

Simple and Trustworthy Cluster-Robust GMM Inference

Jungbin Hwang*

Department of Economics,
University of Connecticut

April 8, 2020

Abstract

This paper develops a new asymptotic theory for GMM estimation and inference in the presence of clustered dependence. The key feature of alternative asymptotics is the number of clusters G is regarded as fixed when the sample size increases. Under the fixed- G asymptotics, we show that Wald and t -tests in two-step GMM are asymptotically pivotal only if we recenter the estimated moment process in clustered covariance estimator (CCE). Also, the J statistic, the trinity of two-step GMM statistics (QLR, LM, and Wald), and the t statistic can be modified to have an asymptotic standard F distribution or t distribution. We suggest a finite sample variance correction to further improve the accuracy of the F and t approximations. Our proposed tests are very appealing to practitioners because our test statistics are simple modifications of the usual test statistics, and critical values are readily available from standard statistical tables. No further simulations or re-sampling methods are needed. A Monte Carlo study shows that our proposed tests are more accurate than the conventional inferences under the large- G asymptotics.

JEL Classification: C12, C21, C23, C31

Keywords: Two-step GMM, Heteroskedasticity and Autocorrelation Robust, Clustered Dependence, t distribution, F distribution

*Email: jungbin.hwang@uconn.edu. Correspondence to: Jungbin Hwang, Department of Economics, University of Connecticut, 365 Fairfield Way U-1063, Storrs, CT 06269-1063. I thank the coeditor, an associate editor, and anonymous referees for constructive comments. I am also indebted to Yixiao Sun, Graham Elliott, Andres Santos, and Min Seong Kim that helped greatly improve this paper. I would also like to thank seminar participants at Boston University, University of California San Diego, University of Connecticut, University of California Irvine, University of Illinois at Urbana-Champaign, National University of Singapore, Korea University, and Pennsylvania State University. I acknowledge partial research support from NSF under Grant No. SES-1530592.

1 Introduction

Clustering is a common feature for many cross-sectional and panel data sets in applied economics. The data often come from a number of independent clusters with a general dependence structure within each cluster. For example, in development economics, data are often clustered by geographical regions, such as village, county and province, and, in empirical finance and industrial organization, firm level data are often clustered at the industry level. Because of learning from daily interactions, the presence of common shocks, and for many other reasons, individuals in the same cluster will be interdependent while those from different clusters tend to be independent. Failure to control for within group or cluster correlation often leads to downwardly biased standard errors and spurious statistical significance.

Seeking to robustify inference, many practical methods employ clustered covariance estimators (CCE). See White (1984, Theorem 6.3, p. 136), Liang and Zeger (1986), Arellano (1987) for seminal methodological contributions, and Wooldridge (2003) and Cameron and Miller (2015) for overviews of the CCE and its applications. It is now well known that standard test statistics based on the CCE are either asymptotically chi-squared or normal. The chi-squared and normal approximations are obtained under the so-called large- G asymptotic specification, which requires the number of clusters G to grow with the sample size. The key ingredient behind these approximations is that the CCE becomes concentrated at the true asymptotic variance as G approaches to infinity. In effect, this type of asymptotics ignores the estimation uncertainty in the CCE despite its high variation in finite samples, especially when the number of clusters is small. In practice, however, it is not unusual to have a data set that has a small number of clusters. For example, if clustering is based on large geographical regions such as U.S. states and regional blocks of neighboring countries, (e.g., Bertrand, Duflo, and Mullainathan, 2004; Ibragimov and Müller, 2015), we cannot convincingly claim that the number of cluster is large so that the large- G asymptotic approximations are applicable. In fact, there is ample simulation evidence that the large- G approximation can be very poor when the number of clusters is not large (e.g., Donald and Lang, 2007; Cameron, Gelbach, and Miller, 2008; Bester, Conley, and Hansen, 2011; Mackinnon and Webb, 2017).

In this paper, we adopt an alternative approach that yields more accurate approximations, and that works well whether or not the number of clusters is large. Our approximations work especially well when the chi-squared and normal approximations are poor. They are obtained from an alternative limiting thought experiment where the number of clusters G is held fixed. Under this fixed (small)- G asymptotics, the CCE no longer asymptotically degenerates; instead, it converges in distribution to a random matrix that is proportional to the true asymptotic variance. The random limit of the CCE has profound implications for the analyses of the asymptotic properties of GMM estimators and the corresponding test statistics.

We start with the first-step GMM estimator where the underlying model is possibly over-identified and show that suitably modified Wald and t statistics converge weakly to standard F and t distributions, respectively. The modification is easy to implement because it involves only a known multiplicative factor. Similar results have been obtained by Hansen (2007) and Bester, Conley and Hansen (2011), which employ a CCE type HAC estimator but consider only linear regressions and M -estimators for an exactly identified model.

We then consider the two-step GMM estimator that uses the CCE as a weighting matrix. Because the weighting matrix is random even in the limit, the two-step estimator is not asymptotically normal. The form of the limiting distribution depends on how the CCE is constructed. If the CCE is based on the uncentered moment process, we obtain the so-called uncentered two-step

GMM estimator. We show that the asymptotic distribution of this two-step GMM estimator is highly nonstandard. As a result, the associated Wald and t statistics are not asymptotically pivotal. However, it is surprising that the J statistic is still asymptotically pivotal, and its limiting distribution can be represented as an increasing function of a standard F random variable.

Next, we establish the asymptotic properties of the “centered” two-step GMM estimator¹ whose weighting matrix is constructed using recentered moment conditions. Invoking centering is not innocuous for an over-identified GMM model because the empirical moment conditions, in this case, are not equal to zero in general. Under the traditional large- G asymptotics, recentering does not matter in large samples because the empirical moment conditions are asymptotically zero and here are ignorable, even though they are not identically zero in finite sample. In contrast, under the fixed- G asymptotics, recentering plays two important roles: it removes the first order effect of the estimation error in the first-step estimator, and it ensures that the weighting matrix is asymptotically independent of the empirical moment conditions. With the recentered CCE as the weighting matrix, the two-step GMM estimator is asymptotically mixed normal. The mixed normality reflects the high variation of the feasible two-step GMM estimator as compared to the infeasible two-step GMM estimator, which is obtained under the assumption that the ‘efficient’ weighing matrix is known. The mixed-normality allows us to construct the Wald and t statistics that are asymptotically nuisance parameter free.

To relate the nonstandard fixed- G asymptotic distributions to standard distributions, we introduce simple modifications to the Wald and t statistics associated with the centered two-step GMM estimator. We show that the modified Wald and t statistics are asymptotically F and t distributed, respectively. This result resembles the corresponding result that is based on the first-step GMM estimator. It is important to point out that the proposed modifications are indispensable for our asymptotic F and t theory. In the absence of the modifications, the Wald and t statistics converge in distribution to nonstandard distributions, and as a result, critical values have to be simulated. The modifications involve only the standard J statistic, and it is very easy to implement because the modified test statistics are scaled versions of the original Wald test statistics with the scaling factor depending on the J statistic. Significantly, the combination of the Wald statistic and the J statistic enables us to develop the F approximation theory. We also find that the uncentered continuously updating (CU) GMM estimators and the centered two-step GMM estimator are asymptotically equivalent under the fixed- G asymptotics. Thus, the CU estimators can be regarded as having a built-in recentering mechanism similar to the centered two-step GMM estimator.

Finally, although the recentering scheme removes the first order effect of the first-step estimation error, the centered two-step GMM still face some extra estimation uncertainty in the first-step estimator. The main source of the problem is that we have to estimate the unobserved moment process based on the first-step estimator. To capture the higher order effect, we propose to retain one more term in our stochastic approximation that is asymptotically negligible. The expansion helps us develop a finite sample correction to the asymptotic variance estimator. Our correction resembles that of Windmeijer (2005), which considers variance correction for a two-step GMM estimator but valid only in an i.i.d. setting. We show that the finite sample variance correction does not change the fixed- G limiting distributions of the test statistics, but they can help improve the finite sample performance of our tests.

Monte Carlo simulations show that our new tests have a much more accurate size than existing

¹Our definition of the centered two-step GMM estimator is originated from the recentered (or demeaned) GMM weighting matrix, and it should not be confused with “centering” the estimator itself.

tests via standard normal and chi-squared critical values, especially when the number of clusters G is not large. An advantage of our procedure is that the test statistics do not entail much extra computational cost because the main ingredient for the modification is the usual J statistic. There is also no need to simulate critical values because the F and t critical values can be readily obtained from standard statistical tables.

Our fixed- G asymptotics is related to fixed-smoothing asymptotics for a long run variance (LRV) estimation in a time series setting. The latter was initiated and developed in econometric literature by Kiefer, Vogelsang and Bunzel (2002), Kiefer and Vogelsang (2005), Müller (2007), Sun, Phillips and Jin (2008), Sun (2013, 2014), Zhang (2016), among others. Our new asymptotics is in the same spirit in that both lines of research attempt to capture the estimation uncertainty in covariance estimation.

A recent paper by Hwang and Sun (2017a) also modifies the two-step GMM statistics using the J statistic. The F and t limit theory in Hwang and Sun (2017a) is seemingly similar to those presented in this paper. However, our asymptotic theory differs from Hwang and Sun (2017a), as it delivers the following sophistication in the CCE-type HAR inference—we can have asymptotically pivotal Wald and t inferences only if we use the centered CCE in the test statistics. We explicitly show that the uncentered two-step test statistics result in highly non-standard and non-pivotal fixed- G limits. This sophistication does not appear in the asymptotics of Hwang and Sun (2017), because its zero-mean basis functions yield the same limiting distributions regardless of using centered or uncentered HAR estimators. The cluster-robust limiting distributions in our paper also differ from those of Hwang and Sun (2017) in terms of the multiplicative adjustment and the degrees of freedom. Moreover, we propose a finite sample variance correction to capture the uncertainty embodied in the estimated moment process adequately. To our knowledge, the finite sample variance correction provided in this paper and its first order asymptotic validity has not been considered in the literature on the fixed-smoothing asymptotics.

