j(\tau) + o_p(1), \tag{82}$$

$$\sqrt{n_j} \omega \hat{\mathcal{J}}_{\pi,\pi}^{-1}(\tau) \left[\mathbb{P}_{n,j} \hat{f}_\tau(D_{i,j}, \beta_n(\tau), \hat{\gamma}(\beta_n(\tau), \tau))\right] = \mathcal{J}_{\theta,\theta}^{-1}(\tau) \mathcal{Z}_j(\tau) + o_p(1),$$

and

$$\hat{A}_{CR}^{-1}(\tau) = \sum_{j \in J} \xi_j \mathcal{J}_{\theta,\theta}^{-1}(\tau) \mathcal{Z}_j(\tau) \mathcal{Z}_j^\top(\tau) \mathcal{J}_{\theta,\theta}^{-1}(\tau) + o_p(1),$$

where the $o_p(1)$ term hold uniformly over $\tau \in \Upsilon$. When $q > d_\phi$, $\sum_{j \in J} \xi_j \mathcal{J}_{\theta,\theta}^{-1}(\tau) \mathcal{Z}_j(\tau) \mathcal{Z}_j^\top(\tau) \mathcal{J}_{\theta,\theta}^{-1}(\tau)$ is invertible with probability one, we have

$$\hat{A}_{CR}(\tau) = \left[\sum_{j \in J} \xi_j \mathcal{J}_{\theta,\theta}^{-1}(\tau) \mathcal{Z}_j(\tau) \mathcal{Z}_j^\top(\tau) \mathcal{J}_{\theta,\theta}^{-1}(\tau)\right]^{-1} + o_p(1), \tag{83}$$

where the $o_p(1)$ term hold uniformly over $\tau \in \Upsilon$. Then, combining (81), (82), and (83), we have

$$
\begin{aligned}
&n(AR^{QR}_{CR,n})^2\\
&= \sup_{\tau\in\Upsilon}\left[\sum_{j\in J}\sqrt{\xi}_j\mathcal{Z}_j^\top(\tau)\right]\mathcal{J}_{\theta,\theta}^{-1}(\tau)\left[\sum_{j\in J}\xi_j\mathcal{J}_{\theta,\theta}^{-1}(\tau)\mathcal{Z}_j(\tau)\mathcal{Z}_j^\top(\tau)\mathcal{J}_{\theta,\theta}^{-1}(\tau)\right]^{-1}\mathcal{J}_{\theta,\theta}^{-1}(\tau)\left[\sum_{j\in J}\sqrt{\xi}_j\mathcal{Z}_j(\tau)\right]\\
&+ o_p(1).
\end{aligned}
$$

Step 2. Next, we consider the limit distribution of the bootstrap test statistic. By the sub-gradient condition of (34), we have

$$
o_p(1/\sqrt{n}) = \mathbb{P}_n\hat{f}_\tau(D_{i,j},\beta_0(\tau),\tilde{\gamma}_g^*(\tau),\tilde{\theta}_g^*(\tau)) + \frac{1}{n}\sum_{j\in J}g_j\sum_{i\in I_{n,j}}\hat{f}_\tau(D_{i,j},\beta_0(\tau),\hat{\gamma}(\beta_0(\tau),\tau),0), \tag{84}
$$

where the $o_p(1/\sqrt{n})$ term on the LHS of the above display holds uniformly over $\tau\in\Upsilon$. In addition,

$$
\begin{aligned}
&\mathbb{P}_n\hat{f}_\tau(D_{i,j},\beta_0(\tau),\tilde{\gamma}_g^*(\tau),\tilde{\theta}_g^*(\tau))\\
&= I_n(\tau) + II_{n,g}(\beta_0(\tau),\tau) - \Gamma_{1,g}(\beta_0(\tau),\tau)\begin{pmatrix}\tilde{\gamma}_g^*(\tau)-\gamma_n(\tau)\\ \tilde{\theta}_g^*(\tau)\end{pmatrix},
\end{aligned} \tag{85}
$$

where

$$
I_n(\tau) = (\mathbb{P}_n - \overline{\mathbb{P}}_n)f_\tau(D_{i,j},\beta_0(\tau),\gamma_n(\tau),0),
$$

$$
II_{n,g}(\beta_0(\tau),\tau) = \sum_{j\in J}\xi_j(\mathbb{P}_{n,j}-\overline{\mathbb{P}}_{n,j})\left(\hat{f}_\tau(D_{i,j},\beta_0(\tau),\tilde{\gamma}_g^*(\tau),\tilde{\theta}_g^*(\tau)) - f_\tau(D_{i,j},\beta_0(\tau),\gamma_n(\tau),0)\right)
$$

$$
\Gamma_{1,g}(\beta_0(\tau),\tau) = \sum_{j\in J}\frac{\xi_j}{n_j}\sum_{i\in I_{n,j}}\mathbb{E}f_{\varepsilon_{i,j}(\tau)}(\hat{\delta}_{i,j,g}(\tau)|W_{i,j},Z_{i,j})\hat{\Psi}_{i,j}(\tau)\hat{\Psi}_{i,j}^\top(\tau),
$$

and

$$
\hat{\delta}_{i,j,g}(\tau)\in(0, W_{i,j}^\top(\tilde{\gamma}_g^*(\tau)-\gamma_n(\tau)) + \hat{\Phi}_{i,j}^\top(\tau)\tilde{\theta}_g^*(\tau)).
$$

Following the same argument in Step 1 of the proof of Lemma S.K.4, we have

$$
\sup_{\tau\in\Upsilon}\left(||\tilde{\gamma}_g^*(\tau)-\gamma_n(\tau)||_2 + ||\tilde{\theta}_g^*(\tau)||_2\right) = o_p(1),
$$

$$
\sup_{\tau\in\Upsilon}\sqrt{n}||II_{n,g}(\beta_0(\tau),\tau)||_2 = o_p(1),
$$

and

$$
\sup_{\tau\in\Upsilon,j\in J}\left\|\frac{1}{n_j}\sum_{i\in I_{n,j}}\mathbb{E}f_{\varepsilon_{i,j}(\tau)}(\hat{\delta}_{i,j,g}(\tau)|W_{i,j},Z_{i,j})\hat{\Psi}_{i,j}(\tau)\hat{\Psi}_{i,j}^\top(\tau) - \mathcal{J}_{\pi,\pi}(\tau)\right\|_{op} = o_p(1).
$$

Note here, to derive the above three displays, we do not need to impose Assumption 8(i)–8(iii) as did in Lemma S.K.4 because $\beta_0(\tau) = \beta_n(\tau)$ under the null. In addition, we have

$$
\begin{aligned}
&\frac{1}{n}\sum_{j\in J} g_j \sum_{i\in I_{n,j}} \hat{f}_\tau(D_{i,j}, \beta_0(\tau), \hat{\gamma}(\beta_0(\tau),\tau), 0) \\
&= \frac{1}{n}\sum_{j\in J} g_j \sum_{i\in I_{n,j}} \Big(\hat{f}_\tau(D_{i,j}, \beta_0(\tau), \hat{\gamma}(\beta_0(\tau),\tau), 0) - f_\tau(D_{i,j}, \beta_0(\tau), \gamma_n(\tau), 0)\Big) \\
&+ \frac{1}{n}\sum_{j\in J} g_j \sum_{i\in I_{n,j}} f_\tau(D_{i,j}, \beta_0(\tau), \gamma_n(\tau), 0) \\
&= \frac{1}{n}\sum_{j\in J} g_j \left[\sum_{i\in I_{n,j}} f_\tau(D_{i,j}, \beta_0(\tau), \gamma_n(\tau), 0) - \mathcal{J}_{\pi,\pi,j}(\tau)\begin{pmatrix}\hat{\gamma}(\beta_0(\tau),\tau) - \gamma_n(\tau)\\ 0\end{pmatrix}\right] + o_p(n^{-1/2}).
\end{aligned}
$$

where the $o_p(n^{-1/2})$ term holds uniformly over $\tau \in \Upsilon$. Therefore, we have

$$
\begin{aligned}
&\begin{pmatrix}\tilde{\gamma}_g^*(\tau) - \gamma_n(\tau)\\ \tilde{\theta}_g^*(\tau)\end{pmatrix} \\
&= \mathcal{J}_{\pi,\pi}^{-1}(\tau)\left[I_n(\tau) + \frac{1}{n}\sum_{j\in J} g_j \left[\sum_{i\in I_{n,j}} f_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0) - \mathcal{J}_{\pi,\pi,j}(\tau)\begin{pmatrix}\hat{\gamma}(\beta_0(\tau),\tau) - \gamma_n(\tau)\\ 0\end{pmatrix}\right]\right] \\
&+ o_p(n^{-1/2})
\end{aligned}
$$

and

$$
\sqrt{n}\tilde{\theta}_g^*(\tau) = \omega \mathcal{J}_{\pi,\pi}^{-1}(\tau)\left[\frac{1}{\sqrt{n}}\sum_{j\in J}(1+g_j)\sum_{i\in I_{n,j}} f_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0)\right] + o_p(1),
$$

where the $o_p(1)$ term holds uniformly over $\tau \in \Upsilon$ and the second display holds because under Assumption 7(i), we have

$$
\omega \mathcal{J}_{\pi,\pi}^{-1}(\tau)\mathcal{J}_{\pi,\pi,j}(\tau)\begin{pmatrix}\hat{\gamma}(\beta_0(\tau),\tau) - \gamma_n(\tau)\\ 0\end{pmatrix} = 0.
$$

Therefore, we have

$$
\begin{aligned}
\sqrt{n}(\tilde{\theta}_g^*(\tau) - \hat{\theta}(\beta_0(\tau),\tau)) &= \omega \mathcal{J}_{\pi,\pi}^{-1}(\tau)\left[\frac{1}{\sqrt{n}}\sum_{j\in J} g_j \sum_{i\in I_{n,j}} f_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0)\right] + o_p(1) \\
&= \mathcal{J}_{\theta,\theta}^{-1}(\tau)\left[\frac{1}{\sqrt{n}}\sum_{j\in J} g_j \sum_{i\in I_{n,j}} \tilde{f}_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0)\right] + o_p(1),
\end{aligned}
$$

where the $o_p(1)$ term holds uniformly over $\tau \in \Upsilon$.

Step 3. We further define

$$
\begin{aligned}
&AR^{QR}_{CR,\infty}(g) \\
&= \sup_{\tau\in\Upsilon}\left\{\sum_{j\in J}\sqrt{\xi}_j g_j \mathcal{Z}_j(\tau)\mathcal{J}^{-1}_{\theta,\theta}(\tau)\left[\sum_{j\in J}\xi_j\mathcal{J}^{-1}_{\theta,\theta}(\tau)g_j\mathcal{Z}_j(\tau)g_j\mathcal{Z}^\top_j(\tau)\mathcal{J}^{-1}_{\theta,\theta}(\tau)\right]^{-1}\mathcal{J}^{-1}_{\theta,\theta}(\tau)\left[\sum_{j\in J}\sqrt{\xi}_j g_j \mathcal{Z}_j(\tau)\right]\right\}^{1/2}.
\end{aligned}
$$

Then, we have, under the null,

$$
(\sqrt{n}AR^{QR}_{CR,n}, \{\sqrt{n}AR^{QR*}_{CR,n}(g)\}_{g\in G}) \rightsquigarrow (AR^{QR}_{W,\infty}(\iota_q), \{AR^{QR}_{CR,\infty}(g)\}_{g\in \mathbf{G}}).
$$

The distribution of $AR^{QR}_{CR,\infty}(g)$ is invariant in g, and $AR^{QR}_{CR,\infty}(g) = AR^{QR}_{CR,\infty}(g')$ if and only if $g \in \{g', -g'\}$. Then, by the same argument in the proofs of Theorem 3.1, we have

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

S.H Proof of Propositions A.1 and A.2

For Proposition A.1, we have

$$
\begin{aligned}
&\sup\left\|\overline{\mathbb{P}}_{n,j} f_{\varepsilon_{i,j}(\tau)}(\delta_{i,j}(v,\tau)|W_{i,j},Z_{i,j})V^2_{i,j}(\tau)W_{i,j}\hat{\Phi}^\top_{i,j}(\tau)\right\|_{op} \\
&\leq \sup\left\|\overline{\mathbb{P}}_{n,j} f_{\varepsilon_{i,j}(\tau)}(\delta_{i,j}(v,\tau)|W_{i,j},Z_{i,j})V^2_{i,j}(\tau)W_{i,j}\underline{\hat{\Phi}}^\top_{i,j}(\tau) - \underline{\mathcal{J}}_{\gamma,\theta,j}(\tau)\right\|_{op} \\
&+ \sup\left\|\left[\overline{\mathbb{P}}_{n,j} f_{\varepsilon_{i,j}(\tau)}(\delta_{i,j}(v,\tau)|W_{i,j},Z_{i,j})V^2_{i,j}(\tau)W_{i,j}W^\top_{i,j}(\tau) - \mathcal{J}_{\gamma,\gamma,j}(\tau)\right]\hat{\chi}(\tau)\right\|_{op} \\
&+ \sup_{\tau\in\Upsilon}\|\mathcal{J}_{\gamma,\gamma,j}(\tau)(\hat{\chi}(\tau)-\chi(\tau))\|_{op} + \sup_{\tau\in\Upsilon}||\mathcal{J}_{\gamma,\gamma,j}(\tau)\chi(\tau) - \underline{\mathcal{J}}_{\gamma,\theta,j}(\tau)||_{op}
\end{aligned}
$$

where the suprema in the first three lines are taken over $\{j \in J, ||v||_2 \leq \delta, \tau \in \Upsilon\}$ for $v = (v_b^\top, v_r^\top, v_t^\top)^\top$. We note that, based on Assumption 13, by letting $n \to \infty$ followed by $\delta \to 0$,

$$
\begin{aligned}
&\sup\left\|\overline{\mathbb{P}}_{n,j} f_{\varepsilon_{i,j}(\tau)}(\delta_{i,j}(v,\tau)|W_{i,j},Z_{i,j})V^2_{i,j}(\tau)W_{i,j}\underline{\hat{\psi}}^\top_{i,j}(\tau) - \underline{\mathcal{J}}_{\gamma,\theta,j}(\tau)\right\|_{op} \xrightarrow{p} 0 \quad \text{and} \\
&\sup\left\|\left[\overline{\mathbb{P}}_{n,j} f_{\varepsilon_{i,j}(\tau)}(\delta_{i,j}(v,\tau)|W_{i,j},Z_{i,j})V^2_{i,j}(\tau)W_{i,j}W^\top_{i,j}(\tau) - \mathcal{J}_{\gamma,\gamma,j}(\tau)\right]\hat{\chi}(\tau)\right\|_{op} \xrightarrow{p} 0.
\end{aligned}
$$

In addition, Assumption 13 implies $\sup_{\tau\in\Upsilon}\|\hat{\chi}(\tau)-\chi(\tau)\| = o_p(1)$. Therefore, in order to show the result, it suffices to show

$$
\sup_{\tau\in\Upsilon}||\underline{\mathcal{J}}_{\gamma,\theta,j}(\tau) - \mathcal{J}_{\gamma,\gamma,j}(\tau)\chi(\tau)||_{op} = o(1).
$$