There is also a growing literature that uses the fixed- G asymptotics to design more accurate cluster-robust inference. For instance, Ibragimov and Müller (2010, 2016) recently proposes a subsample based t test for a scalar parameter that is robust to heterogeneous clusters. Hansen (2007), Stock and Watson (2008), and Bester, Conley and Hansen (2011) propose a cluster-robust F or t tests under cluster-size homogeneity. Imbens and Kolesár (2016) suggest an adjusted t critical value employing data-determined degrees of freedom. Recently, Canay, Romano and Shaikh (2017) and Canay, Santos, and Shaikh (2019), Hagemann (2019) establish a theory of randomization inferences in the context of clustered dependence. For other approaches, see Carter, Schnepel and Steigerwald (2017) which proposes a measure of the effective number of clusters under the large- G asymptotics; Cameron, Gelbach and Miller (2008), MacKinnon and Webb (2017), and Djogbenou, MacKinnon, and Nielsen (2019) which investigate cluster-robust bootstrap approaches. All these studies, however, mainly focus on a simple location model or linear regressions that are special cases of exactly identified models. A recent contribution which develops a large- G cluster asymptotic distribution theory is Hansen and Lee (2019).

The remainder of the paper is organized as follows. Section 2 presents the basic setting and establishes the approximation results for the first-step GMM estimator under the fixed- G asymptotics. Sections 3 and 4 establish the fixed- G asymptotics for two-step GMM estimators and develop the asymptotic F and t tests based on the centered two-step GMM estimator. Section 5 proposes a finite sample variance correction. The next two sections reports a simulation evidence and applies our cluster-robust tests to an empirical study in Emran and Hou (2013). The last section concludes. Proofs the main results are given in Appendix A, and online supplemental

Appendix B available at the author’s website² contains extensions of theories, e.g., clustered dependence in a spatial setting, asymptotically unbalanced sizes in clusters, LM and QLR type GMM tests and continuously updating GMMs, and omitted proofs in the main text.

2 Basic Setting and the First-step GMM Estimator

We want to estimate the $d \times 1$ vector of parameters $\theta \in \Theta$. The true parameter vector θ_0 is assumed to be an interior point of parameter space $\Theta \subseteq \mathbb{R}^d$. Suppose that we observe cross-sectionally dependent triangular array of random vectors $Y_{i,n} \in \mathbb{R}^{d_y}$ for $i = 1, \dots, n$, which satisfy the following moment condition

$$Ef(Y_{i,n}, \theta) = 0 \text{ if and only if } \theta = \theta_0, \quad (1)$$

where $f_{i,n}(\theta) = f(Y_{i,n}, \theta)$ is an $m \times 1$ vector of twice continuously differentiable functions. Throughout the paper, we suppress the dependence of n on $Y_{i,n}$ and $f_{i,n}(\theta)$ for notational simplicity. We assume that $q = m - d \geq 0$ and the rank of $\Gamma = E[\partial f(Y_i, \theta_0)/\partial \theta']$ is d . So the model is possibly over-identified with the degree of over-identification q . The number of observations is n . Define $g_n(\theta) = n^{-1} \sum_{i=1}^n f_i(\theta)$. Given the moment condition in (1), the initial “first-step” GMM estimator of θ_0 is given by

$$\hat{\theta}_1 = \arg \min_{\theta \in \Theta} g_n(\theta)' W_n^{-1} g_n(\theta),$$

where W_n is an $m \times m$ positive definite and a symmetric weighting matrix that does not depend on the unknown parameter θ_0 and $\text{plim}_{n \rightarrow \infty} W_n = W > 0$. In the context of instrumental variable (IV) regression, one popular choice for W_n is $(Z_n' Z_n/n)^{-1}$ where Z_n is the data matrix of instruments.

Let $\hat{\Gamma}(\theta) = n^{-1} \sum_{i=1}^n \frac{\partial f_i(\theta)}{\partial \theta'}$. To establish the asymptotic properties of $\hat{\theta}_1$, we assume that for any \sqrt{n} -consistent estimator $\tilde{\theta}$, $\text{plim}_{n \rightarrow \infty} \hat{\Gamma}(\tilde{\theta}) = \Gamma$ and that Γ is of full column rank. Also, under some regularity conditions, we have the following Central Limit Theorem (CLT)

$$\begin{aligned} \sqrt{n} g_n(\theta_0) &\xrightarrow{d} N(0, \Omega), \text{ where} \\ \Omega &= \lim_{n \rightarrow \infty} \frac{1}{n} E \left(\sum_{i=1}^n f_i(\theta_0) \right) \left(\sum_{i=1}^n f_i(\theta_0) \right)'. \end{aligned}$$

Here Ω is analogous to the long run variance in a time series setting but the components of Ω are contributed by cross-sectional dependences over all locations. For easy reference, we follow Sun and Kim (2015) and call Ω the global variance. Primitive conditions for the above CLT in the presence of weak cross-sectional dependence are provided in supplementary Appendix B.1 of the paper. Under these conditions, we have

$$\sqrt{n}(\hat{\theta}_1 - \theta_0) \xrightarrow{d} N \left[0, (\Gamma' W^{-1} \Gamma)^{-1} \Gamma' W^{-1} \Omega W^{-1} \Gamma (\Gamma' W^{-1} \Gamma)^{-1} \right].$$

Since Γ and W can be accurately estimated by $\hat{\Gamma}(\hat{\theta}_1)$ and W_n relative to Ω , we only need to estimate Ω to make reliable inference about θ_0 . The main issue is how to properly account for

²https://hwang.econ.uconn.edu/wp-content/uploads/sites/1837/2020/04/Supplemental_Appendix_Hwang_March_2020.pdf

cross-sectional dependence in the moment process $\{f_j(\theta_0)\}_{j=1}^n$. In this paper, we assume that the cross-sectional dependence has a cluster structure, which is popular in many microeconomic applications. More specifically, our data consists of a number of independent clusters, each of which has an unknown dependence structure. Let G be the total number of clusters and L_g be the size of cluster g .

Assumption 1 (i) The data $\{Y_i\}_{i=1}^n$ consists of G clusters. (ii) The number of clusters G is fixed, and the size of each cluster $L_g \rightarrow \infty$ with the total sample size $n \rightarrow \infty$

$$\begin{pmatrix} \frac{1}{\sqrt{L_1}} \sum_{i=1}^{L_1} f_i^1(\theta_0) \\ \vdots \\ \frac{1}{\sqrt{L_G}} \sum_{i=1}^{L_G} f_i^G(\theta_0) \end{pmatrix} \xrightarrow{d} N \left(0, \begin{pmatrix} \Omega_1 & & 0 \\ & \ddots & \\ 0 & & \Omega_G \end{pmatrix} \right), \quad (2)$$

where Ω_g is positive definite for $g = 1, \dots, G$. (iii) The cluster sizes are approximately equal, $L_g/\bar{L} \rightarrow 1$ with $\bar{L} = G^{-1} \sum_{g=1}^G L_g$ for every $g = 1, \dots, G$.

Assumption 1-i) implies that the set $\{f_i(\theta_0)\}_{i=1}^n$ can be partitioned into G nonoverlapping clusters, i.e. $\{f_i(\theta_0)\}_{i=1}^n = \cup_{g=1}^G \{f_i^g(\theta_0)\}_{i=1}^{L_g}$, where the notation in $f_i^g(\theta_0)$ indicates the i -th individual in the cluster g which has size L_g . Together with Assumption-i), Assumption 1-ii) assumes that the cluster structure permits the application of joint CLT to the G number of cluster-wise sums. The joint CLT condition is a key device to study a valid inference in our settings where the number of clusters is fixed, but the number of observations per cluster grows with the same rate as the total number of observations increase. It also characterizes the choice of clusters to be asymptotically independent, and the within-cluster dependence is sufficiently weak to apply a suitable cross-sectional central limit theorem. Note that unlike the literature in large- G asymptotics³, our Assumption 1-ii) does not restrict that the clustered data are independent across different clusters.

As one example of the primitive conditions for the joint CLT, supplementary Appendix B.1 considers a spatial mixing setting in Conley (1999), Jenish and Prucha (2009, 2012) and assumes that the number of observations located on the boundaries between two different clusters is dominated by the average number of cluster sizes. By doing so, we show that the different clusters are asymptotically (mean) independent but the units from different clusters are allowed to be weakly dependent.⁴

Assumption 1-iii) states that all of our fixed- G asymptotics are understood to be as the different sizes of clusters L_g to grow to infinity, but at the same rate and relative size, i.e. $L_g/n \rightarrow 1/G$ for $g = 1, \dots, G$. Equivalently, we can express this approximately equal cluster size assumption by $\bar{L} = L_g + o(\bar{L})$ for each $g = 1, \dots, G$. It is important to point out that we do not restrict the size of cluster to be exactly same with each other, i.e. $L_1 = L_2 = \dots = L_G$, but each cluster has approximately same size relative to the average cluster size.

³See for example, Carter et al. (2017), Djogbenou et al. (2019), and Hansen and Lee (2019).

⁴BCH (2011) which is closely related to our clustered structure, considers a spatial setting with group structure and makes an initial contribution toward providing a set of regularity conditions that are sufficient to obtain the fixed- G limiting distributions. However, the asymptotic theory developed in BCH (2011) is only applicable to the exactly linear regression model, and it thus is limited to apply our GMM setting with potentially non-linear moment conditions. Also, BCH (2011) assumes the exactly equal cluster size, whereas we allow the cluster sizes to be unbalanced.

Under Assumption 1, the total (scaled) sum of moment process can be decomposed into cluster-wise sums as follows

$$\sqrt{n}g_n(\theta_0) = \sum_{g=1}^G \sqrt{\frac{L_g}{n}} \frac{1}{\sqrt{L_g}} \sum_{i=1}^{L_g} f_i^g(\theta_0) = \frac{1}{\sqrt{G}} \sum_{g=1}^G \frac{1}{\sqrt{L_g}} \sum_{i=1}^{L_g} f_i^g(\theta_0)(1 + o_p(1)),$$

where the second equation follows by 1-iii).

Let $B_{m,g}$ be an independent $N(0, I_m)$ over $g = 1, \dots, G$. Together with continuous mapping theorem, the joint CLT condition in Assumption 1-ii) further implies

$$\sqrt{n}g_n(\theta_0) \xrightarrow{d} \frac{1}{\sqrt{G}} \sum_{g=1}^G \Lambda_g B_{m,g} \stackrel{d}{=} N(0, \Omega), \quad (3)$$

where $\Omega = \frac{1}{G} \sum_{g=1}^G \Omega_g$ and Λ_g is the matrix square root of Ω_g . Thus, the global covariance matrix of the whole population Ω can be represented as the simple average of Ω_g , $g = 1, \dots, G$, where Ω_g 's are the limiting variances within individual clusters. Motivated by this, we construct the clustered covariance estimator (CCE) as follows:

$$\begin{aligned} \hat{\Omega}(\hat{\theta}_1) &= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n 1(i, j \in \text{the same group}) f_i(\hat{\theta}_1) f_j(\hat{\theta}_1)' \\ &= \frac{1}{G} \sum_{g=1}^G \left\{ \left(\frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} f_i^g(\hat{\theta}_1) \right) \left(\frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} f_i^g(\hat{\theta}_1) \right)' \right\}. \end{aligned}$$

To ensure that $\hat{\Omega}(\hat{\theta}_1)$ is positive definite, we assume that $G \geq m$, and maintain this condition throughout the rest of the paper.

Suppose we want to test the null hypothesis $H_0 : R\theta_0 = r$ against the alternative $H_1 : R\theta_0 \neq r$, where R is a $p \times d$ matrix. In this paper, we focus on linear restrictions without loss of generality because the Delta method can be used to convert nonlinear restrictions into linear ones in an asymptotic sense. The F test version of the Wald test statistic is given by

$$F(\hat{\theta}_1) := \frac{1}{p} (R\hat{\theta}_1 - r)' \left\{ R\widehat{\text{var}}(\hat{\theta}_1)R' \right\}^{-1} (R\hat{\theta}_1 - r), \quad (4)$$

where

$$\widehat{\text{var}}(\hat{\theta}_1) = \frac{1}{n} \left[\hat{\Gamma}(\hat{\theta}_1)' W_n^{-1} \hat{\Gamma}(\hat{\theta}_1) \right]^{-1} \left[\hat{\Gamma}(\hat{\theta}_1)' W_n^{-1} \hat{\Omega}(\hat{\theta}_1) W_n^{-1} \hat{\Gamma}(\hat{\theta}_1) \right] \left[\hat{\Gamma}(\hat{\theta}_1)' W_n^{-1} \hat{\Gamma}(\hat{\theta}_1) \right]^{-1}.$$

When $p = 1$ and the alternative is one sided, we can construct the t statistic $t(\hat{\theta}_1) = (R\hat{\theta}_1 - r) / (R\widehat{\text{var}}(\hat{\theta}_1)R')^{1/2}$. To formally characterize the asymptotic distributions of $F(\hat{\theta}_1)$ and $t(\hat{\theta}_1)$ under the fixed- G asymptotics, we further maintain the following high level conditions.

Assumption 2 $\hat{\theta}_1 \xrightarrow{p} \theta_0$.

Assumption 3 (i) For each $g = 1, \dots, G$, let $\Gamma_g(\theta) := \lim_{L_g \rightarrow \infty} E \left[\frac{1}{L_g} \sum_{i=1}^{L_g} \frac{\partial f_i^g(\theta)}{\partial \theta'} \right]$. Then, there exists $\epsilon > 0$ such that

$$\sup_{\theta \in B(\theta_0, \epsilon)} \left\| \frac{1}{L_g} \sum_{i=1}^{L_g} \frac{\partial f_i^g(\theta)}{\partial \theta'} - \Gamma_g(\theta) \right\| \xrightarrow{p} 0 \text{ holds,}$$

where $B(\theta_0, \epsilon)$ is an open ball of θ_0 of radius ϵ , and $\|\cdot\|$ is the Euclidean norm. (ii) $\Gamma_g(\theta)$ is continuous at $\theta = \theta_0$, and for $\Gamma_g = \Gamma_g(\theta_0)$, $\Gamma = G^{-1} \sum_{g=1}^G \Gamma_g$ has full rank.

Assumption 4 (Homogeneity of Γ_g) For all $g = 1, \dots, G$, $\Gamma_g = \Gamma$.

Assumption 5 (Homogeneity of Ω_g) For all $g = 1, \dots, G$, $\Omega_g = \Omega$.

Assumption 2 is made for convenience, and primitive sufficient conditions are available from the standard GMM asymptotic theory. Assumption 3 is a uniform law of large numbers (ULLN), from which we obtain $\hat{\Gamma}(\hat{\theta}_1) = G^{-1} \sum_{g=1}^G \Gamma_g + o_p(1) = \Gamma + o_p(1)$. Supplementary Appendix B.1 presents the formal assumptions that are sufficient for the spatial ULLN condition in Assumption 3, and proves that each set of assumptions is indeed sufficient.

The homogeneity conditions in Assumptions 4 and 5 guarantee the asymptotic pivotality of the cluster-robust GMM statistics we consider, see Remark 4 below. Similar assumptions are made in BCH (2011) and Sun and Kim (2015), which develop asymptotically valid F tests that are robust to spatial autocorrelation in the same spirit as our fixed- G asymptotics. The homogeneity conditions can be restrictive depending on empirical contexts of the clustered data. However, one of the advantages to impose them in our approach is that one can perform asymptotic t and F inference in a broad class of GMM test statistics such as Wald/LM/LR-type test statistics. Also, our approach allows joint testing for parameter vector as well as J -statistics for testing over-identification. This contrasts with alternative fixed-cluster robust inferences studied in Ibragimov and Müller (2010, 2016) and Canay et al. (2017) and Canay et al. (2019). They relax the homogeneity conditions in Assumptions 4 and 5, but their main focus is testing a single hypothesis in the exactly identified regression problem.

Let

$$\bar{B}_m := G^{-1} \sum_{g=1}^G B_{m,g} \text{ and } \bar{\mathbb{S}} := \frac{1}{G} \sum_{g=1}^G (B_{m,g} - \bar{B}_m) (B_{m,g} - \bar{B}_m)',$$

where $B_{m,g}$ as in (B.9). Also, let $\mathbb{W}_p(K, \Pi)$ denote a Wishart distribution with K degrees of freedom and $p \times p$ positive definite scale matrix Π . By construction, $\sqrt{G} \bar{B}_m \sim N(0, I_m)$, $\bar{\mathbb{S}} \sim G^{-1} \mathbb{W}_p(G-1, I_m)$, and $\bar{B}_m \perp \bar{\mathbb{S}}$. To present our asymptotic results, we partition \bar{B}_m and $\bar{\mathbb{S}}$ as follows:

$$\bar{B}_m = \begin{pmatrix} \bar{B}_d \\ d \times 1 \\ \bar{B}_q \\ q \times 1 \end{pmatrix}, \bar{B}_d = \begin{pmatrix} \bar{B}_p \\ p \times 1 \\ \bar{B}_{d-p} \\ (d-p) \times 1 \end{pmatrix}, \bar{\mathbb{S}} = \begin{pmatrix} \bar{\mathbb{S}}_{dd} & \bar{\mathbb{S}}_{dq} \\ d \times d & d \times q \\ \bar{\mathbb{S}}_{qd} & \bar{\mathbb{S}}_{qq} \\ q \times d & q \times q \end{pmatrix},$$

and similarly define $\bar{\mathbb{S}}_{pp} \in \mathbb{R}^{p \times p}$ and $\bar{\mathbb{S}}_{pq} \in \mathbb{R}^{p \times q}$ as submatrices of $\bar{\mathbb{S}}_{dd}$ and $\bar{\mathbb{S}}_{dq}$, respectively.

Proposition 1 Let Assumptions 1~5 hold. Then,

- (a) $F(\hat{\theta}_1) \xrightarrow{d} \mathbb{F}_{1\infty} := \begin{pmatrix} G \\ p \end{pmatrix} \cdot \bar{B}_p' \bar{\mathbb{S}}_{pp}^{-1} \bar{B}_p$;
- (b) $t(\hat{\theta}_1) \xrightarrow{d} \mathbb{T}_{1\infty} := \frac{N(0,1)}{\sqrt{\chi_{G-1}^2/G}}$ where $N(0,1) \perp \chi_{G-1}^2$.

Let Λ the matrix square root of Ω , that is, $\Lambda\Lambda' = \Omega$. The proof of Proposition 1 shows that $\hat{\Omega}(\hat{\theta}_1)$ converges in distribution to a random matrix $\Omega_{1\infty}$ given by

$$\begin{aligned}\Omega_{1\infty} &= \Lambda \tilde{\mathbb{D}} \Lambda' \text{ where } \tilde{\mathbb{D}} = \frac{1}{G} \sum_{g=1}^G \tilde{D}_g \tilde{D}_g' \text{ and} \\ \tilde{D}_g &= B_{m,g} - \underbrace{\Gamma_\Lambda (\Gamma_\Lambda' W_\Lambda^{-1} \Gamma_\Lambda)^{-1} \Gamma_\Lambda' W_\Lambda^{-1} \bar{B}_m}_{\text{Quasi-demeaning factor}}\end{aligned}\quad (5)$$

for $\Gamma_\Lambda = \Lambda^{-1}\Gamma$ and $W_\Lambda = \Lambda^{-1}W(\Lambda')^{-1}$. \tilde{D}_g is a quasi-demeaned version of $B_{m,g}$ with quasi-demeaning attributable to the estimation error in $\hat{\theta}_1$. Note that the quasi-demeaning factor in (5) depends on all of Γ, Ω and W , and cannot be further simplified in general. The estimation error in $\hat{\theta}_1$ affects $\Omega_{1\infty}$ in a complicated way. However, for the first-step Wald and t statistics, we do not care about $\hat{\Omega}(\hat{\theta}_1)$ *per se*. Instead, we care about the scaled covariance matrix $\hat{\Gamma}(\hat{\theta}_1)' W_n^{-1} \hat{\Omega}(\hat{\theta}_1) W_n^{-1} \hat{\Gamma}(\hat{\theta}_1)$, which converges in distribution to $\Gamma' W^{-1} \Omega_{1\infty} W^{-1} \Gamma$. But

$$\Gamma_\Lambda' W_\Lambda^{-1} \tilde{D}_g = \Gamma_\Lambda' W_\Lambda^{-1} (B_{m,g} - \bar{B}_m),$$

and thus

$$\begin{aligned}\Gamma' W^{-1} \Omega_{1\infty} W^{-1} \Gamma &= \Gamma_\Lambda' W_\Lambda^{-1} \tilde{\mathbb{D}} W_\Lambda^{-1} \Gamma_\Lambda = \frac{1}{G} \sum_{g=1}^G \Gamma_\Lambda' W_\Lambda^{-1} \tilde{D}_g (\Gamma_\Lambda' W_\Lambda^{-1} \tilde{D}_g)' \\ &\stackrel{d}{=} \Gamma_\Lambda' W_\Lambda^{-1} \frac{1}{G} \sum_{g=1}^G (B_{m,g} - \bar{B}_m) (B_{m,g} - \bar{B}_m)' (\Gamma_\Lambda' W_\Lambda^{-1})'.\end{aligned}\quad (6)$$

Therefore, to the first order fixed- G asymptotics, the estimation error in $\hat{\theta}_1$ affects $\Gamma' W^{-1} \Omega_{1\infty} W^{-1} \Gamma$ via simple demeaning only. This is a key result that drives the asymptotic pivotality of $F(\hat{\theta}_1)$ and $t(\hat{\theta}_1)$.