We note that

$$
\sup_{\tau\in\Upsilon}||\underline{\mathcal{J}}_{\gamma,\theta,j}(\tau) - \mathcal{J}_{\gamma,\gamma,j}(\tau)\chi(\tau)||_{op}
$$

$$= \sup_{\tau\in\Upsilon} ||\mathcal{J}_{\gamma,\gamma,j}(\tau)(\chi_{n,j}(\tau) - \chi(\tau))||_{op}$$
$$= \sup_{\tau\in\Upsilon} || \lim_{n\to\infty} \overline{\mathbb{P}}_{n,j} f_{\varepsilon_{i,j}(\tau)}(0|W_{i,j}, Z_{i,j}) W_{i,j} W_{i,j}^\top (\chi_{n,j}(\tau) - \chi(\tau))||_{op}$$
$$\leq \limsup_{n\to\infty} \sup_{\tau\in\Upsilon} \left[\overline{\mathbb{P}}_{n,j} ||f_{\varepsilon_{i,j}(\tau)}(0|W_{i,j}, Z_{i,j}) W_{i,j}||_2^2 \overline{\mathbb{P}}_{n,j} ||W_{i,j}^\top (\chi_{n,j}(\tau) - \chi(\tau))||_{op}^2\right]^{1/2} = o(1).$$

For Proposition A.2, we have

$$\begin{aligned}
&\sup \left\|\overline{\mathbb{P}}_{n,j} f_{\varepsilon_{i,j}(\tau)}(\delta_{i,j}(v,\tau)|W_{i,j}, Z_{i,j}) V_{i,j}^2(\tau) W_{i,j} \hat{\Phi}_{i,j}^\top(\tau)\right\|_{op} \\
&\leq \sup \left\|\overline{\mathbb{P}}_{n,j} f_{\varepsilon_{i,j}(\tau)}(\delta_{i,j}(v,\tau)|W_{i,j}, Z_{i,j}) V_{i,j}^2(\tau) W_{i,j} \underline{\hat{\Phi}}_{i,j}^\top(\tau) - \underline{\mathcal{J}}_{\gamma,\theta,j}(\tau)\right\|_{op} \\
&+ \sup \left\| \left[\overline{\mathbb{P}}_{n,j} f_{\varepsilon_{i,j}(\tau)}(\delta_{i,j}(v,\tau)|W_{i,j}, Z_{i,j}) V_{i,j}^2(\tau) W_{i,j} W_{i,j}^\top(\tau) - \mathcal{J}_{\gamma,\gamma,j}(\tau)\right] \hat{\chi}_j(\tau)\right\|_{op} \\
&+ \sup_{\tau\in\Upsilon} ||\hat{\mathcal{J}}_{\gamma,\gamma,j}(\tau)\hat{\chi}_j(\tau) - \underline{\hat{\mathcal{J}}}_{\gamma,\theta,j}(\tau)||_{op} + \sup_{\tau\in\Upsilon} ||\underline{\mathcal{J}}_{\gamma,\theta,j}(\tau) - \underline{\hat{\mathcal{J}}}_{\gamma,\theta,j}(\tau)||_{op} \\
&+ \sup_{\tau\in\Upsilon} \left\|\left[\mathcal{J}_{\gamma,\gamma,j}(\tau) - \hat{\mathcal{J}}_{\gamma,\gamma,j}(\tau)\right] \hat{\chi}_j(\tau)\right\|_{op}
\end{aligned}$$

where the suprema in the first three lines are taken over $\{j \in J, ||v||_2 \leq \delta, \tau \in \Upsilon\}$ for $v = (v_b^\top, v_r^\top, v_t^\top)^\top$. We note that, by Assumption 14, by letting $n \to \infty$ followed by $\delta \to 0$,

$$\begin{aligned}
&\sup \left\|\overline{\mathbb{P}}_{n,j} f_{\varepsilon_{i,j}(\tau)}(\delta_{i,j}(v,\tau)|W_{i,j}, Z_{i,j}) V_{i,j}^2(\tau) W_{i,j} \underline{\hat{\psi}}_{i,j}^\top(\tau) - \underline{\mathcal{J}}_{\gamma,\theta,j}(\tau)\right\|_{op} \xrightarrow{p} 0 \quad \text{and} \\
&\sup \left\| \left[\overline{\mathbb{P}}_{n,j} f_{\varepsilon_{i,j}(\tau)}(\delta_{i,j}(v,\tau)|W_{i,j}, Z_{i,j}) V_{i,j}^2(\tau) W_{i,j} W_{i,j}^\top(\tau) - \mathcal{J}_{\gamma,\gamma,j}(\tau)\right] \hat{\chi}_j(\tau)\right\|_{op} \xrightarrow{p} 0.
\end{aligned}$$

Assumption 14 also implies

$$\sup_{\tau\in\Upsilon} ||\underline{\mathcal{J}}_{\gamma,\theta,j}(\tau) - \underline{\hat{\mathcal{J}}}_{\gamma,\theta,j}(\tau)||_{op} + \sup_{\tau\in\Upsilon} \left\|\mathcal{J}_{\gamma,\gamma,j}(\tau) - \hat{\mathcal{J}}_{\gamma,\gamma,j}(\tau)\right\|_{op} = o_p(1).$$

Therefore, in order to show the result, it suffices to show

$$\sup_{\tau\in\Upsilon} ||\hat{\chi}_j(\tau)||_{op} = O_p(1) \tag{86}$$

and

$$\hat{\mathcal{J}}_{\gamma,\gamma,j}(\tau)\hat{\chi}_j(\tau) - \underline{\hat{\mathcal{J}}}_{\gamma,\theta,j}(\tau) = 0. \tag{87}$$

To see (86), we note that Assumption 14(ii) implies the generalized inverse is continuous at $\mathcal{J}_{\gamma,\gamma,j}(\tau)$ uniformly over $\tau \in \Upsilon$. Therefore, by the continuous mapping theorem, we have

$$||\hat{\mathcal{J}}_{\gamma,\gamma,j}^-(\tau) - \mathcal{J}_{\gamma,\gamma,j}^-(\tau)||_{op} = o_p(1),$$

and thus

$$\sup_{\tau\in\Upsilon} ||\hat{\chi}_j(\tau)||_{op} \leq \sup_{\tau\in\Upsilon} ||\hat{\chi}_j(\tau) - \mathcal{J}_{\gamma,\gamma,j}(\tau)\mathcal{J}_{\gamma,\gamma,j}^-(\tau)\mathcal{J}_{\gamma,\gamma,j}^-(\tau)\underline{\mathcal{J}}_{\gamma,\theta,j}(\tau)||_{op}$$

$$+ \sup_{\tau \in \Upsilon} ||\mathcal{J}_{\gamma,\gamma,j}(\tau)\mathcal{J}^{-}_{\gamma,\gamma,j}(\tau)\mathcal{J}^{-}_{\gamma,\gamma,j}(\tau)\underline{\mathcal{J}}_{\gamma,\theta,j}(\tau)||_{op} = O_p(1).$$

To show (87), we define $\mathcal{W}_j = (W_{1,j}K^{1/2}\left(\frac{\hat{\varepsilon}_{i,j}}{h}\right)h^{-1/2}, \cdots, W_{n_j,j}K^{1/2}\left(\frac{\hat{\varepsilon}_{n_j,j}}{h}\right)h^{-1/2})^\top$ and $\mathcal{I}_j = (\hat{\Phi}_{1,j}(\tau)K^{1/2}\left(\frac{\hat{\varepsilon}_{i,j}}{h}\right)h^{-1/2}, \cdots, \hat{\Phi}_{n_j,j}(\tau)K^{1/2}\left(\frac{\hat{\varepsilon}_{n_j,j}}{h}\right)h^{-1/2})^\top$. Then, we have

$$\hat{\mathcal{J}}_{\gamma,\gamma,j}(\tau) = \frac{1}{n_j}\mathcal{X}_j^\top \mathcal{X}_j \quad \text{and} \quad \hat{\mathcal{J}}_{\gamma,\theta,j}(\tau) = \frac{1}{n_j}\mathcal{X}_j^\top \mathcal{I}_j.$$

Define the singular value decomposition of $\mathcal{X}_j$ as $\mathcal{X}_j = U_j^\top \Sigma_j V_j$, where U_j and V_j are $n_j \times n_j$ and $d_w \times d_w$ orthonormal matrices and Σ_j is a $n_j \times d_w$ matrix with the first R diagonal elements being positive and all the rest entries in the matrix being zero. Then, we have

$$\begin{aligned}
\hat{\mathcal{J}}_{\gamma,\gamma,j}(\tau)\hat{\chi}_j(\tau) &= \frac{1}{n_j}\mathcal{X}_j^T\mathcal{X}_j\mathcal{X}_j^T\mathcal{X}_j\left(\mathcal{X}_j^T\mathcal{X}_j\right)^-\left(\mathcal{X}_j^T\mathcal{X}_j\right)^-\mathcal{X}_j^\top\mathcal{I}_j \\
&= \frac{1}{n_j}V_j^\top(\Sigma_j^\top\Sigma_j)(\Sigma_j^\top\Sigma_j)(\Sigma_j^\top\Sigma_j)^-(\Sigma_j^\top\Sigma_j)^-\Sigma_j^\top U_j\mathcal{I}_j \\
&= \frac{1}{n_j}V_j^\top\begin{pmatrix} I_R & 0_{R\times(d_w-R)} \\ 0_{(d_w-R)\times R} & 0_{(d_w-R)\times(d_w-R)}\end{pmatrix}\Sigma_j^\top U_j\mathcal{I}_j \\
&= \frac{1}{n_j}V_j^\top\Sigma_j^\top U_j\mathcal{I}_j \\
&= \hat{\mathcal{J}}_{\gamma,\theta,j}(\tau),
\end{aligned}$$

where we use the fact that

$$\begin{pmatrix} I_R & 0_{R\times(d_w-R)} \\ 0_{(d_w-R)\times R} & 0_{(d_w-R)\times(d_w-R)}\end{pmatrix}\Sigma_j^\top = \Sigma_j^\top.$$

This concludes the proof. ■

S.I Proof of Theorems B.1 and B.2

Define $\Upsilon = \{\tau_1, \tau_2\}$. Then,

$$\begin{aligned}
&\sqrt{n}\left(\lambda_1^\top\hat{\beta}(\tau_1) + \lambda_2^\top\hat{\beta}(\tau_2) - \lambda_0\right) \\
&= \lambda_1^\top\sqrt{n}\left(\hat{\beta}(\tau_1) - \beta_n(\tau_1)\right) + \lambda_2^\top\sqrt{n}\left(\hat{\beta}(\tau_2) - \beta_n(\tau_2)\right) + \mu \\
&= \lambda_1^\top\tilde{\Omega}(\tau_1)\left[\sum_{j\in J}\sqrt{\xi}_j\sqrt{n_j}(\mathbb{P}_{n,j} - \overline{\mathbb{P}}_{n,j})\tilde{f}_{\tau_1}(D_{i,j}, \beta_n(\tau_1), \gamma(\tau_1), 0)\right] \\
&+ \lambda_2^\top\tilde{\Omega}(\tau_2)\left[\sum_{j\in J}\sqrt{\xi}_j\sqrt{n_j}(\mathbb{P}_{n,j} - \overline{\mathbb{P}}_{n,j})\tilde{f}_{\tau_2}(D, \beta_n(\tau_2), \gamma(\tau_2), 0)\right] + \mu + o_p(1) \\
&= \sum_{j\in J}\sqrt{\xi_j}\left[\lambda_1^\top\tilde{\Omega}(\tau_1)\mathcal{Z}_j(\tau_1) + \lambda_2^\top\tilde{\Omega}(\tau_2)\mathcal{Z}_j(\tau_2)\right] + \mu + o_p(1),
\end{aligned}$$

where the $o_p(1)$ term holds uniformly over $\tau \in \Upsilon$.

By Lemma S.K.4, we have

$$\begin{aligned}
&\sqrt{n}\left(\lambda_1^\top(\hat{\beta}_g^*(\tau_1)-\hat{\beta}(\tau_1))+\lambda_2^\top(\hat{\beta}_g^*(\tau_2)-\hat{\beta}(\tau_2))\right) \\
&=\lambda_1^\top\tilde{\Omega}(\tau_1)\sum_{j\in J}\sqrt{\xi_j}g_j\sqrt{n_j}(\mathbb{P}_{n,j}-\overline{\mathbb{P}}_{n,j})\tilde{f}_{\tau_1}(D_{i,j},\beta_n(\tau_1),\gamma(\tau_1),0) \\
&+\lambda_2^\top\tilde{\Omega}(\tau_2)\sum_{j\in J}\sqrt{\xi_j}g_j\sqrt{n_j}(\mathbb{P}_{n,j}-\overline{\mathbb{P}}_{n,j})\tilde{f}_{\tau_2}(D,\beta_n(\tau_2),\gamma(\tau_2),0) \\
&-\lambda_1^\top\Omega(\tau_1)\sqrt{n}\sum_{j\in J}\xi_jg_j\mathcal{J}_{\pi,\beta,j}(\tau_1)(\hat{\beta}^r(\tau_1)-\beta_n(\tau_1)) \\
&-\lambda_2^\top\Omega(\tau_2)\sqrt{n}\sum_{j\in J}\xi_jg_j\mathcal{J}_{\pi,\beta,j}(\tau_2)(\hat{\beta}^r(\tau_2)-\beta_n(\tau_2))+o_p(1) \\
&=\lambda_1^\top\tilde{\Omega}(\tau_1)\sum_{j\in J}\sqrt{\xi_j}g_j\sqrt{n_j}(\mathbb{P}_{n,j}-\overline{\mathbb{P}}_{n,j})\tilde{f}_{\tau_1}(D_{i,j},\beta_n(\tau_1),\gamma(\tau_1),0)+\overline{a}_g^*(\tau_1)\mu \qquad (88)
\end{aligned}$$

where the last equality holds because

$$\begin{aligned}
&\lambda_1^\top\Omega(\tau_1)\sqrt{n}\sum_{j\in J}\xi_jg_j\mathcal{J}_{\pi,\beta,j}(\tau_1)(\hat{\beta}^r(\tau_1)-\beta_n(\tau_1))+\lambda_2^\top\Omega(\tau_2)\sqrt{n}\sum_{j\in J}\xi_jg_j\mathcal{J}_{\pi,\beta,j}(\tau_2)(\hat{\beta}^r(\tau_2)-\beta_n(\tau_2)) \\
&=\lambda_1^\top\overline{a}_g^*(\tau_1)\sqrt{n}(\hat{\beta}^r(\tau_1)-\beta_n(\tau_1))+\lambda_2^\top\overline{a}_g^*(\tau_2)\sqrt{n}(\hat{\beta}^r(\tau_2)-\beta_n(\tau_2))=-\overline{a}_g^*(\tau_1)\mu. \qquad (89)
\end{aligned}$$