As an example of the general GMM setting, consider the linear regression model $y_i = x_i' \theta + \epsilon_i$. Under the assumption that $E[x_i \epsilon_i] = 0$, the moment function is $f_i(\theta) = x_i(y_i - x_i' \theta)$. With the moment condition $E f_i(\theta_0) = 0$, the model is exactly identified. This set up was employed in Hansen (2007), Stock and Watson (2008), and BCH (2011); indeed, our F and t approximations in Proposition 1 are identical to what is obtained in these papers.

Remark 2 *The sufficient condition for the within-cluster CLT condition in Assumption 1-ii) is a weak cross-sectional dependence within and across clusters, but we can allow for some type of strong dependence induced by a potentially latent common shock \mathcal{C} as in Andrews (2005).⁵ The modified conditional moment condition is*

$$E[f(Y_i, \theta_0) | \mathcal{C}] = 0,$$

which implies that $f(Y_i, \theta_0)$ is conditionally mean independent of the common shock \mathcal{C} . Correspondingly, the CLT condition in Assumption 1-ii) is to be hold conditional on \mathcal{C} with a (unconditional) mixed normal limit, and the probability limits. Also, Γ_g and Ω_g , in Assumptions 3, 4 and 5 are random processes which are conditionally independent across g . Then, it is straightforward

⁵We thank an anonymous referee for pointing this out.

to check that the conditional limits $\mathbb{F}_{1\infty}$ and $\mathbb{T}_{1\infty}$ do not depend on the conditioning common shock \mathcal{C} , and thus the fixed- G asymptotics in Proposition 1 are still asymptotically valid. The conditioning arguments used here are entirely analogous to that used in Theorem 6 in Andrews (2005) and shall be maintained in the rest of this paper.

Remark 3 The limiting distribution $\mathbb{F}_{1\infty}$ follows Hotelling's T^2 distribution. Using the well-known relationship between the T^2 and standard F distributions, we obtain $\mathbb{F}_{1\infty} \stackrel{d}{=} G/(G-p) \mathcal{F}_{p, G-p}$ where $\mathcal{F}_{p, G-p}$ is a random variable that follows the F distribution with degree of freedom $(p, G-p)$. Similarly, $\mathbb{T}_{1\infty} \stackrel{d}{=} \sqrt{G/(G-1)} t_{G-1}$ where t_{G-1} is a random variable that follows the t distribution with degree of freedom $G-1$.

Remark 4 When the cluster homogeneity conditions in Assumptions 4 and 5 are relaxed, one can re-investigate the proof of Proposition 1 and check that

$$\begin{aligned} & \hat{\Gamma}(\hat{\theta}_1)' W_n^{-1} \frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} f_i^g(\hat{\theta}_1) \xrightarrow{d} \Gamma' W^{-1} [\Lambda_g B_{m,g} - \Gamma_g (\Gamma' W^{-1} \Gamma)^{-1} \Gamma' W^{-1} \Lambda \bar{B}_m] \\ & = \Gamma' W \Lambda_g B_{m,g} - \Gamma' W^{-1} \Gamma_g (\Gamma' W^{-1} \Gamma)^{-1} \Gamma' W^{-1} \Lambda \bar{B}_m. \end{aligned}$$

Without imposing $\Lambda_g = \Lambda$ and $\Gamma_g = \Gamma$, we cannot proceed with the above expressions and further show that the CCE has the fixed- G Wishart limiting distributions in (6). In Section 6, we investigate how violation of these assumptions impacts the finite sample performance of the cluster-robust tests and compare our results to those of Ibragimov and Müller (2010, 2016) which do not require the homogeneity conditions. See also Remark 5 below where we justify our tests in the context of large- G asymptotics which does not require the cluster homogeneity.

Remark 5 Under the large- G asymptotics where $G \rightarrow \infty$ but L_g is fixed, one can show that the CCE $\hat{\Omega}(\hat{\theta}_1)$ converges in probability to Ω for

$$\Omega = \lim_{G \rightarrow \infty} \frac{1}{G} \sum_{g=1}^G \text{var} \left(\frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} f_i^g(\theta_0) \right).$$

The convergence of $\hat{\Omega}(\hat{\theta}_1)$ to Ω does not require the homogeneity of Ω_g in Assumption 5 (Hansen, 2007; Carter et al., 2017). Under this type of asymptotics, the test statistics $F(\hat{\theta}_1)$ and $t(\hat{\theta}_1)$ are asymptotically χ_p^2/p and $N(0, 1)$. Let $\mathcal{F}_{p, G-p}^{1-\alpha}$ and $\chi_p^{1-\alpha}$ be the $1-\alpha$ quantiles of the $\mathcal{F}_{p, G-p}$ and χ_p^2 distributions, respectively. As $G/(G-p) > 1$ and $\mathcal{F}_{p, G-p}^{1-\alpha} > \chi_p^{1-\alpha}/p$, it is easy to see that

$$\frac{G}{G-p} \mathcal{F}_{p, G-p}^{1-\alpha} > \chi_p^{1-\alpha}/p.$$

However, the difference between the two critical values $G(G-p)^{-1} \mathcal{F}_{p, G-p}^{1-\alpha}$ and $\chi_p^{1-\alpha}/p$ shrinks to zero as G increases. Therefore, the fixed- G critical value $G/(G-p)^{-1} \mathcal{F}_{p, G-p}^{1-\alpha}$ is asymptotically valid under the large- G asymptotics. This asymptotic validity holds even if the homogeneity conditions of Assumptions 4 and 5 are not satisfied. The fixed- G critical value is robust in the sense that it works whether G is small or large.

3 Two-step GMM Estimation and Inference

In an overidentified GMM framework, we often employ a two-step procedure to improve the efficiency of the initial GMM estimator and the power of the associated tests. It is now well-known that the optimal weighting matrix is the (inverted) asymptotic variance of the sample moment conditions, see Hansen (1982). There are at least two different ways to estimate the asymptotic variance, and these lead to two different estimators $\hat{\Omega}(\hat{\theta}_1)$ and $\hat{\Omega}^c(\hat{\theta}_1)$, where

$$\hat{\Omega}(\hat{\theta}_1) = \frac{1}{G} \sum_{g=1}^G \left(\frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} f_i^g(\hat{\theta}_1) \right) \left(\frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} f_i^g(\hat{\theta}_1) \right)', \quad (7)$$

and

$$\hat{\Omega}^c(\hat{\theta}_1) = \frac{1}{G} \sum_{g=1}^G \left\{ \frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} [f_i^g(\hat{\theta}_1) - g_n(\hat{\theta}_1)] \right\} \left\{ \frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} [f_i^g(\hat{\theta}_1) - g_n(\hat{\theta}_1)] \right\}'. \quad (8)$$

While $\hat{\Omega}(\hat{\theta}_1)$ employs the uncentered moment process $\cup_{g=1}^G \cup_{i=1}^{L_g} \{f_i^g(\hat{\theta}_1)\}$, $\hat{\Omega}^c(\hat{\theta}_1)$ employs the recentered moment process $\cup_{g=1}^G \cup_{i=1}^{L_g} \{f_i^g(\hat{\theta}_1) - g_n(\hat{\theta}_1)\}$. For inference based on the first-step estimator $\hat{\theta}_1$, it does not matter which asymptotic variance estimator is used. This is so because for any asymptotic variance estimator $\hat{\Omega}(\hat{\theta}_1)$, the Wald statistic depends on $\hat{\Omega}(\hat{\theta}_1)$ only via $\hat{\Gamma}(\hat{\theta}_1)' W_n^{-1} \hat{\Omega}(\hat{\theta}_1) W_n^{-1} \hat{\Gamma}(\hat{\theta}_1)$. It is easy to show that the following asymptotic equivalence:

$$\begin{aligned} \hat{\Gamma}(\hat{\theta}_1)' W_n^{-1} \hat{\Omega}(\hat{\theta}_1) W_n^{-1} \hat{\Gamma}(\hat{\theta}_1) &= \hat{\Gamma}(\hat{\theta}_1)' W_n^{-1} \hat{\Omega}^c(\hat{\theta}_1) W_n^{-1} \hat{\Gamma}(\hat{\theta}_1) + o_p(1) \\ &= \Gamma' W_n^{-1} \hat{\Omega}^c(\theta_0) W_n^{-1} \Gamma + o_p(1). \end{aligned}$$

Thus, the limiting distribution of the Wald statistic is the same whether the estimated moment process is recentered or not. In the next subsections, however, we show that the two asymptotic variance estimators in (7) and (8) are not asymptotically equivalent by themselves under the fixed- G asymptotics.

Depending on whether we use $\hat{\Omega}(\hat{\theta}_1)$ or $\hat{\Omega}^c(\hat{\theta}_1)$, we have different two-step GMM estimators:

$$\hat{\theta}_2 = \arg \min_{\theta \in \Theta} g_n(\theta)' \left[\hat{\Omega}(\hat{\theta}_1) \right]^{-1} g_n(\theta) \quad \text{and} \quad \hat{\theta}_2^c = \arg \min_{\theta \in \Theta} g_n(\theta)' \left[\hat{\Omega}^c(\hat{\theta}_1) \right]^{-1} g_n(\theta).$$

Given that $\hat{\Omega}(\hat{\theta}_1)$ and $\hat{\Omega}^c(\hat{\theta}_1)$ are not asymptotically equivalent and that they enter the definitions of $\hat{\theta}_2$ and $\hat{\theta}_2^c$ by themselves, the two estimators have different asymptotic behaviors, as proved in the next two subsections.

3.1 Uncentered Two-step GMM estimator

In this subsection, we consider the two-step GMM estimator $\hat{\theta}_2$ based on the uncentered moment process. We investigate the asymptotic properties of $\hat{\theta}_2$ and establish fixed- G limits of the associated Wald statistic and J statistic. We show that the J statistic is asymptotically pivotal, even though the Wald statistic is not.

It follows from standard asymptotic arguments and Assumption 1-iii) that

$$\sqrt{n}(\hat{\theta}_2 - \theta_0) = - \left[\Gamma' \hat{\Omega}^{-1}(\hat{\theta}_1) \Gamma \right]^{-1} \Gamma' \hat{\Omega}^{-1}(\hat{\theta}_1) \frac{1}{\sqrt{G}} \sum_{g=1}^G \left(\frac{1}{\sqrt{L_g}} \sum_{i=1}^{L_g} f_i^g(\theta_0) \right) (1 + o_p(1)). \quad (9)$$

Using the joint convergence of the followings

$$\hat{\Omega}(\hat{\theta}_1) \xrightarrow{d} \Omega_{1\infty} = \Lambda \tilde{\mathbb{D}} \Lambda' \text{ and } \frac{1}{\sqrt{G}} \sum_{g=1}^G \left(\frac{1}{\sqrt{L_g}} \sum_{i=1}^{L_g} f_i^g(\theta_0) \right) \xrightarrow{d} \sqrt{G} \Lambda \bar{B}_m, \quad (10)$$

we obtain

$$\sqrt{n}(\hat{\theta}_2 - \theta_0) \xrightarrow{d} - \left[\Gamma'_\Lambda \tilde{\mathbb{D}}^{-1} \Gamma_\Lambda \right]^{-1} \Gamma'_\Lambda \tilde{\mathbb{D}}^{-1} \sqrt{G} \bar{B}_m,$$

where $\tilde{\mathbb{D}} = G^{-1} \sum_{i=1}^G \tilde{D}_g \tilde{D}'_g$ is defined in (5).

Since $\tilde{\mathbb{D}}$ is random, the limiting distribution is not normal. Even though both \tilde{D}_g and \bar{B}_m are normal, there is a nonzero correlation between them. As a result, $\tilde{\mathbb{D}}$ and \bar{B}_m are correlated, too. This makes the limiting distribution of $\sqrt{n}(\hat{\theta}_2 - \theta_0)$ highly nonstandard. Define the F statistic and variance estimate for the two-step estimator $\hat{\theta}_2$ as

$$F_{\hat{\Omega}(\hat{\theta}_1)}(\hat{\theta}_2) = \frac{1}{p} (R\hat{\theta}_2 - r)' \left(R \widehat{\text{var}}_{\hat{\Omega}(\hat{\theta}_1)}(\hat{\theta}_2) R' \right)^{-1} (R\hat{\theta}_2 - r);$$

$$\widehat{\text{var}}_{\hat{\Omega}(\hat{\theta}_1)}(\hat{\theta}_2) = \frac{1}{n} \left(\hat{\Gamma}(\hat{\theta}_2)' \hat{\Omega}^{-1}(\hat{\theta}_1) \hat{\Gamma}(\hat{\theta}_2) \right)^{-1}.$$

On the basis of $\hat{\theta}_2$, the J statistic for testing over-identification restrictions is

$$J(\hat{\theta}_2) := n g_n(\hat{\theta}_2)' \left(\hat{\Omega}(\hat{\theta}_1) \right)^{-1} g_n(\hat{\theta}_2). \quad (11)$$

In the above definitions, we use a subscript notation $\hat{\Omega}(\hat{\theta}_1)$ to clarify the choice of CCE in F and J statistics. Now the question is, is the above F statistic asymptotically pivotal? How about the J statistics? To answer this, U to be the $m \times m$ matrix of the eigen vectors of $\Gamma'_\Lambda \Gamma_\Lambda = \Gamma' \Omega^{-1} \Gamma$ and $U \Sigma V'$ be a singular value decomposition (SVD) of Γ_Λ . By construction, $U'U = UU' = I_m$, $V'V = VV' = I_d$, and $\Sigma' = \begin{bmatrix} A_{d \times d} & O_{d \times q} \end{bmatrix}$. We then define $\tilde{W} = U'W_\Lambda U$ and partition \tilde{W} as before to define $\beta_{\tilde{W}} = \tilde{W}_{dq} \tilde{W}_{qq}^{-1}$ which is the regression coefficient induced by the constant matrix \tilde{W} . We use the following additional notation:

$$\mathbb{E}_{p+q,p+q} := \begin{pmatrix} \mathbb{E}_{pp} & \mathbb{E}_{pq} \\ \mathbb{E}'_{pq} & \mathbb{E}_{qq} \end{pmatrix} = \begin{pmatrix} \bar{\mathbb{S}}_{pp} & \bar{\mathbb{S}}_{pq} \\ \bar{\mathbb{S}}'_{pq} & \bar{\mathbb{S}}_{qq} \end{pmatrix} + \begin{pmatrix} \tilde{\beta}_{\tilde{W}}^p \bar{B}_q \bar{B}'_q (\tilde{\beta}_{\tilde{W}}^p)' & \tilde{\beta}_{\tilde{W}}^p \bar{B}_q \bar{B}'_q \\ \bar{B}_q \bar{B}'_q (\tilde{\beta}_{\tilde{W}}^p)' & \bar{B}_q \bar{B}'_q \end{pmatrix}, \quad (12)$$

where $\tilde{\beta}_{\tilde{W}}^p$ is the $p \times q$ matrix and consists of the first p rows of $\tilde{V}' \beta_{\tilde{W}}$ where \tilde{V} is the $d \times d$ matrix of the eigen vector of $(RVA^{-1})' RVA^{-1}$. By construction, the matrix $\tilde{\beta}_{\tilde{W}}^p$ depends on R, Γ, W , and Ω .

Proposition 6 *Let Assumptions 1~5 hold. Then*

- (a) $F_{\hat{\Omega}(\hat{\theta}_1)}(\hat{\theta}_2) \xrightarrow{d} \frac{G}{p} \cdot [\bar{B}'_{p+q} \mathbb{E}_{p+q,p+q}^{-1} \bar{B}_{p+q} - \bar{B}'_q \mathbb{S}_{qq}^{-1} \bar{B}_q]$;
- (b) $J(\hat{\theta}_2) \xrightarrow{d} G \cdot \bar{B}'_q \mathbb{S}_{qq}^{-1} \bar{B}_q$, where the convergences in (a) and (b) holds jointly.

The proofs of Proposition 6 are available from supplementary Appendix 6. Due to the presence of the term of $\mathbb{E}_{p+q,p+q}$, the result of Proposition-(a) indicates that the F statistic is not asymptotically pivotal, and it depends on several nuisance parameters including Ω . This is because $\mathbb{E}_{p+q,p+q}$ is a nonconstant function of $\tilde{\beta}_{\tilde{W}}^p$, which, in turn, depends on R, Γ, W and Ω .

While the asymptotic distributions of $F_{\hat{\Omega}(\hat{\theta}_1)}(\hat{\theta}_2)$ is complicated and nonstandard, it is perhaps surprising that the limiting distribution of the J-statistic is free of nuisance parameters. Using the Sherman–Morrison formula⁶, we check that the weak limit of $J(\hat{\theta}_2)$ satisfies

$$G \cdot \bar{B}'_q \bar{S}_{qq}^{-1} \bar{B}_q = G \cdot \frac{\bar{B}'_q \bar{S}_{qq}^{-1} \bar{B}_q}{1 + \bar{B}'_q \bar{S}_{qq}^{-1} \bar{B}_q} \stackrel{d}{=} G \cdot \frac{q \mathcal{F}_{q, G-q}}{(G-q) + q \mathcal{F}_{q, G-q}},$$

which implies that we have we have

$$\tilde{J}(\hat{\theta}_2) := \frac{G-q}{q} \frac{q J(\hat{\theta}_2)}{G-q J(\hat{\theta}_2)} \stackrel{d}{\rightarrow} \mathcal{F}_{q, G-q}.$$

That is, the transformed J-statistic $\tilde{J}(\hat{\theta}_2)$ is asymptotically F distributed, which is a very appealing result for empirical applications. It is important to point out that the convenient F limit of $J(\hat{\theta}_2)$ holds only if the J-statistic is equal to the GMM criterion function evaluated at the two-step GMM estimator $\hat{\theta}_2$. This effectively imposes a constraint on the weighting matrix. If we use a weighting matrix that is different from $\hat{\Omega}(\hat{\theta}_1)$, then the resulting J-statistic may not be asymptotically pivotal any longer.

3.2 Centered Two-step GMM estimator

Given that the estimation error in $\hat{\theta}_1$ affects the limiting distribution of $\hat{\Omega}(\hat{\theta}_1)$, the Wald statistic based on the uncentered two-step GMM estimator $\hat{\theta}_2$ is not asymptotically pivotal. In view of (5), the effect of the estimation error is manifested via a location shift in \tilde{D}_g ; the shifting amount depends on $\hat{\theta}_1$. A key observation is that the location shift is the same for all groups under the homogeneity Assumptions 4 and 5. Therefore, if we demean the empirical moment process, we can remove the location shift that is caused by the estimator error in $\hat{\theta}_1$. This leads to the recentered asymptotic variance estimator and a pivotal inference for both the Wald test and J test.

It is important to note that the recentering is not innocuous for an over-identified GMM model because $n^{-1} \sum_{i=1}^n f_i(\hat{\theta}_1)$ is not zero in general. In the time series HAR variance estimation, the recentering is known to have several advantages. For example, as Hall (2000) observes, in the conventional increasing smoothing asymptotic theory, the recentering can potentially improve the power of the J test using a HAR variance estimator when the model is misspecified.