Let $T_{2,\infty}^{QR}(g)=\left\|\sum_{j\in J}\sqrt{\xi_j}g_j\left[\lambda_1^\top\tilde{\Omega}(\tau_1)\mathcal{Z}_j(\tau_1)+\lambda_2^\top\tilde{\Omega}(\tau_2)\mathcal{Z}_j(\tau_2)+\sum_{j\in J}\xi_jg_ja_j(\tau_1)\mu\right]\right\|_{A_r}$. We have,

$$\left(T_{2,n}^{QR},\{T_{2,n}^{QR*}(g)\}_{g\in\mathbf{G}}\right)\rightsquigarrow\left(T_{2,\infty}^{QR}(\iota),\{T_{2,\infty}^{QR}(g)\}_{g\in\mathbf{G}}\right).$$

Then, following the same argument above, we have, under the null,

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

and

$$\lim_{||\mu||_2\to\infty}\liminf_{n\to\infty}\mathbb{P}\{T_{2,n}^{QR}>\hat{c}_{2,n}^{QR}(1-\alpha)\}=1.$$

The proof of Theorem B.2 is similar, and thus, is omitted for brevity. ■

S.J Proof of Theorem C.1

The proof for the bootstrap LM test follows similar arguments as those for the bootstrap AR tests. Let $\mathbb{S}\equiv\otimes_{j\in J}\mathbf{R}^{d_z\times d_x}\times\otimes_{j\in J}\mathbf{R}^{d_z}$, 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

$$T_{LM}(s) = \left\| \left(D(s)^\top \left(\sum_{j \in J} s_{2,j} s_{2,j}^\top \right)^{-1} D(s) \right)^{-1/2} D(s)^\top \left(\sum_{j \in J} s_{2,j} s_{2,j}^\top \right)^{-1} \sum_{j \in J} s_{2,j} \right\|^2, \tag{90}$$

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 $T_{LM}(s) = 0$ otherwise, where

$$\begin{aligned} D(s) &= (D_1(s), ..., D_{d_x}(s)), \\ D_l(s) &= \sum_{j \in J} s_{1,j,l} - \left(\sum_{j \in J} s_{1,j,l} s_{2,j}^\top \right) \left(\sum_{j \in J} s_{2,j} s_{2,j}^\top \right)^{-1} \sum_{j \in J} s_{2,j}, \end{aligned} \tag{91}$$

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

$$S_n = \left(\left\{ \frac{1}{n} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} X_{i,j}^\top : j \in J \right\}, \left\{ \frac{1}{\sqrt{n}} \sum_{i \in I_{n,j}} \tilde{Z}_{i,j} \varepsilon_{i,j} : j \in J \right\} \right), \tag{92}$$

and define the statistic $\widehat{S}_n$ by replacing $\varepsilon_{i,j}$ in S_n with $\hat{\varepsilon}_{i,j}^r = y_{i,j} - X_{i,j}^\top \beta_0 - W_{i,j}^\top \bar{\gamma}^r$. Then, by Assumptions 1(ii), 2(i)-(ii), and the continuous mapping theorem we have

$$S_n \xrightarrow{d} \left(\left\{ \xi_j Q_{\tilde{Z}X,j} : j \in J \right\}, \left\{ \sqrt{\xi_j} \mathcal{Z}_{\varepsilon,j} : j \in J \right\} \right) \equiv S, \tag{93}$$

where $\xi_j > 0$ for all $j \in J$. Furthermore, by Assumptions 1(i), Assumptions 2(iii)-(iv), and $\beta_n = \beta_0$, we have for every $g \in \mathbf{G}$,

$$T_{LM}(g\widehat{S}_n) = T_{LM}(gS_n) + o_P(1). \tag{94}$$

We then obtain from (93)-(94) and the continous mapping theorem that

$$\left(T_{LM}(S_n), \{T_{LM}(g\widehat{S}_n) : g \in \mathbf{G}\} \right) \xrightarrow{d} \left(T_{LM}(S), \{T_{LM}(gS) : g \in \mathbf{G}\} \right). \tag{95}$$

Then, the upper and lower bounds follow by using similar arguments as those for Theorem 2.1.

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

$$\begin{aligned} LR_n &= \frac{1}{2} \left\{ AR_{CR,n} - rk_n + \sqrt{(AR_{CR,n} - rk_n)^2 + 4 \cdot LM_n \cdot rk_n} \right\} \\ &= \frac{1}{2} \left\{ AR_{CR,n} - rk_n + |AR_{CR,n} - rk_n| \sqrt{1 + \frac{4 \cdot LM_n \cdot rk_n}{(AR_{CR,n} - rk_n)^2}} \right\} \\ &= \frac{1}{2} \left\{ AR_{CR,n} - rk_n + |AR_{CR,n} - rk_n| \left(1 + 2 \cdot LM_n \frac{rk_n}{(AR_{CR,n} - rk_n)^2} (1 + o_P(1)) \right) \right\} \\ &= LM_n \frac{rk_n}{rk_n - AR_{CR,n}} (1 + o_P(1)) = LM_n + o_P(1), \end{aligned} \tag{96}$$

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 1(ii). Using arguments similar to those in (96), we obtain that for each $g \in \mathbf{G}$,

$$LR_n^*(g) = LM_n^*(g)\frac{rk_n}{rk_n - AR^*_{CR,n}(g)}(1 + o_P(1)) = LM_n^*(g) + o_P(1), \tag{97}$$

by $AR^*_{CR,n}(g) - rk_n < 0$ w.p.a.1 since $AR^*_{CR,n}(g) = O_P(1)$ for each $g \in \mathbf{G}$. The desired result for the CQLR test follows. ■

S.K Technical Lemmas used in the Proofs of Results in Section 3

S.K.1 Linear Expansion of $\hat{\gamma}(b_n(\tau), \tau)$

Lemma S.K.1. *Let $\mathcal{B}(\delta) = \{b(\cdot) \in \ell^\infty(\Upsilon) : \sup_{\tau\in\Upsilon} ||b(\tau) - \beta_n(\tau)||_2 \leq \delta\}$. Suppose Assumptions 5 and 6 hold. Let $b_n(\tau)$ be a generic point in $\mathcal{B}(\delta)$. Then, for any $\varepsilon > 0$, there exists $\overline{\delta}$ and constants $c', c > 0$ that are independent of $(n, \delta, \varepsilon, \overline{\delta})$ such that for $\delta \leq \overline{\delta}$, with probability greater than $1 - c\varepsilon$,*

$$\begin{aligned}
&\begin{pmatrix} \sqrt{n}\left(\hat{\gamma}(b_n(\tau), \tau) - \gamma_n(\tau)\right) \\ \sqrt{n}\hat{\theta}(b_n(\tau), \tau) \end{pmatrix} \\
&= \left[\Gamma_1(b_n(\tau), \tau)\right]^{-1} \sqrt{n}\left[I_n(\tau) + II_n(b_n(\tau), \tau) + o_p(1/\sqrt{n}) - \Gamma_2(b_n(\tau), \tau)(b_n(\tau) - \beta_n(\tau))\right]
\end{aligned} \tag{98}$$

where the $o_p(1/\sqrt{n})$ term on the RHS of the above display holds uniformly over $\tau \in \Upsilon, b_n(\cdot) \in B(\delta)$,

$$I_n(\tau) = (\mathbb{P}_n - \overline{\mathbb{P}}_n) f_\tau(D, \beta_n(\tau), \gamma_n(\tau), 0), \quad \sup_{\tau\in\Upsilon} ||\sqrt{n} I_n(\tau)||_2 = O_p(1),$$

$$II_n(b_n(\tau), \tau) = (\mathbb{P}_n - \overline{\mathbb{P}}_n)\left(\hat{f}_\tau(D, b_n(\tau), \hat{\gamma}(b_n(\tau), \tau), \hat{\theta}_n(b_n(\tau), \tau)) - f_\tau(D, \beta_n(\tau), \gamma_n(\tau), 0)\right),$$

$$\sup_{b_n(\cdot)\in B(\delta), \tau\in\Upsilon} ||\sqrt{n} II_n(b_n(\tau), \tau)||_2 \leq c'\varepsilon,$$

$$\Gamma_1(b_n(\tau), \tau) = \overline{\mathbb{P}}_n f_{\varepsilon_{i,j}(\tau)}(\hat{\delta}_{i,j}(\tau)|W_{i,j}, Z_{i,j})\hat{\Psi}_{i,j}(\tau)\hat{\Psi}^\top_{i,j}(\tau), \qquad \sup_{b_n(\cdot)\in B(\delta), \tau\in\Upsilon} ||\Gamma_1(b_n(\tau), \tau) - \mathcal{J}_{\pi,\beta}(\tau)||_{op} \leq \varepsilon,$$

$$\Gamma_2(b_n(\tau), \tau) = \overline{\mathbb{P}}_n f_{\varepsilon_{i,j}(\tau)}(\hat{\delta}_{i,j}(\tau)|W_{i,j}, Z_{i,j})\hat{\Psi}_{i,j}(\tau)X^\top_{i,j}, \qquad \sup_{b_n(\cdot)\in B(\delta), \tau\in\Upsilon} ||\Gamma_2(b_n(\tau), \tau) - \mathcal{J}_{\pi,\beta}(\tau)||_{op} \leq \varepsilon,$$

and $\hat{\delta}_{i,j}(\tau) \in (0, X^\top_{i,j}(b_n(\tau) - \beta_n(\tau)) + W^\top_{i,j}(\hat{\gamma}(b_n(\tau), \tau) - \gamma_n(\tau)) + \hat{\Phi}^\top_{i,j}(\tau)\hat{\theta}(b_n(\tau), \tau))$.

Proof. By Assumption 5(vi), the subgradient condition for $(\hat{\gamma}(b_n(\tau),\tau), \hat{\theta}(b_n(\tau),\tau))$ implies

$$\begin{aligned}
o_p(1/\sqrt{n}) &= \mathbb{P}_n \hat{f}_\tau(D_{i,j}, b_n(\tau), \hat{\gamma}(b_n(\tau),\tau), \hat{\theta}(b_n(\tau),\tau)) \\
&= (\mathbb{P}_n - \overline{\mathbb{P}}_n)\hat{f}_\tau(D_{i,j}, b_n(\tau), \hat{\gamma}(b_n(\tau),\tau), \hat{\theta}(b_n(\tau),\tau)) + \overline{\mathbb{P}}_n \hat{f}_\tau(D_{i,j}, b_n(\tau), \hat{\gamma}(b_n(\tau),\tau), \hat{\theta}(b_n(\tau),\tau)) \\
&= (\mathbb{P}_n - \overline{\mathbb{P}}_n) f_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0) \\
&+ (\mathbb{P}_n - \overline{\mathbb{P}}_n)\Big(\hat{f}_\tau(D_{i,j}, b_n(\tau), \hat{\gamma}(b_n(\tau),\tau), \hat{\theta}(b_n(\tau),\tau)) - f_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0)\Big) \\
&+ \overline{\mathbb{P}}_n \hat{f}_\tau(D_{i,j}, b_n(\tau), \hat{\gamma}(b_n(\tau),\tau), \hat{\theta}(b_n(\tau),\tau)) \\
&= I_n(\tau) + II_n(b_n(\tau),\tau) + III_n(b_n(\tau),\tau), \qquad (99)
\end{aligned}$$

where the $o_p(1/\sqrt{n})$ term on the LHS of the above display holds uniformly over $\tau \in \Upsilon, b_n(\cdot) \in B(\delta)$. For the first term, we note that $\overline{\mathbb{P}}_n f_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0) = 0$. Then, by Assumption 6(vi), we have,

$$\sup_{\tau \in \Upsilon} ||\sqrt{n} I_n(\tau)||_2 = O_p(1).$$

For the second term, note

$$\begin{aligned}
&\sup_{\tau \in \Upsilon} ||\hat{\gamma}(b_n(\tau),\tau) - \gamma_n(\tau)||_2 \\
&\leq \sup_{\tau \in \Upsilon, b \in \mathcal{B}} ||\hat{\gamma}(b,\tau) - \gamma_n(b,\tau)||_2 + \sup_{\tau \in \Upsilon, b, b' \in \mathcal{B}, ||b-b'||_2 \leq \delta} ||\gamma_n(b,\tau) - \gamma_n(b',\tau)||_2 \\
&\leq \sup_{\tau \in \Upsilon, b \in \mathcal{B}} ||\hat{\gamma}(b,\tau) - \gamma_n(b,\tau)||_2 + \sup_{\tau \in \Upsilon, b, b' \in \mathcal{B}, ||b-b'||_2 \leq \delta} ||\gamma_\infty(b,\tau) - \gamma_\infty(b',\tau)||_2 \\
&+ 2 \sup_{\tau \in \Upsilon, b \in \mathcal{B}} ||\gamma_\infty(b,\tau) - \gamma_n(b,\tau)||_2,
\end{aligned}$$

where the RHS of the above display vanishes as $n \to \infty$ followed by $\delta \downarrow 0$ and it holds by the same argument in Step 1 of the proof of Lemma S.K.2. Therefore, for any $\delta' > 0$ and $\varepsilon > 0$, there exist $\underline{n}$ and $\overline{\delta}$ such that for $n \geq \underline{n}$ and $\delta \leq \overline{\delta}$, with probability greater than $1 - \varepsilon$,

$$\sup_{\tau \in \Upsilon} ||\hat{\gamma}(b_n(\tau),\tau) - \gamma_n(\tau)||_2 \leq \delta'.$$

Similarly, we have

$$\sup_{\tau \in \Upsilon} ||\hat{\theta}(b_n(\tau),\tau) - 0||_2 \leq \delta'.$$

Then, for any $\varepsilon > 0$, there exist $\underline{n}$ and $\overline{\delta}$ such that for $n \geq \underline{n}$ and $\delta', \delta \leq \overline{\delta}$, with probability greater than $1 - \varepsilon$, we have,

$$\begin{aligned}
&\sup_{b_n(\cdot) \in B(\delta), \tau \in \Upsilon} ||\sqrt{n} II_n(b_n(\tau),\tau)||_2 \\
&\leq \sup_{||v||_2 \leq 2\delta' + \delta} \left\| \sqrt{n}(\mathbb{P}_n - \overline{\mathbb{P}}_n)\Big(\hat{f}_\tau(D_{i,j}, \beta_n(\tau) + v_b, \gamma_n(\tau) + v_r, v_t) - f_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0)\Big) \right\|_2.
\end{aligned}$$

Then, by Assumption 6(ii), for any $\varepsilon > 0$, there exist $\underline{n}$ and $\overline{\delta}$ such that for $n \geq \underline{n}$ and $\delta, \delta' \leq \overline{\delta}$, we have, with probability greater than $1 - c\varepsilon$,

$$\sup_{b_n(\cdot)\in B(\delta),\tau\in\Upsilon} ||\sqrt{n}II_n(b_n(\tau),\tau)||_2 \leq c'\varepsilon,$$

for some constants (c, c') that are independent of $(n, \delta, \delta', \varepsilon)$.

For the third term in (99), we have

$$\begin{aligned}
&\overline{\mathbb{P}}_n \hat{f}_\tau(D_{i,j}, b_n(\tau), \hat{\gamma}(b_n(\tau),\tau), \hat{\theta}(b_n(\tau),\tau)) \\
&= \sum_{j\in J} \xi_j \overline{\mathbb{P}}_{n,j}(\tau - 1\{y_{i,j} - X_{i,j}^\top b_n(\tau) - W_{i,j}^\top \hat{\gamma}(b_n(\tau),\tau) - \hat{\Phi}_{i,j}^\top(\tau)\hat{\theta}(b_n(\tau),\tau) \leq 0\})\hat{\Psi}_{i,j}(\tau) \\
&= -\overline{\mathbb{P}}_{n,j} f_{\varepsilon_{i,j}(\tau)}(\hat{\delta}_{i,j}(\tau)|W_{i,j}, Z_{i,j})\hat{\Psi}_{i,j}(\tau)X_{i,j}^\top(b_n(\tau) - \beta_n(\tau)) \\
&- \overline{\mathbb{P}}_{n,j} f_{\varepsilon_{i,j}(\tau)}(\hat{\delta}_{i,j}(\tau)|W_{i,j}, Z_{i,j})\hat{\Psi}_{i,j}(\tau)\hat{\Psi}_{i,j}^\top(\tau) \begin{pmatrix} \hat{\gamma}(b_n(\tau),\tau) - \gamma_n(\tau) \\ \hat{\theta}(b_n(\tau),\tau) \end{pmatrix},
\end{aligned} \tag{100}$$

where $\hat{\delta}_{i,j}(\tau) \in (0, X_{i,j}^\top(b_n(\tau) - \beta_n(\tau)) + W_{i,j}^\top(\hat{\gamma}(b_n(\tau),\tau) - \gamma_n(\tau)) + \hat{\Phi}_{i,j}^\top(\tau)\hat{\theta}(b_n(\tau),\tau))$. For any $\varepsilon > 0$, there exist $\underline{n}$ and $\overline{\delta}$ such that for $n \geq \underline{n}$ and $\delta, \delta' \leq \overline{\delta}$, we have, with probability greater than $1 - c\varepsilon$,

$$\sup_{\tau\in\Upsilon}\left(||b_n(\tau) - \beta_n(\tau)||_2 + ||\hat{\gamma}(b_n(\tau),\tau) - \gamma_n(\tau)||_2 + ||\hat{\theta}(b_n(\tau),\tau)||_2\right) \leq \delta + 2\delta'.$$

This implies, with probability greater than $1 - c\varepsilon$,

$$\begin{aligned}
&\sup_{b_n(\cdot)\in B(\delta),\tau\in\Upsilon,j\in J} ||\Gamma_2(b_n(\tau),\tau) - \mathcal{J}_{\pi,\beta}(\tau)||_{op} \\
&\leq \left(\max_{j\in J}\xi_j\right) \sup_{b_n(\cdot)\in B(\delta),\tau\in\Upsilon,j\in J} \left\|\overline{\mathbb{P}}_{n,j} f_{\varepsilon_{i,j}(\tau)}(\hat{\delta}_{i,j}(\tau)|W_{i,j}, Z_{i,j})\hat{\Psi}_{i,j}(\tau)X_{i,j}^\top - \mathcal{J}_{\pi,\beta}(\tau)\right\|_{op} \\
&\leq \left(\max_{j\in J}\xi_j\right) \sup \left\|\overline{\mathbb{P}}_{n,j} f_{\varepsilon_{i,j}(\tau)}(\delta_{i,j}(v,\tau)|W_{i,j}, Z_{i,j})\hat{\Psi}_{i,j}(\tau)X_{i,j}^\top - \mathcal{J}_{\pi,\beta,j}(\tau)\right\|_{op},
\end{aligned}$$

where the supremum in the second inequality is taken over $\{(j, v, \tau) : j \in J, ||v||_2 \leq \delta + 2\delta', \tau \in \Upsilon\}$. Then, Assumption 6(iii) implies, with probability greater than $1 - c\varepsilon$,

$$\sup_{b_n(\cdot)\in B(\delta),\tau\in\Upsilon} ||\Gamma_2(b_n(\tau),\tau) - \mathcal{J}_{\pi,\beta}(\tau)||_{op} \leq \varepsilon.$$

Similarly, we have, with probability greater than $1 - c\varepsilon$,

$$\sup_{b_n(\cdot)\in B(\delta),\tau\in\Upsilon} ||\Gamma_1(b_n(\tau),\tau) - \mathcal{J}_{\pi,\pi}(\tau)||_{op} \leq \varepsilon.$$

Then, by Assumption 6 and the fact that ε can be made arbitrarily small, we have, with probability greater than $1 - c\varepsilon$, $\Gamma_1(b_n(\tau),\tau)$ is invertible. Therefore, (99) implies, with probability greater than $1 - c\varepsilon$,

$$\begin{pmatrix} \sqrt{n}\left(\hat{\gamma}(b_n(\tau),\tau) - \gamma_n(\tau)\right) \\ \sqrt{n}\hat{\theta}(b_n(\tau),\tau) \end{pmatrix}$$
$$= \left[\Gamma_1(b_n(\tau),\tau)\right]^{-1}\sqrt{n}\left[I_n(\tau) + II_n(b_n(\tau),\tau) + o_p(1/\sqrt{n}) - \Gamma_2(b_n(\tau),\tau)(b_n(\tau) - \beta_n(\tau))\right]$$

where the $o_p(1/\sqrt{n})$ term on the RHS of the above display holds uniformly over $\tau \in \Upsilon, b_n(\cdot) \in B(\delta)$. □

S.K.2 Technical Results for the IVQR Estimator

Lemma S.K.2. *Suppose Assumptions 5, 6, and 8(i)–8(iii) hold. Then,*

$$\begin{pmatrix} \sqrt{n}\left(\hat{\beta}(\tau) - \beta_n(\tau)\right) \\ \sqrt{n}\left(\hat{\gamma}(\tau) - \gamma_n(\tau)\right) \\ \sqrt{n}\hat{\theta}(\tau) \end{pmatrix}$$
$$= \begin{pmatrix} \left[\mathcal{J}^\top_{\pi,\beta}(\tau)\overline{\mathcal{J}}^\top_\theta(\tau)A_\phi(\tau)\overline{\mathcal{J}}_\theta(\tau)\mathcal{J}_{\pi,\beta}(\tau)\right]^{-1}\mathcal{J}^\top_{\pi,\beta}(\tau)\overline{\mathcal{J}}^\top_\theta(\tau)A_\phi(\tau)\overline{\mathcal{J}}_\theta(\tau) \\ \overline{\mathcal{J}}_{\pi,\pi}(\tau)\left[\mathbb{I}_{d_w+d_\phi} - \mathcal{J}_{\pi,\beta}(\tau)\left[\mathcal{J}^\top_{\pi,\beta}(\tau)\overline{\mathcal{J}}^\top_\theta(\tau)A_\phi(\tau)\overline{\mathcal{J}}_\theta(\tau)\mathcal{J}_{\pi,\beta}(\tau)\right]^{-1}\mathcal{J}^\top_{\pi,\beta}(\tau)\overline{\mathcal{J}}^\top_\theta(\tau)A_\phi(\tau)\overline{\mathcal{J}}_\theta(\tau)\right] \end{pmatrix}$$
$$\times \sqrt{n}\mathbb{P}_n f_\tau(D_{i,j},\beta_n(\tau),\gamma_n(\tau),0) + o_p(1),$$

where $o_p(1)$ term holds uniformly over $\tau \in \Upsilon$.

Proof. We divide the proof into three steps. In the first step, we show $(\hat{\beta}(\tau),\hat{\gamma}(\tau),\hat{\theta}(\tau))$ are consistent. In the second step, we derive convergence rates of $(\hat{\beta}(\tau),\hat{\gamma}(\tau),\hat{\theta}(\tau))$. In the third step, we derive linear expansions for $(\hat{\beta}(\tau),\hat{\gamma}(\tau),\hat{\theta}(\tau))$.

Step 1. We first show the consistency of $(\hat{\beta}(\tau),\hat{\gamma}(\tau),\hat{\theta}(\tau))$. Note by construction, we have $\gamma(\beta_n(\tau),\tau) = \gamma_n(\tau)$, $\theta_n(\beta_n(\tau),\tau) = 0$, $\hat{\gamma}(\tau) = \hat{\gamma}(\hat{\beta}(\tau),\tau)$, and $\hat{\theta}(\tau) = \hat{\theta}(\hat{\beta}(\tau),\tau)$. By Kato (2009, Theorem 1) and the fact that both $\hat{Q}_n(b,r,t,\tau)$ and $Q_n(b,r,t,\tau)$ are convex, we have

$$\sup_{(b,\tau)\in\mathcal{B}\times\Upsilon}\left(||\hat{\gamma}(b,\tau) - \gamma_n(b,\tau)||_2 + ||\hat{\theta}(b,\tau) - \theta_n(b,\tau)||_2\right) = o_p(1).$$

Similarly, we have

$$\sup_{(b,\tau)\in\mathcal{B}\times\Upsilon}\left(||\gamma_\infty(b,\tau) - \gamma_n(b,\tau)||_2 + ||\theta_\infty(b,\tau) - \theta_n(b,\tau)||_2\right) = o(1).$$

This implies

$$\sup_{(b,\tau)\in\mathcal{B}\times\Upsilon}\left|||\hat{\theta}(b,\tau)||_{\hat{A}_\phi(\tau)} - ||\theta_\infty(b,\tau)||_{A_\phi(\tau)}\right| = o_p(1)$$

and $0 = \lim_{n\to\infty}\theta_n(\beta_n(\tau),\tau) = \theta_\infty(\beta_0(\tau),\tau)$, which implies $||\theta_\infty(b,\tau)||_{A_\phi(\tau)}$ is uniquely minimized at $b = \beta_0(\tau)$. Then, Chernozhukov and Hansen (2006, Lemma B.1) implies

$$\sup_{\tau \in \Upsilon} ||\hat{\beta}(\tau) - \beta_0(\tau)||_2 = o_p(1),$$

and thus,

$$\sup_{\tau \in \Upsilon} ||\hat{\beta}(\tau) - \beta_n(\tau)||_2 = o_p(1).$$

Then, we have

$$\sup_{\tau \in \Upsilon} ||\hat{\gamma}(\hat{\beta}(\tau), \tau) - \gamma_n(\tau)||_2 \leq \sup_{\tau \in \Upsilon, b \in \mathcal{B}} ||\hat{\gamma}(b, \tau) - \gamma_\infty(b, \tau)||_2 + \sup_{\tau \in \Upsilon} ||\gamma_\infty(\hat{\beta}(\tau), \tau) - \gamma_\infty(\beta_n(\tau), \tau)||_2 = o_p(1),$$

and similarly,

$$\sup_{\tau \in \Upsilon} ||\hat{\theta}(\hat{\beta}(\tau), \tau) - 0||_2 = o_p(1).$$

Step 2. We derive the convergence rates of $\hat{\beta}(\tau)$, $\hat{\gamma}(\tau)$, and $\hat{\theta}(\tau)$. Let $\mathcal{B}(\delta) = \{b(\cdot) \in \ell^\infty(\Upsilon) : \sup_{\tau \in \Upsilon} ||b(\tau) - \beta_n(\tau)||_2 \leq \delta\}$. For any $\delta > 0$, we have $\hat{\beta}(\cdot) \in \mathcal{B}(\delta)$ w.p.a.1. Let $b_n(\tau)$ be a generic point in $\mathcal{B}(\delta)$. Recall $\omega = (0_{d_w \times d_\phi}, \mathbb{I}_{d_\phi})$. Then, Lemma S.K.1 implies, with probability greater than $1 - c\varepsilon$,

$$\begin{aligned}
&||\sqrt{n}\hat{\theta}(b_n(\tau), \tau)||^2_{\hat{A}_\phi(\tau)} \\
&= n\left[I_n(\tau) + II_n(b_n(\tau), \tau) + O_p(1/n) - \Gamma_2(b_n(\tau), \tau)(b_n(\tau) - \beta_n(\tau))\right]^\top \\
&\times \left[\Gamma_1(b_n(\tau), \tau)\right]^{-1} \omega^\top \hat{A}_\phi(\tau)\omega \left[\Gamma_1(b_n(\tau), \tau)\right]^{-1} \\
&\times \left[I_n(\tau) + II_n(b_n(\tau), \tau) + O_p(1/n) - \Gamma_2(b_n(\tau), \tau)(b_n(\tau) - \beta_n(\tau))\right]^\top \\
&\geq n(b_n(\tau) - \beta_n(\tau))^\top \left\{\Gamma_2^\top(b_n(\tau), \tau)\left[\Gamma_1(b_n(\tau), \tau)\right]^{-1} \omega^\top \hat{A}_\phi(\tau)\omega \left[\Gamma_1(b_n(\tau), \tau)\right]^{-1} \Gamma_2(b_n(\tau), \tau)\right\}(b_n(\tau) - \beta_n(\tau)) \\
&- O_p(1),
\end{aligned}$$

where the $O_p(1)$ term on the RHS of the above display holds uniformly over $\tau \in \Upsilon, b_n(\cdot) \in B(\delta)$. In addition, Lemma S.K.1 implies that there exists a constant $C > 0$ such that for any $\varepsilon > 0$, with probability greater than $1 - c\varepsilon$

$$\begin{aligned}
\sup_{b_n(\cdot) \in B(\delta), \tau \in \Upsilon} \Bigg\| &\Gamma_2^\top(b_n(\tau), \tau)\left[\Gamma_1(b_n(\tau), \tau)\right]^{-1} \omega^\top \hat{A}_\phi(\tau)\omega \left[\Gamma_1(b_n(\tau), \tau)\right]^{-1} \Gamma_2(b_n(\tau), \tau) \\
&- \mathcal{J}_{\pi,\beta}^\top(\tau)\mathcal{J}_{\theta,\theta}^{-1}(\tau)^T A_\phi(\tau)\mathcal{J}_{\theta,\theta}^{-1}(\tau)\mathcal{J}_{\pi,\beta}(\tau)\Bigg\|_{op} \leq C\varepsilon.
\end{aligned}$$

Therefore, by Assumption 6, there exists a constant $\underline{c}$ independent of τ and $b_n(\cdot)$ such that

$$\begin{aligned}
&n(b_n(\tau) - \beta_n(\tau))^\top \left\{\Gamma_2^\top(b_n(\tau), \tau)\left[\Gamma_1(b_n(\tau), \tau)\right]^{-1} \omega^\top \hat{A}_\phi(\tau)\omega \left[\Gamma_1(b_n(\tau), \tau)\right]^{-1} \Gamma_2(b_n(\tau), \tau)\right\}(b_n(\tau) - \beta_n(\tau)) \\
&- O_p(1)
\end{aligned}$$

$$\geq \underline{c}||\sqrt{n}(b_n(\tau) - \beta_n(\tau))||_2^2 - O_p(1),$$

where the $O_p(1)$ term on the RHS of the above display holds uniformly over $\tau \in \Upsilon, b_n(\cdot) \in B(\delta)$.