In our fixed- G asymptotic framework, the recentering plays an important role in the CCE estimation. It ensures that the limiting distribution of $\hat{\Omega}^c(\hat{\theta}_1)$ is invariant to the initial estimator $\hat{\theta}_1$. The following lemma proves a more general result and characterizes the fixed- G limiting distribution of the centered CCE matrix for any \sqrt{n} -consistent estimator $\tilde{\theta}$.

Lemma 7 *Let Assumptions 1~5 hold. Let $\tilde{\theta}$ be any \sqrt{n} -consistent estimator of θ_0 . Then*

- (a) $\hat{\Omega}^c(\tilde{\theta}) = \hat{\Omega}^c(\theta_0) + o_p(1)$;
- (b) $\hat{\Omega}^c(\theta_0) \stackrel{d}{\rightarrow} \Omega_\infty^c$ where $\Omega_\infty^c = \Lambda \bar{S} \Lambda'$.

Lemma 7 indicates that the centered CCE $\Omega^c(\hat{\theta}_1)$ converges in distribution to the random matrix limit $\Omega_\infty^c = \Lambda \bar{S} \Lambda'$, which follows a (scaled) Wishart distribution $G^{-1} \mathbb{W}_m(G-1, \Omega)$. Using

⁶ $(C + ab')^{-1} = C^{-1} - \frac{C^{-1} ab' C^{-1}}{1 + b' C^{-1} a}$ for any invertable square matrix C and conforming column vectors such that $1 + b' C^{-1} a \neq 0$.

Lemma 7, it is possible to show

$$\sqrt{n}(\hat{\theta}_2^c - \theta_0) = - \left(\Gamma' \left[\hat{\Omega}^c(\hat{\theta}_1) \right]^{-1} \Gamma \right)^{-1} \Gamma' \left[\hat{\Omega}^c(\hat{\theta}_1) \right]^{-1} \sqrt{n}g_n(\theta_0) + o_p(1) \quad (13)$$

$$\xrightarrow{d} - \left[\Gamma' (\Omega_\infty^c)^{-1} \Gamma \right]^{-1} \Gamma' (\Omega_\infty^c)^{-1} \Lambda \sqrt{G} \bar{B}_m. \quad (14)$$

Since $(\Omega_\infty^c)^{-1}$ is independent with $\sqrt{G}\Lambda\bar{B}_m \sim N(0, \Omega)$, the limiting distribution of $\hat{\theta}_2^c$ is mixed normal.

On the basis of $\hat{\theta}_2^c$, we can construct the following normalized Wald statistic

$$F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) := \frac{1}{p} (R\hat{\theta}_2^c - r)' \{ R \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) R' \}^{-1} (R\hat{\theta}_2^c - r), \text{ where} \quad (15)$$

$$\widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) = \frac{1}{n} \left(\hat{\Gamma}(\hat{\theta}_2^c)' \left(\hat{\Omega}^c(\hat{\theta}_1) \right)^{-1} \hat{\Gamma}(\hat{\theta}_2^c) \right)^{-1}.$$

When $p = 1$, corresponding t statistic $t_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)$ can be constructed similarly. We also construct the standard J statistic based on $\hat{\theta}_2^c$

$$J(\hat{\theta}_2^c) := ng_n(\hat{\theta}_2^c)' \left(\hat{\Omega}^c(\hat{\theta}_1) \right)^{-1} g_n(\hat{\theta}_2^c),$$

where $\hat{\Omega}^c(\hat{\theta}_1)$ can be replaced by $\hat{\Omega}^c(\hat{\theta}_2^c)$ without affecting the limiting distribution of the J statistic.

Using (13) and Lemma 7, we have $F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) \xrightarrow{d} \mathbb{F}_{2\infty}$ where

$$\mathbb{F}_{2\infty} = \frac{G}{p} \cdot \left[R (\Gamma'_\Lambda \bar{\mathbb{S}}^{-1} \Gamma_\Lambda)^{-1} \Gamma'_\Lambda \bar{\mathbb{S}}^{-1} \bar{B}_m \right]' \left[R (\Gamma'_\Lambda \bar{\mathbb{S}}^{-1} \Gamma_\Lambda)^{-1} R' \right]^{-1} \quad (16)$$

$$\times \left[R (\Gamma'_\Lambda \bar{\mathbb{S}}^{-1} \Gamma_\Lambda)^{-1} \Gamma'_\Lambda \bar{\mathbb{S}}^{-1} \bar{B}_m \right].$$

Also, it follows in a similar way that

$$J(\hat{\theta}_2^c) \xrightarrow{d} \mathbb{J}_\infty := G \cdot \left\{ \bar{B}_m - \Gamma_\Lambda (\Gamma'_\Lambda \bar{\mathbb{S}}^{-1} \Gamma_\Lambda)^{-1} \Gamma'_\Lambda \bar{\mathbb{S}}^{-1} \bar{B}_m \right\}' \bar{\mathbb{S}}^{-1} \quad (17)$$

$$\times \left\{ \bar{B}_m - \Gamma_\Lambda (\Gamma'_\Lambda \bar{\mathbb{S}}^{-1} \Gamma_\Lambda)^{-1} \Gamma'_\Lambda \bar{\mathbb{S}}^{-1} \bar{B}_m \right\}.$$

The remaining question is whether the above representations for $\mathbb{F}_{2\infty}$ and \mathbb{J}_∞ are free of nuisance parameters. The following proposition provides a positive answer. The proofs of Proposition 8 are available from supplementary Appendix 8.

Proposition 8 *Let Assumptions 1~5 hold and define $\bar{\mathbb{S}}_{pp\cdot q} = \bar{\mathbb{S}}_{pp} - \bar{\mathbb{S}}_{pq} \bar{\mathbb{S}}_{qq}^{-1} \bar{\mathbb{S}}_{qp}$. Then,*

- (a) $F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) \xrightarrow{d} \frac{G}{p} \cdot (\bar{B}_p - \bar{\mathbb{S}}_{pq} \bar{\mathbb{S}}_{qq}^{-1} \bar{B}_q)' \bar{\mathbb{S}}_{pp\cdot q}^{-1} (\bar{B}_p - \bar{\mathbb{S}}_{pq} \bar{\mathbb{S}}_{qq}^{-1} \bar{B}_q)' \stackrel{d}{=} \mathbb{F}_{2\infty}$;
- (b) $t_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) \xrightarrow{d} \sqrt{G} (\bar{B}_p - \bar{\mathbb{S}}_{pq} \bar{\mathbb{S}}_{qq}^{-1} \bar{B}_q) / \sqrt{\bar{\mathbb{S}}_{pp\cdot q}} \stackrel{d}{=} \mathbb{T}_{2\infty}$ for $p = 1$;
- (c) $J(\hat{\theta}_2^c) \xrightarrow{d} G \cdot \bar{B}'_q \bar{\mathbb{S}}_{qq}^{-1} \bar{B}_q \stackrel{d}{=} \mathbb{J}_\infty$, where the convergences hold jointly.

To simplify the representations of $\mathbb{F}_{2\infty}$ and $\mathbb{T}_{2\infty}$ in the above proposition, we note that

$$G \begin{bmatrix} \bar{\mathbb{S}}_{pp} & \bar{\mathbb{S}}_{pq} \\ \bar{\mathbb{S}}_{qp} & \bar{\mathbb{S}}_{qq} \end{bmatrix} \stackrel{d}{=} \sum_{g=1}^G (B_{p+q,g} - \bar{B}_{p+q}) (B_{p+q,g} - \bar{B}_{p+q})',$$

where $B_{p+q,g} := (B'_{p,g}, B'_{q,g})'$. The above random matrix has a standard Wishart distribution $\mathbb{W}_{p+q}(G-1, I_{p+q})$. It follows from the well-known properties of a Wishart distribution that $\bar{\mathbb{S}}_{pp \cdot q} \sim \mathbb{W}_p(G-1-q, I_p)/G$ and $\bar{\mathbb{S}}_{pp \cdot q}$ is independent of $\bar{\mathbb{S}}_{pq}$ and $\bar{\mathbb{S}}_{qq}$. See Bilodeau and Brenner (2008, Proposition 7.9). Therefore, if we condition on $\Delta := \bar{\mathbb{S}}_{pq} \bar{\mathbb{S}}_{qq}^{-1} \sqrt{G} \bar{B}_q$, the limiting distribution $\mathbb{F}_{2\infty}$ satisfies

$$\frac{G-p-q}{G} \mathbb{F}_{2\infty} \stackrel{d}{=} \frac{G-p-q}{G} \frac{(\sqrt{G} \bar{B}_p + \Delta)' \bar{\mathbb{S}}_{pp \cdot q}^{-1} (\sqrt{G} \bar{B}_p + \Delta)}{p} \stackrel{d}{=} \mathcal{F}_{p, G-p-q}(\|\Delta\|^2),$$

where $\mathcal{F}_{p, G-p-q}(\|\Delta\|^2)$ is a noncentral F distribution with random noncentrality parameter $\|\Delta\|^2$. Similarly, the limiting distribution of (scaled) $\mathbb{T}_{2\infty}$ can be represented as $t_{G-1-q}(\Delta)$ which is a noncentral t distribution with a noncentrality parameter $t_{p, G-1-q}(\Delta)$ which is a noncentral t distribution with a noncentrality parameter Δ . These nonstandard limiting distributions are similar to those in Sun (2014) which provides the fixed-smoothing asymptotic result in the case of the series LRV estimation. However, in our setting of clustered dependence, the scale adjustment and degrees of freedom parameter in $\mathcal{F}_{p, G-p-q}(\|\Delta\|^2)$ and $t_{p, G-1-q}(\Delta)$ are different from those in Sun (2014). The critical values from the nonstandard limiting distribution $\mathbb{T}_{2\infty}$ and $\mathbb{F}_{2\infty}$ can be obtained through simulation.

For the J statistic $J(\hat{\theta}_2^c)$, it follows from Proposition 8-(c) that

$$\left(\frac{G-q}{Gq} \right) \cdot J(\hat{\theta}_2^c) \stackrel{d}{\rightarrow} \left(\frac{G-q}{Gq} \right) \cdot \bar{B}'_q \bar{\mathbb{S}}_{qq}^{-1} \bar{B}_q \stackrel{d}{=} \mathcal{F}_{q, G-q}.$$

This is consistent with Sun and Kim's (2012) results except that our adjustment and degrees of freedom parameter are different.