On the other hand, we have $\hat{\beta}(\tau) \in B(\delta)$ w.p.a.1 for any $\delta > 0$ and

$$||\sqrt{n}\hat{\theta}(\hat{\beta}(\tau), \tau)||_{\hat{A}_\phi(\tau)} \leq ||\sqrt{n}\hat{\theta}(\beta_n(\tau), \tau)||_{\hat{A}_\phi(\tau)} = O_p(1),$$

where the $O_p(1)$ term on the RHS of the above display holds uniformly over $\tau \in \Upsilon, b(\cdot) \in B(\delta)$ and the last equality holds by Lemma S.K.1. This implies, for any $\varepsilon > 0$, there exist $\underline{n}$ and $\overline{\delta}$ such that for $n \geq \underline{n}$ and $\delta, \delta' \leq \overline{\delta}$, we have, with probability greater than $1 - c\varepsilon$,

$$\underline{c}||\sqrt{n}(\hat{\beta}(\tau) - \beta_n(\tau))||_2^2 - O_p(1) \leq ||\sqrt{n}\hat{\theta}(\hat{\beta}(\tau), \tau)||_{\hat{A}_\phi(\tau)} \leq O_p(1),$$

where the $O_p(1)$ term on the RHS of the above display holds uniformly over $\tau \in \Upsilon$. This further implies

$$\sup_{\tau \in \Upsilon} ||\sqrt{n}(\hat{\beta}(\tau) - \beta_n(\tau))||_2 = O_p(1). \tag{101}$$

Plugging (101) into (98), we obtain that

$$\sup_{\tau \in \Upsilon} ||\sqrt{n}(\hat{\gamma}(\tau) - \gamma_n(\tau))||_2 = O_p(1) \quad \text{and} \quad \sup_{\tau \in \Upsilon} ||\sqrt{n}(\hat{\theta}(\tau) - 0)||_2 = O_p(1).$$

Step 3. Next, we derive the linear expansions for $\hat{\beta}(\tau)$ and $\hat{\gamma}(\tau)$. Let $\hat{u}(\tau) = \sqrt{n}(\hat{\beta}(\tau) - \beta_n(\tau))$. Then, Step 2 shows $\sup_{\tau \in \Upsilon} ||\hat{u}(\tau)||_2 = O_p(1)$. For any $\varepsilon > 0$, there exists a constant $C > 0$ such that with probability greater than $1 - \varepsilon$, we have, for all $\tau \in \Upsilon$,

$$\hat{u}(\tau) = \arg \inf_{u:||u||_2 \leq C} ||\hat{\theta}(\beta_n(\tau) + u/\sqrt{n}, \tau)||_{\hat{A}_\tau}.$$

Denote $b_n(\tau) = \beta_n(\tau) + u/\sqrt{n}$ for $||u||_2 \leq C$. Then, by Lemma S.K.1, we have

$$||\hat{\theta}(\beta_n(\tau) + u/\sqrt{n}, \tau)||_{\hat{A}_\tau}$$
$$= \left[\sqrt{n}\mathbb{P}_n f_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0) - \Gamma_2(\beta_n(\tau) + u/\sqrt{n}, \tau)u + II_n(\beta_n(\tau) + u/\sqrt{n}, \tau) + O_p(1/\sqrt{n})\right]^\top$$
$$\times \left[\Gamma_1^{-1}(\beta_n(\tau) + u/\sqrt{n}, \tau)\omega^\top A_\phi(\tau)\omega\Gamma_1^{-1}(\beta_n(\tau) + u/\sqrt{n}, \tau)\right]$$
$$\times \left[\sqrt{n}\mathbb{P}_n f_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0) - \Gamma_2(\beta_n(\tau) + u/\sqrt{n}, \tau)u + II_n(\beta_n(\tau) + u/\sqrt{n}, \tau) + O_p(1/\sqrt{n})\right],$$

where the $O_p(1/\sqrt{n})$ term is uniform over $\tau \in \Upsilon$ and $|u| \leq C$. In addition, by Assumption 6, we have

$$\sup_{\tau \in \Upsilon, ||u||_2 \leq C} \left\|\Gamma_1^{-1}(\beta_n(\tau) + u/\sqrt{n}, \tau) - \mathcal{J}_{\pi,\pi}^{-1}(\tau)\right\|_{op} = o_p(1), \tag{102}$$

$$\sup_{\tau \in \Upsilon, ||u||_2 \leq C} \left\|\Gamma_2(\beta_n(\tau) + u/\sqrt{n}, \tau) - \mathcal{J}_{\pi,\beta}(\tau)\right\|_{op} = o_p(1),$$

and

$$\sup_{\tau \in \Upsilon, ||u||_2 \leq C} ||II_n(\beta_n(\tau) + u/\sqrt{n}, \tau)||_2 = o_p(1). \tag{103}$$

Further recall $\overline{J}_\theta(\tau) = \omega J^{-1}_{\pi,\pi}(\tau)$. Then, we have

$$\left| ||\hat{\theta}(\beta_n(\tau) + u/\sqrt{n}, \tau)||_{\hat{A}_\tau} - ||\overline{J}_\theta(\tau) \left[\sqrt{n}\mathbb{P}_n f_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0) - \mathcal{J}_{\pi,\beta}(\tau) u\right] ||_{A_\phi(\tau)} \right| = o_p(1).$$

Then, Chernozhukov and Hansen (2006, Lemma B.1) implies

$$\begin{aligned} \hat{u}(\tau) = &\left[\mathcal{J}^\top_{\pi,\beta}(\tau) \overline{\mathcal{J}}^\top_\theta(\tau) A_\phi(\tau) \overline{\mathcal{J}}_\theta(\tau) \mathcal{J}_{\pi,\beta}(\tau)\right]^{-1} \\ &\times \mathcal{J}^\top_{\pi,\beta}(\tau) \overline{\mathcal{J}}^\top_\theta(\tau) A_\phi(\tau) \overline{\mathcal{J}}_\theta(\tau) \sqrt{n}\mathbb{P}_n f_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0) + o_p(1), \end{aligned} \tag{104}$$

where $o_p(1)$ term holds uniformly over $\tau \in \Upsilon$. Plugging (104) into (98), we have

$$\begin{aligned} &\begin{pmatrix} \sqrt{n}\left(\hat{\gamma}(\tau) - \gamma_n(\tau)\right) \\ \sqrt{n}\hat{\theta}(\tau) \end{pmatrix} \\ &= \left[\Gamma_1(\hat{\beta}(\tau), \tau)\right]^{-1} \sqrt{n}\left[I_n(\tau) + II_n(\hat{\beta}(\tau), \tau) + o_p(1/\sqrt{n}) - \Gamma_2(\hat{\beta}(\tau), \tau)(\hat{\beta}(\tau) - \beta_n(\tau))\right] \\ &= \mathcal{J}^{-1}_{\pi,\pi}(\tau) \left[\mathbb{I}_{d_w + d_\phi} - \mathcal{J}_{\pi,\beta}(\tau) \left[\mathcal{J}^\top_{\pi,\beta}(\tau) \overline{\mathcal{J}}^\top_\theta(\tau) A_\phi(\tau) \overline{\mathcal{J}}_\theta(\tau) \mathcal{J}_{\pi,\beta}(\tau)\right]^{-1} \mathcal{J}^\top_{\pi,\beta}(\tau) \overline{\mathcal{J}}^\top_\theta(\tau) A_\phi(\tau) \overline{\mathcal{J}}_\theta(\tau)\right] \\ &\times \sqrt{n}\mathbb{P}_n f_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0) + o_p(1), \end{aligned}$$

where both $o_p(1/\sqrt{n})$ and $o_p(1)$ terms hold uniformly over $\tau \in \Upsilon$. This concludes the proof. □

Lemma S.K.3. *If Assumptions 5 and 6 hold, then*

$$\begin{pmatrix} \sqrt{n}\left(\hat{\gamma}(\beta_n(\tau), \tau) - \gamma_n(\tau)\right) \\ \sqrt{n}\hat{\theta}(\beta_n(\tau), \tau) \end{pmatrix} = \mathcal{J}^{-1}_{\pi,\pi}(\tau) \sqrt{n}(\mathbb{P}_n - \overline{\mathbb{P}}_n) f_\tau(D, \beta_n(\tau), \gamma_n(\tau), 0) + o_p(1),$$

where the $o_p(1)$ terms hold uniformly over $\tau \in \Upsilon$.

Proof. By Lemma S.K.1 with $b_n(\tau) = \beta_n(\tau)$, we have

$$\begin{pmatrix} \sqrt{n}\left(\hat{\gamma}(\beta_n(\tau), \tau) - \gamma_n(\tau)\right) \\ \sqrt{n}\hat{\theta}(\beta_n(\tau), \tau) \end{pmatrix} = \left[\Gamma_1(\beta_n(\tau), \tau)\right]^{-1} \sqrt{n}\left[I_n(\tau) + II_n(\beta_n(\tau), \tau) + o_p(1/\sqrt{n})\right],$$

where the $o_p(1/\sqrt{n})$ term holds uniformly over $\tau \in \Upsilon$. Then, by (102) and (103), we have

$$\begin{pmatrix} \sqrt{n}\left(\hat{\gamma}(\beta_n(\tau), \tau) - \gamma_n(\tau)\right) \\ \sqrt{n}\hat{\theta}(\beta_n(\tau), \tau) \end{pmatrix} = \mathcal{J}^{-1}_{\pi,\pi}(\tau) \sqrt{n} I_n(\tau) + o_p(1),$$

where the $o_p(1)$ terms hold uniformly over $\tau \in \Upsilon$. This concludes the proof. □

S.K.3 Technical Results for the Bootstrap IVQR Estimator

Lemma S.K.4. *Suppose Assumptions 5–7, 8(i)–8(iii) hold. Then,*

$$\begin{aligned}
&\sqrt{n}(\hat{\beta}_g^*(\tau) - \hat{\beta}(\tau)) \\
&= \tilde{\Omega}(\tau)\sum_{j\in J}\sqrt{\xi_j}g_j\sqrt{n_j}(\mathbb{P}_{n,j} - \overline{\mathbb{P}}_{n,j})\tilde{f}_\tau(D, \beta_n(\tau), \gamma_n(\tau), 0) \\
&- \sum_{j\in J}\Omega(\tau)\xi_j g_j \mathcal{J}_{\pi,\beta,j}(\tau)\sqrt{n}(\hat{\beta}^r(\tau) - \beta_n(\tau)) + o_p(1),
\end{aligned}$$

where $o_p(1)$ term holds uniformly over $\tau \in \Upsilon$ and

$$\Omega(\tau) = \left[\mathcal{J}_{\pi,\beta}^\top(\tau)\overline{\mathcal{J}}_\theta^\top(\tau)A_\phi(\tau)\overline{\mathcal{J}}_\theta(\tau)\mathcal{J}_{\pi,\beta}(\tau)\right]^{-1}\mathcal{J}_{\pi,\beta}^\top(\tau)\overline{\mathcal{J}}_\theta^\top(\tau)A_\phi(\tau)\overline{\mathcal{J}}_\theta(\tau).$$

Proof. We divide the proof into three steps. In the first step, we show the consistency of $(\hat{\beta}_g^*(\tau), \hat{\gamma}_g^*(\tau), \hat{\theta}_g^*(\tau))$. In the second step, we show

$$\begin{aligned}
&\sup_{\tau\in\Upsilon}||\sqrt{n}(\hat{\beta}_g^*(\tau) - \beta_n(\tau))||_2 = O_p(1), \\
&\sup_{\tau\in\Upsilon}||\sqrt{n}(\hat{\gamma}_g^*(\tau) - \gamma_n(\tau))||_2 = O_p(1), \\
&\sup_{\tau\in\Upsilon}||\sqrt{n}(\hat{\theta}_g^*(\tau) - 0)||_2 = O_p(1).
\end{aligned} \tag{105}$$

In the third step, we show the desired result.

Step 1. Note

$$\begin{aligned}
&\left\|\frac{1}{n}\sum_{j\in J}g_j\sum_{i\in I_{n,j}}\left(\hat{f}_\tau(D_{i,j}, \hat{\beta}^r(\tau), \hat{\gamma}^r(\tau), 0) - f_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0)\right)\right\|_2 \\
&\leq \sum_{j\in J}\xi_j\left\|(\mathbb{P}_{n,j} - \overline{\mathbb{P}}_{n,j})(\hat{f}_\tau(D_{i,j}, \hat{\beta}^r(\tau), \hat{\gamma}^r(\tau), 0) - \hat{f}_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0))\right\|_2 \\
&+ \sum_{j\in J}\xi_j\left\|\overline{\mathbb{P}}_{n,j}\hat{f}_\tau(D_{i,j}, \hat{\beta}^r(\tau), \hat{\gamma}^r(\tau), 0)\right\|_2.
\end{aligned} \tag{106}$$

Because $(\hat{\beta}^r(\tau), \hat{\gamma}^r(\tau), 0)$ are consistent as shown in Lemma S.K.5, Assumption 6 implies

$$\sup_{\tau\in\Upsilon}\sum_{j\in J}\xi_j\left\|(\mathbb{P}_{n,j} - \overline{\mathbb{P}}_{n,j})(\hat{f}_\tau(D_{i,j}, \hat{\beta}^r(\tau), \hat{\gamma}^r(\tau), 0) - \hat{f}_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0))\right\|_2 = o_p(n^{-1/2}). \tag{107}$$

In addition, due to Assumption 6 and Lemma S.K.5, following the same argument in (100), we can show that

$$\sup_{\tau\in\Upsilon}\left\|\overline{\mathbb{P}}_{n,j}\hat{f}_\tau(D_{i,j}, \hat{\beta}^r(\tau), \hat{\gamma}^r(\tau), 0) + \mathcal{J}_{\pi,\beta,j}(\tau)(\hat{\beta}^r(\tau) - \beta_n(\tau)) + \mathcal{J}_{\pi,\pi,j}(\tau)\begin{pmatrix}\hat{\gamma}^r(\tau) - \gamma_n(\tau) \\ 0\end{pmatrix}\right\|_2$$

$$= o_p(n^{-1/2}). \tag{108}$$

This further implies

$$\sup_{\tau \in \Upsilon} \sum_{j \in J} \xi_j \left\| \overline{\mathbb{P}}_{n,j} \hat{f}_\tau(D_{i,j}, \hat{\beta}^r(\tau), \hat{\gamma}^r(\tau), 0) \right\|_2 = o_p(1). \tag{109}$$

Note that

$$\begin{aligned}
(\hat{\gamma}_g^*(b,\tau), \hat{\theta}_g^*(b,\tau)) &= \arg\inf_{r,t} \Bigg[\sum_{j \in J} \sum_{i \in I_{n,j}} \rho_\tau(y_{i,j} - X_{i,j}^\top b - W_{i,j}^\top r - \hat{\Phi}_{i,j}^\top(\tau) t) \hat{V}_{i,j}(\tau) \\
&\quad - \sum_{j \in J} g_j \sum_{i \in I_{n,j}} \hat{f}_\tau^\top(D_{i,j}, \hat{\beta}^r(\tau), \hat{\gamma}^r(\tau), 0) \begin{pmatrix} r \\ t \end{pmatrix} \Bigg] \\
&= \arg \inf_{(r,t) \times \mathcal{B} \times \mathcal{R}} \hat{Q}_n(b,r,t,\tau) \\
&\quad - \frac{1}{n} \sum_{j \in J} g_j \sum_{i \in I_{n,j}} \left(\hat{f}_\tau^\top(D_{i,j}, \hat{\beta}^r(\tau), \hat{\gamma}^r(\tau), 0) - f^\top(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0) \right) \begin{pmatrix} r \\ t \end{pmatrix}.
\end{aligned}$$

Based on (107) and (109), uniformly over $(\tau, b) \in \Upsilon \times \mathcal{B}$,

$$\begin{aligned}
&\tilde{Q}_n(b,r,t,\tau) \\
\equiv &\hat{Q}_n(b,r,t,\tau) - \frac{1}{n} \sum_{j \in J} g_j \sum_{i \in I_{n,j}} \left(\hat{f}_\tau^\top(D_{i,j}, \hat{\beta}^r(\tau), \hat{\gamma}^r(\tau), 0) - f^\top(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0) \right) \begin{pmatrix} r \\ t \end{pmatrix} \\
\xrightarrow{p} &Q_\infty(b,r,t,\tau).
\end{aligned}$$

Then, because $\tilde{Q}_n(b,r,t,\tau)$ is convex in (r,t), by Kato (2009, Theorem 1), we have

$$\sup_{(b,\tau) \in \mathcal{B} \times \Upsilon} \left(||\hat{\gamma}_g^*(b,\tau) - \gamma_\infty(b,\tau)||_2 + ||\hat{\theta}_g^*(b,\tau) - \theta_\infty(b,\tau)||_2 \right) = o_p(1).$$

The rest of the proof is the same as that in Step 1 of the proof of Lemma S.K.2. We can show

$$\sup_{\tau \in \Upsilon} ||\hat{\beta}_g^*(\tau) - \beta_n(\tau)||_2 = o_p(1),$$

which further implies

$$\sup_{\tau \in \Upsilon} ||\hat{\gamma}_g^*(\tau) - \gamma_n(\tau)||_2 = o_p(1) \quad \text{and} \quad \sup_{\tau \in \Upsilon} ||\hat{\theta}_g^*(\tau) - 0||_2 = o_p(1).$$

Step 2. For any $b_n(\cdot) \in \mathcal{B}(\delta)$, the sub-gradient condition for $(\hat{\gamma}_g^*(b_n(\tau),\tau), \hat{\theta}_g^*(b_n(\tau),\tau))$ is

$$o_p(1/\sqrt{n}) = \mathbb{P}_n \hat{f}_\tau(D_{i,j}, b_n(\tau), \hat{\gamma}_g^*(b_n(\tau),\tau), \hat{\theta}_g^*(b_n(\tau),\tau), \tau) + \frac{1}{n} \sum_{j \in J} g_j \sum_{i \in I_{n,j}} \hat{f}_\tau(D_{i,j}, \hat{\beta}^r(\tau), \hat{\gamma}^r(\tau), 0), \tag{110}$$

where the $o_p(1/\sqrt{n})$ term on the LHS of the above display holds uniformly over $\tau \in \Upsilon, b_n(\cdot) \in$

$B(\delta)$.

Following the same argument in Lemma S.K.1, for any $\varepsilon > 0$, there exists $\overline{\delta}$ such that for $\delta, \delta' \leq \overline{\delta}$, we have, with probability greater than $1-\varepsilon$,

$$\begin{aligned}
&\mathbb{P}_n \hat{f}_\tau(D_{i,j}, b_n(\tau), \hat{\gamma}_g^*(b_n(\tau),\tau), \hat{\theta}_g^*(b_n(\tau),\tau),\tau) \\
&= I_n(\tau) + II_{n,g}(b_n(\tau),\tau) - \Gamma_{2,g}(b_n(\tau),\tau)(b_n(\tau)-\beta_n(\tau)) - \Gamma_{1,g}(b_n(\tau),\tau) \begin{pmatrix} \hat{\gamma}_g^*(b_n(\tau),\tau) - \gamma_n(\tau) \\ \hat{\theta}_g^*(b_n(\tau),\tau) \end{pmatrix},
\end{aligned} \tag{111}$$

where

$$I_n(\tau) = (\mathbb{P}_n - \overline{\mathbb{P}}_n) f_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0),$$

$$II_{n,g}(b_n(\tau),\tau) = \sum_{j\in J} \xi_j (\mathbb{P}_{n,j} - \overline{\mathbb{P}}_{n,j}) \left(\hat{f}_\tau(D_{i,j}, b_n(\tau), \hat{\gamma}_g^*(b_n(\tau),\tau), \hat{\theta}_g^*(b_n(\tau),\tau),\tau) - f_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0)\right)$$

such that $\sup_{b_n(\cdot)\in B(\delta), \tau\in\Upsilon} \sqrt{n}||II_{n,g}(b_n(\tau),\tau)||_2 \leq \varepsilon$,

$$\Gamma_{1,g}(b_n(\tau),\tau) = \sum_{j\in J} \frac{\xi_j}{n_j} \sum_{i\in I_{n,j}} \mathbb{E} f_{\varepsilon_{i,j}(\tau)}(\hat{\delta}_{i,j,g}(\tau)|W_{i,j}, Z_{i,j}) \hat{\Psi}_{i,j}(\tau) \hat{\Psi}_{i,j}^\top(\tau),$$

$$\Gamma_{2,g}(b_n(\tau),\tau) = \sum_{j\in J} \frac{\xi_j}{n_j} \sum_{i\in I_{n,j}} \mathbb{E} f_{\varepsilon_{i,j}(\tau)}(\hat{\delta}_{i,j,g}(\tau)|W_{i,j}, Z_{i,j}) \hat{\Psi}_{i,j}(\tau) X_{i,j}^\top,$$

$$\hat{\delta}_{i,j,g}(\tau) \in (0, X_{i,j}^\top(b_n(\tau)-\beta_n(\tau)) + W_{i,j}^\top(\hat{\gamma}_g^*(b_n(\tau),\tau) - \gamma_n(\tau)) + \hat{\Phi}_{i,j}^\top(\tau)\hat{\theta}_g^*(b_n(\tau),\tau)),$$

$$\sup_{b_n(\cdot)\in B(\delta), \tau\in\Upsilon, j\in J} \left\| \frac{1}{n_j} \sum_{i\in I_{n,j}} \mathbb{E} f_{\varepsilon_{i,j}(\tau)}(\hat{\delta}_{i,j,g}(\tau)|W_{i,j}, Z_{i,j}) \hat{\Psi}_{i,j}(\tau) \hat{\Psi}_{i,j}^\top(\tau) - \mathcal{J}_{\pi,\pi,j}(\tau) \right\|_{op} \leq \varepsilon,$$

and

$$\sup_{b_n(\cdot)\in B(\delta), \tau\in\Upsilon, j\in J} \left\| \frac{1}{n_j} \sum_{i\in I_{n,j}} \mathbb{E} f_{\varepsilon_{i,j}(\tau)}(\hat{\delta}_{i,j,g}(\tau)|W_{i,j}, Z_{i,j}) \hat{\Psi}_{i,j}(\tau) X_{i,j}^\top - \mathcal{J}_{\pi,\beta,j}(\tau) \right\|_{op} \leq \varepsilon.$$

In addition, by (106), (107), and (108), we have

$$\begin{aligned}
&\frac{1}{n}\sum_{j\in J} g_j \sum_{i\in I_{n,j}} \hat{f}_\tau(D_{i,j}, \hat{\beta}^r(\tau), \hat{\gamma}^r(\tau), 0) \\
&= \frac{1}{n}\sum_{j\in J} g_j \sum_{i\in I_{n,j}} \left(\hat{f}_\tau(D_{i,j}, \hat{\beta}^r(\tau), \hat{\gamma}^r(\tau), 0) - f_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0)\right) \\
&+ \frac{1}{n}\sum_{j\in J} g_j \sum_{i\in I_{n,j}} f_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0)
\end{aligned}$$

$$= \frac{1}{n}\sum_{j\in J} g_j \left[\sum_{i\in I_{n,j}} f_\tau(D_{i,j},\beta_n(\tau),\gamma_n(\tau),0) - \mathcal{J}_{\pi,\beta,j}(\tau)(\hat{\beta}^r(\tau)-\beta_n(\tau)) - \mathcal{J}_{\pi,\pi,j}(\tau)\begin{pmatrix}\hat{\gamma}^r(\tau)-\gamma_n(\tau)\\ 0\end{pmatrix}\right]$$
$$+ o_p(n^{-1/2}).$$

where the $o_p(n^{-1/2})$ term holds uniformly over $\tau \in \Upsilon$. Combining this with Assumption 6 and Lemma S.K.5, we have

$$\sup_{\tau\in\Upsilon}\left\|\frac{1}{n}\sum_{j\in J} g_j \sum_{i\in I_{n,j}} \hat{f}_\tau(D_{i,j},\hat{\beta}^r(\tau),\hat{\gamma}^r(\tau),0)\right\|_2 = O_p(n^{-1/2}).$$

In addition, (110) implies

$$o_p(1) = -\Gamma_{2,g}(b_n(\tau),\tau)\sqrt{n}(b_n(\tau)-\beta_n(\tau)) - \Gamma_{1,g}(b_n(\tau),\tau)\begin{pmatrix}\sqrt{n}(\hat{\gamma}^*_g(b_n(\tau),\tau)-\gamma_n(\tau))\\ \sqrt{n}\hat{\theta}^*_g(b_n(\tau),\tau)\end{pmatrix} + O_p(1), \tag{112}$$

where the $o_p(1)$ and $O_p(1)$ terms hold uniformly over $\{b_n(\cdot)\in B(\delta), \tau\in\Upsilon\}$. Then,

$$||\sqrt{n}\hat{\theta}^*_g(b_n(\tau),\tau)||^2_{\hat{A}_\phi(\tau)} = ||\omega\left[\Gamma_{1,g}(b_n(\tau),\tau)\right]^{-1}\left[\Gamma_{2,g}(b_n(\tau),\tau)\sqrt{n}(b_n(\tau)-\beta_n(\tau)) + O_p(1)\right]||^2_{\hat{A}_\phi(\tau)}.$$

Let $b_n(\tau) = \hat{\beta}^*_g(\tau)$, we have $\hat{\beta}^*_g(\tau)\in\mathcal{B}(\delta)$ w.p.a.1 for any $\delta>0$. Therefore,

$$||\omega\left[\Gamma_{1,g}(\hat{\beta}^*_g(\tau),\tau)\right]^{-1}\left[\Gamma_{2,g}(\hat{\beta}^*_g(\tau),\tau)\sqrt{n}(\hat{\beta}^*_g(\tau)-\beta_n(\tau)) + O_p(1)\right]||^2_{\hat{A}_\phi(\tau)} \le ||\sqrt{n}\hat{\theta}^*_g(\beta_n(\tau),\tau)||^2_{\hat{A}_\phi(\tau)}.$$

Because, w.p.a.1,

$$\inf_{\tau\in\Upsilon}\lambda_{\min}([\Gamma_{2,g}(\hat{\beta}^*_g(\tau),\tau)]^\top\left[\Gamma_{1,g}(\hat{\beta}^*_g(\tau),\tau)\right]^{-1}\omega^T\hat{A}_\phi(\tau)\omega\left[\Gamma_{1,g}(\hat{\beta}^*_g(\tau),\tau)\right]^{-1}\Gamma_{2,g}(\hat{\beta}^*_g(\tau),\tau)) \ge \underline{c} > 0$$

we have

$$\underline{c}n||\hat{\beta}^*_g(\tau)-\beta_n(\tau)||^2_2 - O_p(1) \le \sup_{\tau\in\Upsilon}||\sqrt{n}\hat{\theta}^*_g(\beta_n(\tau),\tau)||^2_{\hat{A}_\phi(\tau)} \le O_p(1),$$

where the $O_p(1)$ term holds uniformly over $\{\tau\in\Upsilon\}$. Therefore, we have

$$\sup_{\tau\in\Upsilon}\sqrt{n}||\hat{\beta}^*_g(\tau)-\beta_n(\tau)||_2 = O_p(1).$$

Plugging this into (112), we have

$$\sup_{\tau\in\Upsilon}\sqrt{n}||\hat{\gamma}^*_g(\tau)-\gamma_n(\tau)||_2 = O_p(1) \quad \text{and} \quad \sup_{\tau\in\Upsilon}\sqrt{n}||\hat{\theta}^*_g(\tau)||_2 = O_p(1).$$

Step 3. Let $b_n(\tau) = \beta_n(\tau) + u/\sqrt{n}$ in the above step, we have

$$o_p(1) = \sqrt{n}I_n(\tau) - \mathcal{J}_{\pi,\beta}(\tau)u - \mathcal{J}_{\pi,\pi}(\tau)\begin{pmatrix}\sqrt{n}(\hat{\gamma}^*_g(\beta_n(\tau)+u/\sqrt{n},\tau)-\gamma_n(\tau))\\ \sqrt{n}\hat{\theta}^*_g(\beta_n(\tau)+u/\sqrt{n},\tau)\end{pmatrix}$$

$$+\sqrt{n}I_{n,g}(\tau)-\sum_{j\in J}g_j\xi_j\mathcal{J}_{\pi,\beta,j}(\tau)\sqrt{n}(\hat{\beta}^r(\tau)-\beta_n(\tau))$$
$$-\sum_{j\in J}g_j\xi_j\mathcal{J}_{\pi,\pi,j}(\tau)\begin{pmatrix}\sqrt{n}(\hat{\gamma}^r(\tau)-\gamma_n(\tau))\\0\end{pmatrix}-o_p(1),$$

where

$$I_{n,g}(\tau)=\sum_{j\in J}\xi_j g_j(\mathbb{P}_{n,j}-\overline{\mathbb{P}}_{n,j})f_\tau(D_{i,j},\beta_n(\tau),\gamma_n(\tau),0)$$

such that $\sqrt{n}I_{n,g}(\tau)=\sum_{j\in J}\sqrt{\xi}_j g_j\sqrt{n_j}(\mathbb{P}_{n,j}-\overline{\mathbb{P}}_{n,j})f_\tau(D_{i,j},\beta_n(\tau),\gamma_n(\tau),0)$ and the $o_p(1)$ term holds uniformly over $\tau\in\Upsilon,|u|\leq M$. This implies

$$\begin{aligned}&\sqrt{n}\hat{\theta}^*_g(\beta_n(\tau)+u/\sqrt{n},\tau)\\&=\omega\mathcal{J}^{-1}_{\pi,\pi}(\tau)\Big[I_n(\tau)+I_{n,g}(\tau)-\mathcal{J}_{\pi,\beta}(\tau)u-\sum_{j\in J}g_j\xi_j\mathcal{J}_{\pi,\beta,j}(\tau)\sqrt{n}(\hat{\beta}^r(\tau)-\beta_n(\tau))\\&-\sum_{j\in J}g_j\xi_j\mathcal{J}_{\pi,\pi,j}(\tau)\begin{pmatrix}\sqrt{n}(\hat{\gamma}^r(\tau)-\gamma_n(\tau))\\0\end{pmatrix}-o_p(1)\Big]\\&=G_n(u,\tau)+o_p(1),\end{aligned}$$

where

$$\begin{aligned}G_n(u,\tau)=\omega\mathcal{J}^{-1}_{\pi,\pi}(\tau)\Big[&\sqrt{n}I_n(\tau)+\sqrt{n}I_{n,g}(\tau)-\mathcal{J}_{\pi,\beta}(\tau)u-\sum_{j\in J}g_j\xi_j\mathcal{J}_{\pi,\beta,j}(\tau)\sqrt{n}(\hat{\beta}^r(\tau)-\beta_n(\tau))\\&-\sum_{j\in J}g_j\xi_j\mathcal{J}_{\pi,\pi,j}(\tau)\begin{pmatrix}\sqrt{n}(\hat{\gamma}^r(\tau)-\gamma_n(\tau))\\0\end{pmatrix}\Big]\end{aligned}$$

and the $o_p(1)$ term holds uniformly over $\tau\in\Upsilon,|u|\leq M$. Let $\hat{u}^*_g(\tau)=\sqrt{n}(\hat{\beta}^*_g(\tau)-\beta_n(\tau))$. Because $\sup_{\tau\in\Upsilon}||\hat{u}_g||_2=O_p(1)$, for any $\varepsilon>0$, there exists an integer $\underline{n}$ such that for $n\geq\underline{n}$, there exists a sufficiently large constant $M>0$ such that

$$\hat{u}^*_g(\tau)=\arg\inf_{||u||_2\leq M}||\sqrt{n}\hat{\theta}^*_g(\beta_n(\tau)+u/\sqrt{n},\tau)||^2_{\hat{A}_\phi(\tau)}.$$