Remark 9 *In supplemental Appendix B.4, we extend the formulation of Wald tests to the QLR and LM types of GMM statistics and show that they are asymptotically equivalent to $F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)$. This implies that all three types of test statistics share the same fixed- G limit as given in Proposition 8. Similar results are obtained by Hwang and Sun (2017a), which focus on the two-step GMM estimation and HAR inference in a time series setting.*

4 Asymptotic F and t Tests for Centered Two-step GMM Procedures

The limiting distributions of the centered two-step GMM test statistics in Section 3 are nonstandard under the fixed- G asymptotics, and hence the corresponding critical values have to be simulated in practice. This contrasts with the conventional large- G asymptotics, where the limiting distributions are the standard chi-squared and normal distributions. In this section, we show that a simple modification of the two-step Wald and t statistics enables us to develop the standard F and t asymptotic theory under the fixed- G asymptotics. The asymptotic F and t tests are more appealing in empirical applications because the standard F and t distributions are more accessible than the nonstandard $\mathbb{F}_{2\infty}$ and $\mathbb{T}_{2\infty}$ distributions.

The modified two-step Wald and t statistic are

$$\begin{aligned}\tilde{F}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) &: = \frac{G-p-q}{G} \cdot \frac{F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)}{1 + \frac{1}{G}J(\hat{\theta}_2^c)}; \\ \tilde{t}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) &: = \sqrt{\frac{G-1-q}{G}} \cdot \frac{t_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)}{\sqrt{1 + \frac{1}{G}J(\hat{\theta}_2^c)}}.\end{aligned}\tag{18}$$

The modified test statistics involve a scale multiplication factor that uses the usual J statistic and a constant factor that adjusts the degrees of freedom.

It follows from Proposition 8 that

$$\left(F_{\hat{\Omega}^c(\hat{\theta}_2^c)}(\hat{\theta}_2^c), J(\hat{\theta}_2^c)\right) \xrightarrow{d} (\mathbb{F}_{2\infty}, \mathbb{J}_\infty)\tag{19}$$

$$\stackrel{d}{=} \left(\frac{G}{p} (\bar{B}_p - \bar{\mathbb{S}}_{pq} \bar{\mathbb{S}}_{qq}^{-1} \bar{B}_q)' \bar{\mathbb{S}}_{pp\cdot q}^{-1} (\bar{B}_p - \bar{\mathbb{S}}_{pq} \bar{\mathbb{S}}_{qq}^{-1} \bar{B}_q)', G \cdot \bar{B}_q' \bar{\mathbb{S}}_{qq}^{-1} \bar{B}_q\right)\tag{20}$$

Thus,

$$\tilde{F}_{\hat{\Omega}^c(\hat{\theta}_2^c)}(\hat{\theta}_2^c) \xrightarrow{d} \frac{G-p-q}{G} \frac{\mathbb{F}_{2\infty}}{1 + \frac{1}{G}\mathbb{J}_\infty} \stackrel{d}{=} \frac{G-p-q}{pG} \xi_p' \bar{\mathbb{S}}_{pp\cdot q}^{-1} \xi_p,$$

where

$$\xi_p := \frac{\sqrt{G}(\bar{B}_p - \bar{\mathbb{S}}_{pq} \bar{\mathbb{S}}_{qq}^{-1} \bar{B}_q)}{\sqrt{1 + \bar{B}_q' \bar{\mathbb{S}}_{qq}^{-1} \bar{B}_q}}.$$

Similarly,

$$\tilde{t}_{\hat{\Omega}^c(\hat{\theta}_2^c)}(\hat{\theta}_2^c) \xrightarrow{d} \sqrt{\frac{G-1-q}{G}} \cdot \frac{\mathbb{T}_{2\infty}}{\sqrt{1 + \frac{1}{G}\mathbb{J}_\infty}} \stackrel{d}{=} \frac{\xi_p}{\sqrt{\tilde{\mathbb{S}}_{pp\cdot q}}}.$$

In the proof of Theorem 10 we show that ξ_p follows a standard normal distribution $N(0, I_p)$, and ξ_p is independent of $\bar{\mathbb{S}}_{pp\cdot q}^{-1}$. Thus, the limiting distribution of $\tilde{F}_{\hat{\Omega}^c(\hat{\theta}_2^c)}(\hat{\theta}_2^c)$ is proportional to a quadratic form in the standard normal vector ξ_p with an independent inverse-Wishart distributed weighting matrix $\bar{\mathbb{S}}_{pp\cdot q}^{-1}$. It follows from a theory of multivariate statistics that the limiting distribution of $\tilde{F}_{\hat{\Omega}^c(\hat{\theta}_2^c)}(\hat{\theta}_2^c)$ is $\mathcal{F}_{p, G-p-q}$. Similarly, the limiting distribution of $\tilde{t}_{\hat{\Omega}^c(\hat{\theta}_2^c)}(\hat{\theta}_2^c)$ is t_{G-1-q} . This is formalized in the following theorem.

Theorem 10 *Let Assumptions 1~5 hold. Then the modified Wald statistics converges in distribution to $\mathcal{F}_{p, G-p-q}$. Also, the modified t statistics has limiting distribution t_{G-1-q} .*

The limiting t and F results in Theorem are consistent with the recent paper by Hwang and Sun (2017a) which establishes a similar F and t limit theory of two-step GMM in a time series setting. But our cluster-robust limiting distributions in Theorem 10 are different from Hwang and Sun (2017a) in terms of the multiplicative adjustment and the degrees of freedom correction.

It follows from the proofs of Theorem 10 and Proposition 8 that

$$\begin{aligned}\sqrt{n}(\hat{\theta}_2^c - \theta_0) &\xrightarrow{d} MN\left(0, (\Gamma' \Omega^{-1} \Gamma)^{-1} \cdot (1 + \bar{B}_q' \bar{\mathbb{S}}_{qq}^{-1} \bar{B}_q)\right) \text{ and} \\ J(\hat{\theta}_2^c) &\xrightarrow{d} G \cdot \bar{B}_q' \bar{\mathbb{S}}_{qq}^{-1} \bar{B}_q\end{aligned}\tag{21}$$

hold jointly under fixed- G asymptotics. Here, $MN(0, \mathbb{V})$ denotes a random variable that follows a mixed normal distribution with conditional variance \mathbb{V} . The random multiplication term $(1 + \bar{B}'_q \tilde{S}_{qq}^{-1} \bar{B}_q)$ in (21) reflects the estimation uncertainty of CCE weighting matrix on the limiting distribution of $\sqrt{n}(\hat{\theta}_2^c - \theta_0)$. The fixed- G limiting distribution in (21) is in sharp contrast to that of under the conventional large- G asymptotics as the latter completely ignores the variability in the cluster-robust GMM weighting matrix. By continuous mapping theorem,

$$\frac{\sqrt{n}(\hat{\theta}_2^c - \theta_0)}{\sqrt{1 + \frac{1}{G}J(\hat{\theta}_2^c)}} \xrightarrow{d} N\left(0, (\Gamma'\Omega^{-1}\Gamma)^{-1}\right). \quad (22)$$

and this shows that the J statistic modification factor in the denominator effectively cancels out the uncertainty of CCE to recover the limiting distribution of $\sqrt{n}(\hat{\theta}_2^c - \theta_0)$ under the conventional large- G asymptotics. In view of (22), the finite sample distribution of $\sqrt{n}(\hat{\theta}_2^c - \theta_0)$ conditional on the J statistic $J(\hat{\theta}_2^c)$, can be well-approximated by $N(0, \widehat{\text{var}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c))$ where

$$\widehat{\text{var}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) := \widehat{\text{var}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) \cdot \left(1 + \frac{1}{G}J(\hat{\theta}_2^c)\right). \quad (23)$$

The modification term $(1 + (1/G)J(\hat{\theta}_2^c))^{-1}$ degenerates to one as G increases so that the two variance estimates in (23) become close to each other. Thus, the multiplicative term $(1 + (1/G)J(\hat{\theta}_2^c))^{-1}$ in (18) can be regarded as a finite sample modification to the standard variance estimate $\widehat{\text{var}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)$ under the large- G asymptotics. For more discussions about the role of J statistic modification, see Hwang and Sun (2017b) which casts the two-step GMM problems into OLS estimation and inference in classical normal linear regression.

Remark 11 *When the cluster sizes are asymptotically unbalanced, i.e. $L_g/n \rightarrow \lambda_g > 0$ and λ_g does not have to equal to $1/G$, Supplemental Appendix B.5 shows that the centered CCE weakly converges to a random matrix $\Lambda\bar{M}\Lambda'$, where*

$$\bar{M} := \bar{M}(\lambda) = \sum_{g=1}^G \left(\sqrt{\lambda_g} B_{m,g} - \lambda_g \sum_{h=1}^G \sqrt{\lambda_h} B_{m,h} \right) \left(\sqrt{\lambda_g} B_{m,g} - \lambda_g \sum_{h=1}^G \sqrt{\lambda_h} B_{m,h} \right)'$$

Clearly, the submatrices \bar{M}_{pp} and \bar{M}_{qq} do not follow (scaled) Wishart distributions unless for $\lambda_1 = \dots = \lambda_G = 1/G$. As a result, $\bar{M}_{pp \cdot q} = \bar{M}_{pp} - \bar{M}_{pq}\bar{M}_{qq}^{-1}\bar{M}_{qp}$ does not follow a (scaled) Wishart distribution and is not independent of \bar{M}_{pq} and \bar{M}_{qq} . In addition, ξ_p is not distributed as $N(0, I_p)$ which are the key conditions that drive the F and t limit theory. Thus, in the absence of the approximately equal cluster size in Assumption 1-iii), an exact asymptotic F theory or t theory can not be developed.