Because

$$\sup_{\tau\in\Upsilon,||u||_2\leq M}\left|||\sqrt{n}\hat{\theta}^*_g(\beta_n(\tau)+u/\sqrt{n},\tau)||^2_{\hat{A}_\phi(\tau)}-||G_n(u,\tau)||^2_{A_\phi(\tau)}\right|=o_p(1),$$

Chernozhukov and Hansen (2006, Lemma B.1) implies

$$\begin{aligned}\hat{u}^*_g(\tau)=&\left[\mathcal{J}^\top_{\pi,\beta}(\tau)\overline{\mathcal{J}}^\top_\theta(\tau)A_\phi(\tau)\overline{\mathcal{J}}_\theta(\tau)\mathcal{J}_{\pi,\beta}(\tau)\right]^{-1}\mathcal{J}^\top_{\pi,\beta}(\tau)\overline{\mathcal{J}}^\top_\theta(\tau)A_\phi(\tau)\overline{\mathcal{J}}_\theta(\tau)\\&\times\Big[\sqrt{n}I_n(\tau)+\sqrt{n}I_{n,g}(\tau)-\sum_{j\in J}g_j\xi_j\mathcal{J}_{\pi,\beta,j}(\tau)\sqrt{n}(\hat{\beta}^r(\tau)-\beta_n(\tau))\end{aligned}$$

$$- \sum_{j \in J} g_j \xi_j \mathcal{J}_{\pi,\pi,j}(\tau) \begin{pmatrix} \sqrt{n}(\hat{\gamma}^r(\tau) - \gamma_n(\tau)) \\ 0 \end{pmatrix} \Bigg] + o_p(1), \tag{113}$$

where the $o_p(1)$ term holds uniformly over $\tau \in \Upsilon$. Subtracting (104) from (113), we have

$$\sqrt{n}(\hat{\beta}_g^*(\tau) - \hat{\beta}(\tau))$$
$$= \Omega(\tau) \Bigg[\sqrt{n} I_{n,g}(\tau) - \sum_{j \in J} g_j \xi_j \mathcal{J}_{\pi,\beta,j}(\tau) \sqrt{n}(\hat{\beta}^r(\tau) - \beta_n(\tau)) - \sum_{j \in J} g_j \xi_j \mathcal{J}_{\pi,\pi,j}(\tau) \begin{pmatrix} \sqrt{n}(\hat{\gamma}^r(\tau) - \gamma_n(\tau)) \\ 0 \end{pmatrix} \Bigg]$$
$$+ o_p(1).$$

Assumption 7(i) implies $\mathcal{J}_{\pi,\pi,j}(\tau)$, and thus, $[\mathcal{J}_{\pi,\pi}(\tau)]^{-1} \mathcal{J}_{\pi,\pi,j}(\tau)$ are block diagonal, i.e.,

$$[\mathcal{J}_{\pi,\pi}(\tau)]^{-1} \mathcal{J}_{\pi,\pi,j}(\tau) = \begin{pmatrix} \mathcal{J}_{d_w \times d_w} & 0_{d_w \times d_\phi} \\ 0_{d_\phi \times d_w} & \mathcal{J}_{d_\phi \times d_\phi} \end{pmatrix}.$$

Then,

$$\overline{\mathcal{J}}_\theta(\tau) \mathcal{J}_{\pi,\pi,j}(\tau) \begin{pmatrix} \sqrt{n}(\hat{\gamma}^r(\tau) - \gamma_n(\tau)) \\ 0 \end{pmatrix} = \omega [\mathcal{J}_{\pi,\pi}(\tau)]^{-1} \mathcal{J}_{\pi,\pi,j}(\tau)$$
$$= [0_{d_\phi \times d_w}, \mathbb{I}_{d_\phi}] \begin{pmatrix} \mathcal{J}_{d_w \times d_w} & 0_{d_w \times d_\phi} \\ 0_{d_\phi \times d_w} & \mathcal{J}_{d_\phi \times d_\phi} \end{pmatrix} \begin{pmatrix} \sqrt{n}(\hat{\gamma}^r(\tau) - \gamma_n(\tau)) \\ 0_{d_\phi \times 1} \end{pmatrix} = 0.$$

In addition, for the same reason, we have

$$\Omega(\tau) f_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0) = \tilde{\Omega}(\tau) \tilde{f}_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0).$$

Therefore, we have

$$\sqrt{n}(\hat{\beta}_g^*(\tau) - \hat{\beta}(\tau))$$
$$= \Omega(\tau) \sqrt{n} I_{n,g}(\tau) - \sum_{j \in J} g_j \xi_j \Omega(\tau) \mathcal{J}_{\pi,\pi,j}(\tau) \sqrt{n}(\hat{\beta}^r(\tau) - \beta_n(\tau)) + o_p(1),$$
$$= \tilde{\Omega}(\tau) \sum_{j \in J} \sqrt{\xi_j} g_j \sqrt{n_j} (\mathbb{P}_{n,j} - \overline{\mathbb{P}}_{n,j}) \tilde{f}_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0)$$
$$- \sum_{j \in J} g_j \xi_j \Omega(\tau) \mathcal{J}_{\pi,\pi,j}(\tau) \sqrt{n}(\hat{\beta}^r(\tau) - \beta_n(\tau)) + o_p(1),$$

where the $o_p(1)$ term holds uniformly over $\tau \in \Upsilon$.

□

S.K.4 Technical Results for the Restricted Estimator

Lemma S.K.5. *Suppose Assumptions 5, 6, 8(i)–8(iv) hold. Then,*

$$\sup_{\tau\in\Upsilon}\left[||\hat{\beta}^r(\tau)-\beta_n(\tau)||_2+||\hat{\gamma}^r(\tau)-\gamma_n(\tau)||_2\right]=O_p(n^{-1/2}).$$

Proof. Suppose $h_0(\tau)$ satisfies $\lambda_\beta^\top(\tau)h_0(\tau)=\lambda_0(\tau)$, $L=d_x-d_r$, and $(h_1(\tau),\cdots,h_L(\tau))$ be L $d_x\times 1$ vectors that are the bases of a linear space that is orthogonal to the column space of $\lambda_\beta(\tau)$. Then, we have $\hat{\beta}^r(\tau)=h_0(\tau)+\sum_{l\in[L]}\hat{s}_l(\tau)h_l(\tau)$ and $\hat{s}(\tau)=(\hat{s}_1(\tau),\cdots,\hat{s}_L(\tau))^\top\in\mathbf{R}^L$, where

$$\hat{s}(\tau)=\arg\inf_{s\in\mathbf{R}^L,||s||_2\leq C}||\hat{\theta}(h_0(\tau)+\sum_{l\in[L]}s_lh_l(\tau),\tau)||_{\hat{A}_\phi(\tau)}.$$

In addition, we can write $\beta_0(\tau)=h_0(\tau)+\sum_{l\in[L]}s_l^*(\tau)h_l(\tau)$ such that $||s^*||_2\leq C$ with $s^*=(s_1^*,\cdots,s_L^*)^\top$. [28] Following the same argument in Step 1 in the proof of Lemma S.K.2, we have

$$\sup_{s:||s||_2\leq C}\left|||\hat{\theta}(h_0(\tau)+\sum_{l\in[L]}s_lh_l(\tau),\tau)||_{\hat{A}_\phi(\tau)}-||\theta_\infty(h_0(\tau)+\sum_{l\in[L]}s_lh_l(\tau),\tau)||_{A_\phi(\tau)}\right|=o_p(1).$$

As $||\theta_\infty(h_0(\tau)+\sum_{l\in[L]}s_lh_l(\tau),\tau)||_{A_\phi(\tau)}$ is uniquely minimized at $s^*(\tau)$, by Chernozhukov and Hansen (2006, Lemma B.1), we have

$$\sup_{\tau\in\Upsilon}||\hat{s}(\tau)-s^*(\tau)||_2=o_p(1),$$

which implies

$$\sup_{\tau\in\Upsilon}||\hat{\beta}^r(\tau)-\beta_n(\tau)||_2=o_p(1)\quad\text{and}\quad\sup_{\tau\in\Upsilon}||\hat{\gamma}^r(\tau)-\gamma_n(\tau)||_2=o_p(1).$$

Next, we turn to the convergence rate of $\hat{\beta}^r(\tau)$. By Step 2 in the proof of Lemma S.K.2 with $b_n(\cdot)=\hat{\beta}^r(\cdot)$, we have

$$\underline{c}||\sqrt{n}(\hat{\beta}^r(\tau)-\beta_n(\tau))||_2^2\leq||\sqrt{n}\hat{\theta}(\beta_0(\tau),\tau)||^2_{\hat{A}_\phi(\tau)},$$

because $\beta_0(\tau)=h_0(\tau)+\sum_{l\in[L]}s_l^*(\tau)h_l(\tau)$ satisfies the constraint. In addition, because $\sup_{\tau\in\Upsilon}||\beta_0(\tau)-\beta_n(\tau)||_2=O_p(n^{-1/2})$. By Step 3 in the proof of Lemma S.K.2, we have

$$||\sqrt{n}\hat{\theta}(\beta_0(\tau),\tau)||_{\hat{A}_\phi(\tau)}\leq||\sqrt{n}\hat{\theta}(\beta_n(\tau),\tau)||^2_{\hat{A}_\phi(\tau)}+O_p(1)=O_p(1),$$

where the $O_p(1)$ terms hold uniformly over $\tau\in\Upsilon$. This implies

$$\underline{c}||\sqrt{n}(\hat{\beta}^r(\tau)-\beta_n(\tau))||_2^2=O_p(1),$$

[28] It is w.o.l.g. to impose that $||s^*||_2\leq C$ for some constant $C>0$. We have already shown the unrestricted estimator $\hat{\beta}(\tau)$ is uniformly convergent over $\tau\in\Upsilon$. Then, the constant C can be calibrated by the length of the projection of $\hat{\beta}(\tau)-h_0(\tau)$ onto $\mathcal{M}(\tau)$.

which further implies

$$\sup_{\tau\in\Upsilon}||\hat{\beta}^r(\tau)-\beta_n(\tau)||_2 = O_p(n^{-1/2}).$$

□

S.K.5 Lemma used in the Proof of Theorem 3.2

Lemma S.K.6. *Suppose Assumptions 5–7, 8(i)–8(iv), and 11 hold. When Assumption 9(i) holds, we define $a_j(\tau)=\Omega(\tau)\mathcal{J}_{\pi,\beta,j}(\tau)=\tilde{\Omega}(\tau)\mathcal{J}_{\theta,\beta,j}(\tau)$, which is a scalar. When Assumption 9(ii) holds, $a_j(\tau)$ is as defined in Assumption 9(ii). Then,*

$$\begin{aligned}&\hat{\Omega}(\tau)\mathbb{P}_{n,j}\hat{f}_\tau(D_{i,j},\hat{\beta}(\tau),\hat{\gamma}(\tau),0)\\&=\tilde{\Omega}(\tau)\mathbb{P}_{n,j}\tilde{f}_\tau(D_{i,j},\beta_n(\tau),\gamma_n(\tau),0)-a_j(\tau)\tilde{\Omega}(\tau)\mathbb{P}_n\tilde{f}_\tau(D_{i,j},\beta_n(\tau),\gamma_n(\tau),0)+o_p(n^{-1/2}),\end{aligned}$$

$$\begin{aligned}&\hat{\Omega}(\tau)\mathbb{P}_{n,j}\hat{f}_\tau(D_{i,j},\hat{\beta}^*_g(\tau),\hat{\gamma}^*_g(\tau),0)\\&=\tilde{\Omega}(\tau)\mathbb{P}_{n,j}\tilde{f}_\tau(D_{i,j},\beta_n(\tau),\gamma_n(\tau),0)-\tilde{\Omega}(\tau)a_j(\tau)\mathbb{P}_n\tilde{f}_\tau(D_{i,j},\beta_n(\tau),\gamma_n(\tau),0)\\&-\tilde{\Omega}(\tau)a_j(\tau)\sum_{\tilde{j}\in J}g_{\tilde{j}}\xi_{\tilde{j}}\mathbb{P}_{n,j}\tilde{f}_\tau(D_{i,j},\beta_n(\tau),\gamma_n(\tau),0)+a_j(\tau)\overline{a}^*_g(\tau)(\hat{\beta}^r(\tau)-\beta_n(\tau))+o_p(n^{-1/2}).\end{aligned}$$

$$\hat{\Omega}(\tau)\mathbb{P}_{n,j}\hat{f}_\tau(D_{i,j},\hat{\beta}^r(\tau),\hat{\gamma}^r(\tau),0)=\Omega(\tau)\mathbb{P}_{n,j}f_\tau(D_{i,j},\beta_n(\tau),\gamma_n(\tau),0)-a_j(\tau)(\hat{\beta}^r(\tau)-\beta(\tau))+o_p(n^{-1/2}),$$

$$\begin{aligned}\hat{\Omega}(\tau)\mathbb{P}_{n,j}\overline{f}^*_{\tau,g}(D_{i,j})&=g_j\tilde{\Omega}(\tau)\mathbb{P}_{n,j}\tilde{f}_\tau(D_{i,j},\beta_n(\tau),\gamma_n(\tau),0)-g_ja_j(\tau)(\hat{\beta}^r(\tau)-\beta_n(\tau))\\&-\tilde{\Omega}(\tau)a_j(\tau)\sum_{\tilde{j}\in J}g_{\tilde{j}}\xi_{\tilde{j}}\mathbb{P}_{n,\tilde{j}}\tilde{f}_\tau(D_{i,\tilde{j}},\beta_n(\tau),\gamma_n(\tau),0)+a_j(\tau)\overline{a}^*_g(\tau)(\hat{\beta}^r(\tau)-\beta_n(\tau))\\&+o_p(n^{-1/2}),\end{aligned}$$

where all the $o_p(n^{-1/2})$ terms hold uniformly over $\tau\in\Upsilon$ and $\overline{a}^*_g(\tau)=\sum_{j\in J}\xi_jg_ja_j(\tau)$.

Proof. Recall

$$\Omega(\tau)=\left[\mathcal{J}^\top_{\pi,\beta}(\tau)\overline{\mathcal{J}}^\top_\theta(\tau)A_\phi(\tau)\overline{\mathcal{J}}_\theta(\tau)\mathcal{J}_{\pi,\beta}(\tau)\right]^{-1}\mathcal{J}^\top_{\pi,\beta}(\tau)\overline{\mathcal{J}}^\top_\theta(\tau)A_\phi(\tau)\overline{\mathcal{J}}_\theta(\tau).$$

We have $\sup_{\tau\in\Upsilon}||\hat{\Omega}(\tau)-\Omega(\tau)||_{op}=o_p(1)$. For the first result, we have

$$\begin{aligned}&\mathbb{P}_{n,j}\hat{f}_\tau(D_{i,j},\hat{\beta}(\tau),\hat{\gamma}(\tau),0)\\&=\overline{\mathbb{P}}_{n,j}\hat{f}_\tau(D_{i,j},\hat{\beta}(\tau),\hat{\gamma}(\tau),0)+(\mathbb{P}_{n,j}-\overline{\mathbb{P}}_{n,j})\hat{f}_\tau(D_{i,j},\hat{\beta}(\tau),\hat{\gamma}(\tau),0)\\&=\overline{\mathbb{P}}_{n,j}\hat{f}_\tau(D_{i,j},\hat{\beta}(\tau),\hat{\gamma}(\tau),0)+(\mathbb{P}_{n,j}-\overline{\mathbb{P}}_{n,j})f_\tau(D_{i,j},\beta_n(\tau),\gamma_n(\tau),0)+o_p(n^{-1/2})\\&=\mathbb{P}_{n,j}f_\tau(D_{i,j},\beta_n(\tau),\gamma_n(\tau),0)-\mathcal{J}_{\pi,\beta,j}(\tau)(\hat{\beta}(\tau)-\beta_n(\tau))-\mathcal{J}_{\pi,\pi,j}(\tau)\begin{pmatrix}\hat{\gamma}(\tau)-\gamma_n(\tau)\\0\end{pmatrix}+o_p(n^{-1/2})\\&=\mathbb{P}_{n,j}f_\tau(D_{i,j},\beta_n(\tau),\gamma_n(\tau),0)-\mathcal{J}_{\pi,\beta,j}(\tau)\tilde{\Omega}(\tau)\mathbb{P}_n\tilde{f}_\tau(D_{i,j},\beta_n(\tau),\gamma_n(\tau),0)\end{aligned}$$

$$- \mathcal{J}_{\pi,\pi,j}(\tau) \begin{pmatrix} \hat{\gamma}(\tau) - \gamma_n(\tau) \\ 0 \end{pmatrix} + o_p(n^{-1/2}), \tag{114}$$

where the $o_p(n^{-1/2})$ terms hold uniformly over $\tau \in \Upsilon$, the second equality is by Assumption 6(iii), the third equality is by Assumption 6(ii) and the fact that

$$\sup_{\tau\in\Upsilon} \left(||\hat{\beta}(\tau) - \beta_n(\tau)||_2 + ||\hat{\gamma}(\tau) - \gamma_n(\tau)||_2 \right) = O_p(n^{-1/2})$$

as shown in Lemma S.K.2, and the last equality is by Lemma S.K.2 and the fact that, by Assumption 7(i),

$$\Omega(\tau) f_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0) = \tilde{\Omega}(\tau) \tilde{f}_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0).$$

In addition, due to the argument in Step 3 in the proof of Lemma S.K.4, under Assumption 7(i), we have

$$\overline{\mathcal{J}}_\theta(\tau) \mathcal{J}_{\pi,\pi,j}(\tau) \begin{pmatrix} \hat{\gamma}(\tau) - \gamma_n(\tau) \\ 0_{d_\phi \times 1} \end{pmatrix} = 0.$$

This implies

$$\Omega(\tau) \mathcal{J}_{\pi,\pi,j}(\tau) \begin{pmatrix} \hat{\gamma}(\tau) - \gamma_n(\tau) \\ 0 \end{pmatrix} = 0,$$

and thus,

$$\begin{aligned}
&\hat{\Omega}(\tau) \mathbb{P}_{n,j} \hat{f}_\tau(D_{i,j}, \hat{\beta}(\tau), \hat{\gamma}(\tau), 0) \\
&= \Omega(\tau) \mathbb{P}_{n,j} f_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0) - \Omega(\tau) \mathcal{J}_{\pi,\beta,j}(\tau) \tilde{\Omega}(\tau) \mathbb{P}_n \tilde{f}_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0) + o_p(n^{-1/2}) \\
&= \tilde{\Omega}(\tau) \mathbb{P}_{n,j} \tilde{f}_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0) - a_j(\tau) \tilde{\Omega}(\tau) \mathbb{P}_n \tilde{f}_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0) + o_p(n^{-1/2}),
\end{aligned}$$

where the $o_p(n^{-1/2})$ terms hold uniformly over $\tau \in \Upsilon$ and the last equality holds due to the fact that

$$\Omega(\tau) \mathcal{J}_{\pi,\beta,j}(\tau) = \tilde{\Omega}(\tau) \mathcal{J}_{\theta,\beta,j}(\tau) = a_j(\tau) \mathbb{I}_{\beta,\beta}.$$

For the second result, we have

$$\begin{aligned}
&\mathbb{P}_{n,j} \hat{f}_\tau(D_{i,j}, \hat{\beta}_g^*(\tau), \hat{\gamma}_g^*(\tau), 0) \\
&= \mathbb{P}_{n,j} f_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0) - \mathcal{J}_{\pi,\beta,j}(\tau)(\hat{\beta}_g^*(\tau) - \beta_n(\tau)) - \mathcal{J}_{\pi,\pi,j}(\tau) \begin{pmatrix} \hat{\gamma}_g^*(\tau) - \gamma_n(\tau) \\ 0 \end{pmatrix} + o_p(n^{-1/2}) \\
&= \mathbb{P}_{n,j} f_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0) - \mathcal{J}_{\pi,\beta,j}(\tau) \Omega(\tau) \sum_{\tilde{j} \in J} g_{\tilde{j}} \xi_{\tilde{j}} \mathbb{P}_{n,j} f_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0) \\
&- \mathcal{J}_{\pi,\beta,j}(\tau)(\hat{\beta}(\tau) - \beta_n(\tau)) + \mathcal{J}_{\pi,\beta,j}(\tau) \overline{a}_g^*(\tau)(\hat{\beta}^r(\tau) - \beta_n(\tau))
\end{aligned}$$

$$- \mathcal{J}_{\pi,\pi,j}(\tau) \begin{pmatrix} \hat{\gamma}_g^*(\tau) - \gamma_n(\tau) \\ 0 \end{pmatrix} + o_p(n^{-1/2}),$$

where the first equality is due to the same argument in (114) and the second equality is due to Lemma S.K.4. Then, we have

$$\begin{aligned}
&\hat{\Omega}(\tau)\mathbb{P}_{n,j}\hat{f}_\tau(D_{i,j}, \hat{\beta}_g^*(\tau), \hat{\gamma}_g^*(\tau), 0) \\
&= \tilde{\Omega}(\tau)\mathbb{P}_{n,j}\tilde{f}_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0) - \tilde{\Omega}(\tau)a_j(\tau)\mathbb{P}_n\tilde{f}_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0) \\
&- \tilde{\Omega}(\tau)a_j(\tau)\sum_{\tilde{j}\in J} g_{\tilde{j}}\xi_{\tilde{j}}\mathbb{P}_{n,\tilde{j}}\tilde{f}_\tau(D_{i,\tilde{j}}, \beta_n(\tau), \gamma_n(\tau), 0) + a_j(\tau)\overline{a}_g^*(\tau)(\hat{\beta}^r(\tau) - \beta_n(\tau)) + o_p(n^{-1/2}).
\end{aligned}$$

For the third result, we have

$$\begin{aligned}
&\mathbb{P}_{n,j}\hat{f}_\tau(D_{i,j}, \hat{\beta}^r(\tau), \hat{\gamma}^r(\tau), 0) \\
&= \mathbb{P}_{n,j}f_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0) - \mathcal{J}_{\pi,\beta,j}(\tau)(\hat{\beta}^r(\tau) - \beta_n(\tau)) - \mathcal{J}_{\pi,\pi,j}(\tau) \begin{pmatrix} \hat{\gamma}^r(\tau) - \gamma_n(\tau) \\ 0 \end{pmatrix} + o_p(n^{-1/2})
\end{aligned}$$

Then,

$$\begin{aligned}
&\hat{\Omega}(\tau)\mathbb{P}_{n,j}\hat{f}_\tau(D_{i,j}, \hat{\beta}^r(\tau), \hat{\gamma}^r(\tau), 0) \\
&= \tilde{\Omega}(\tau)\mathbb{P}_{n,j}\tilde{f}_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0) - a_j(\tau)(\hat{\beta}^r(\tau) - \beta_n(\tau)) + o_p(n^{-1/2}).
\end{aligned}$$

Combining the previous three results, we have

$$\begin{aligned}
\hat{\Omega}(\tau)\mathbb{P}_{n,j}\overline{f}_{\tau,g}^*(D_{i,j}) &= g_j\tilde{\Omega}(\tau)\mathbb{P}_{n,j}\tilde{f}_\tau(D_{i,j}, \beta_n(\tau), \gamma_n(\tau), 0) - g_j a_j(\tau)(\hat{\beta}^r(\tau) - \beta_n(\tau)) \\
&- \tilde{\Omega}(\tau)a_j(\tau)\sum_{\tilde{j}\in J} g_{\tilde{j}}\xi_{\tilde{j}}\mathbb{P}_{n,\tilde{j}}\tilde{f}_\tau(D_{i,\tilde{j}}, \beta_n(\tau), \gamma_n(\tau), 0) + a_j(\tau)\overline{a}_g^*(\tau)(\hat{\beta}^r(\tau) - \beta_n(\tau)) \\
&+ o_p(n^{-1/2}).
\end{aligned}$$

All the $o_p(n^{-1/2})$ terms in this proof hold uniformly over $\tau \in \Upsilon$. □

References

Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge (2017): "When should you adjust standard errors for clustering?" Tech. rep., National Bureau of Economic Research.

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

Alejo, J., A. F. Galvao, and G. Montes-Rojas (2021): "A first-stage representation for instrumental variables quantile regression," *arXiv preprint arXiv:2102.01212*.

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., G. Imbens, and A. Krueger (1999): "Jackknife instrumental variables estimates," *Journal of Applied Econometrics*, 14, 57–67.

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.

Brodeur, A., N. Cook, and A. Heyes (2020): "Methods matter: P-hacking and publication bias in causal analysis in economics," *American Economic Review*.

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

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

Chernozhukov, V. and C. Hansen (2004): "The effects of 401(k) participation on the wealth distribution: an instrumental quantile regression analysis," *Review of Economics and statistics*, 86, 735–751.

——— (2005): "An IV model of quantile treatment effects," *Econometrica*, 73, 245–261.

——— (2006): "Instrumental quantile regression inference for structural and treatment effect models," *Journal of Econometrics*, 132, 491–525.

——— (2008a): "Instrumental variable quantile regression: A robust inference approach," *Journal of Econometrics*, 142, 379–398.

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

Chernozhukov, V., C. Hansen, and K. Wüthrich (2020): "Instrumental variable quantile regression," *arXiv preprint arXiv:2009.00436*.

Chetverikov, D., B. Larsen, and C. Palmer (2016): "IV quantile regression for group-level treatments, with an application to the distributional effects of trade," *Econometrica*, 84, 809–833.

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

de Castro, L., A. F. Galvao, D. M. Kaplan, and X. Liu (2019): "Smoothed GMM for quantile models," *Journal of Econometrics*, 213, 121–144.

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

——— (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.

Galvao, A. F., J. Gu, and S. Volgushev (2020): "On the unbiased asymptotic normality of quantile regression with fixed effects," *Journal of Econometrics*, 218, 178–215.

Galvao, A. F. and K. Kato (2016): "Smoothed quantile regression for panel data," *Journal of Econometrics*, 193, 92–112.

Goldsmith-Pinkham, P., I. Sorkin, and H. Swift (2020): "Bartik instruments: What, when, why, and how," *American Economic Review*, 110, 2586–2624.

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

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*.

Hahn, J. and Z. Liao (2021): "Bootstrap standard error estimates and inference," *Econometrica*, 89, 1963–1977.

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

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

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

Jiang, L., X. Liu, P. Phillips, and Y. Zhang (2021): "Bootstrap inference for quantile treatment effects in randomized experiments with matched pairs," *Review of Economics and Satistics*, 1–43.

Kaido, H. and K. Wüthrich (2021): "Decentralization estimators for instrumental variable quantile regression models," *Quantitative Economics*, 12, 443–475.

Kato, K. (2009): "Asymptotics for argmin processes: Convexity arguments," *Journal of Multivariate Analysis*, 100, 1816–1829.

Kato, K., A. F. Galvao, and G. V. Montes-Rojas (2012): "Asymptotics for panel quantile regression models with individual effects," *Journal of Econometrics*, 170, 76–91.

Kiefer, N. M. and T. J. Vogelsang (2002): "Heteroskedasticity-autocorrelation robust standard errors using the Bartlett kernel without truncation," *Econometrica*, 70, 2093–2095.

——— (2005): "A new asymptotic theory for heteroskedasticity-autocorrelation robust tests," *Econometric Theory*, 1130–1164.

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.

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

Lehmann, E. L. and J. P. Romano (2006): *Testing statistical hypotheses*, Springer Science & Business Media.

MacKinnon, J. G., M. Nielsen, and M. D. Webb (2020): "Cluster-robust inference: A guide to empirical practice," Tech. rep., Qed working paper, Queen's University.

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

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.

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

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

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

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

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

Phillips, G. and C. Hale (1977): "The bias of instrumental variable estimators of simultaneous equation systems," *International Economic Review*, 18, 219–228.

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

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: journal of the Econometric Society*, 557–586.

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

Su, L. and T. Hoshino (2016): "Sieve instrumental variable quantile regression estimation of functional coefficient models," *Journal of Econometrics*, 191, 231–254.

van der Vaart, A. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer, New York.

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.

Wüthrich, K. (2019): "A closed-form estimator for quantile treatment effects with endogeneity," *Journal of Econometrics*, 210, 219–235.

Young, A. (2021): "Leverage, Heteroskedasticity and Instrumental Variables in Practical Application," Tech. rep., London School of Economics.