Remark 12 *Another class of popular GMM estimators is the continuous updating (CU) estimators, which are designed to improve the poor finite sample performance of two-step GMM estimators, e.g. Hansen, Heaton, and Yaron (1996) and Newey and Smith (2004). Supplemental Appendix B.6 presents two types of CU schemes, CU-GMM and CU-GEE, and shows that the uncentered CU estimators and the centered two-step GMM estimator are asymptotically equivalent under the fixed- G asymptotics. Thus, the CU estimators can be regarded as having a built-in recentering mechanism similar to the centered two-step GMM estimator.*

5 Finite Sample Variance Correction

Although the recentering scheme we investigate in the previous sections enables us to remove the first order effect of the first-step estimation error, the centered two-step GMM estimator still faces some extra estimation uncertainty when the unobserved moment process in the GMM weight matrix is estimated using the first-step estimator. To solve this problem, Windmeijer (2005) proposes a finite sample corrected variance formula which is derived by a second order stochastic approximation of linear GMM model. Windmeijer (2005)'s corrected variance formula has been widely implemented in applied work with high impact for linear GMM models such as instrumental regression models and dynamic panel data models.⁷ See, for example, Roodman (2009), Brown et al. (2009), Oberholzer-Gee and Strumpf (2007), and many others. However, the original corrected variance formula in Windmeijer (2005) is not applicable in our clustered dependence setting because its key assumption is that the true moment process is i.i.d.

In this section, we overcome the limited applicability of Windmeijer (2005)'s one in the presence of clustered dependence and develop a finite-sample corrected variance formula for the feasible two-step GMM estimator. The new variance formula can be thought of as refining the first-order fixed-G asymptotics. We first derive a second order stochastic expansion of linear GMM two-step estimator which uses the CCE weight matrix. To be more specific, define the infeasible two-step GMM estimator with the centered CCE weighting matrix $\hat{\Omega}^c(\theta_0)$ as

$$\tilde{\theta}_2^c = \arg \min_{\theta \in \Theta} g_n(\theta)' \left(\hat{\Omega}^c(\theta_0) \right)^{-1} g_n(\theta).$$

When the moment condition is linear in parameter vector, we obtain

$$\sqrt{n}(\tilde{\theta}_2^c - \theta_0) = - \left[\hat{\Gamma}' \left(\hat{\Omega}^c(\theta_0) \right)^{-1} \hat{\Gamma} \right]^{-1} \hat{\Gamma}' \left(\hat{\Omega}^c(\theta_0) \right)^{-1} \sqrt{n}g_n(\theta_0),$$

and

$$\sqrt{n}(\hat{\theta}_2^c - \theta_0) = - \left[\hat{\Gamma}' \left(\hat{\Omega}^c(\hat{\theta}_1) \right)^{-1} \hat{\Gamma} \right]^{-1} \hat{\Gamma}' \left(\hat{\Omega}^c(\hat{\theta}_1) \right)^{-1} \sqrt{n}g_n(\theta_0). \quad (24)$$

Together with the result in Lemma 7, $\hat{\Omega}^c(\hat{\theta}_1) = \hat{\Omega}^c(\theta_0) + o_p(1)$, this implies that

$$\sqrt{n}(\hat{\theta}_2^c - \theta_0) = \sqrt{n}(\tilde{\theta}_2^c - \theta_0) + o_p(1), \quad (25)$$

as $n \rightarrow \infty$ holding G is fixed. That is, the estimation error in $\hat{\theta}_1$ has no effect on the asymptotic distribution of $\sqrt{n}(\hat{\theta}_2^c - \theta_0)$ in the first-order asymptotic analysis. However, in finite samples, $\hat{\theta}_2^c$ does have higher variation than $\tilde{\theta}_2^c$, and this can be attributed to the high variation in $\hat{\Omega}^c(\hat{\theta}_1)$ than $\hat{\Omega}^c(\theta_0)$. However, we can be more explicit in the extra variation of $\sqrt{n}(\hat{\theta}_2^c - \theta_0)$ by keeping the leading order term of $o_p(1)$ in (25) as

$$\begin{aligned} \sqrt{n}(\hat{\theta}_2^c - \theta_0) &= - \left[\hat{\Gamma}' \left(\hat{\Omega}^c(\hat{\theta}_1) \right)^{-1} \hat{\Gamma} \right]^{-1} \hat{\Gamma}' \left(\hat{\Omega}^c(\hat{\theta}_1) \right)^{-1} \sqrt{n}g_n(\theta_0) \\ &\quad + (\mathcal{E}_{1n} + \mathcal{E}_{1n})\sqrt{n}(\hat{\theta}_1 - \theta_0) + o_p\left(\frac{1}{\sqrt{n}}\right), \end{aligned} \quad (26)$$

⁷More than 5,000 citations according to Google Scholar in March, 2020.

where \mathcal{E}_{1n} and \mathcal{E}_{2n} are $d \times d$ matrices with

$$\mathcal{E}_{1n} = - \left. \frac{\partial \left\{ \hat{\Gamma}' \left[\hat{\Omega}^c(\theta) \right]^{-1} \hat{\Gamma} \right\}^{-1}}{\partial \theta'} \right|_{\theta=\theta_0} \hat{\Gamma}' \left[\hat{\Omega}^c(\theta_0) \right]^{-1} g_n(\theta_0);$$

$$\mathcal{E}_{2n} = - \left\{ \hat{\Gamma}' \left[\hat{\Omega}^c(\theta) \right]^{-1} \hat{\Gamma} \right\}^{-1} \left. \frac{\partial \hat{\Gamma}' \left[\hat{\Omega}^c(\theta) \right]^{-1} g(\theta)}{\partial \theta'} \right|_{\theta=\theta_0},$$

respectively. It is easy to check both \mathcal{E}_{1n} and \mathcal{E}_{2n} are $O_p(n^{-1/2})$, so that the term $(\mathcal{E}_{1n} + \mathcal{E}_{2n})\sqrt{n}(\hat{\theta}_1 - \theta_0)$ in (26) exactly captures the leading order of errors in (25).

Considering the leading higher order term in (26), our goal is to come up with a finite sample corrected variance formula. To do so, we seek for a refined distributional approximation of $\sqrt{n}(\hat{\theta}_2^c - \theta_0)$ based on (26). Note that, from the FOC of $\hat{\theta}_2^c$ if we try to estimate the term $\hat{\Gamma}'[\hat{\Omega}^c(\theta_0)]^{-1}g_n(\theta_0)$ in \mathcal{E}_{1n} by $\hat{\Gamma}'[\hat{\Omega}^c(\hat{\theta}_1)]^{-1}g_n(\hat{\theta}_2^c)$, the estimate will be identically zero almost surely. For this reason, we can drop \mathcal{E}_{1n} and consider only \mathcal{E}_{2n} , and consider the following distributional approximation

$$\sqrt{n}(\hat{\theta}_2^c - \theta_0) \stackrel{a}{\sim} - \left(\left[\Gamma'(\Omega_\infty^c)^{-1} \Gamma \right]^{-1} \tilde{\mathcal{E}}_{2n}(\Gamma'W^{-1}\Gamma)^{-1} \right) \begin{pmatrix} \Gamma'(\Omega_\infty^c)^{-1} \Lambda Z \\ \Gamma'W\Lambda Z \end{pmatrix}, \quad (27)$$

where $Z \sim N(0, I_d)$, Z is independent of Ω_∞^c , and $\tilde{\mathcal{E}}_{2n}$ has the same marginal distribution as \mathcal{E}_{2n} and is independent of Z and Ω_∞^c . Here, the notation $\stackrel{a}{\sim}$ in (27) denotes that the stochastically bounded sequences of random vectors ξ_n and η_n converge in distribution to the same limit.

Since the $\tilde{\mathcal{E}}_{2n}$ (\mathcal{E}_{2n}) converges to zero in probability, the multiplicative term $\tilde{\mathcal{E}}_{2n}(\Gamma'W^{-1}\Gamma)^{-1}$ on the right-hand side in (27) has no first order effect to characterize the first-order asymptotic distribution of $\sqrt{n}(\hat{\theta}_2^c - \theta_0)$. However, the smaller order term at the approximating distribution in (27) motivates us to develop a finite sample correction to the asymptotic variance estimator. From (27), it follows that $\sqrt{n}(\hat{\theta}_2^c - \theta_0)$ is asymptotically equivalent in distribution to the mixed normal distribution whose conditional variance is given by

$$\Xi_n = \left(\begin{pmatrix} \left[\Gamma'(\Omega_\infty^c)^{-1} \Gamma \right]^{-1} \\ (\Gamma'W^{-1}\Gamma)^{-1} \tilde{\mathcal{E}}'_{2n} \end{pmatrix} \right)' \begin{pmatrix} \Gamma'(\Omega_\infty^c)^{-1} \Omega (\Omega_\infty^c)^{-1} \Gamma & \Gamma'(\Omega_\infty^c)^{-1} \Omega W^{-1} \Gamma \\ \Gamma'W^{-1} \Omega' (\Omega_\infty^c)^{-1} \Gamma & \Gamma'W^{-1} \Omega W^{-1} \Gamma \end{pmatrix} \begin{pmatrix} \left[\Gamma'(\Omega_\infty^c)^{-1} \Gamma \right]^{-1} \\ (\Gamma'W^{-1}\Gamma)^{-1} \tilde{\mathcal{E}}'_{2n} \end{pmatrix}.$$

From this, we propose to use the following corrected variance estimator:

$$\begin{aligned} \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}^w(\hat{\theta}_2^c) &= \frac{1}{n} \hat{\Xi}_n \\ &= \frac{1}{n} \left(\left[\hat{\Gamma}' \left[\hat{\Omega}^c(\hat{\theta}_1) \right]^{-1} \hat{\Gamma} \right]^{-1} \hat{\mathcal{E}}_n(\hat{\Gamma}'W_n^{-1}\hat{\Gamma})^{-1} \right) \times \begin{pmatrix} \hat{\Gamma}' \left[\hat{\Omega}^c(\hat{\theta}_1) \right]^{-1} \hat{\Gamma} & \hat{\Gamma}'W_n^{-1}\hat{\Gamma} \\ \hat{\Gamma}'W_n^{-1}\hat{\Gamma} & \hat{\Gamma}'W_n^{-1}\hat{\Omega}^c(\hat{\theta}_1)W_n^{-1}\hat{\Gamma} \end{pmatrix} \\ &\times \left(\begin{pmatrix} \left[\hat{\Gamma}' \left[\hat{\Omega}^c(\hat{\theta}_1) \right]^{-1} \hat{\Gamma} \right]^{-1} \\ (\hat{\Gamma}'W_n^{-1}\hat{\Gamma})^{-1} \hat{\mathcal{E}}'_n \end{pmatrix} \right) \\ &= \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) + \hat{\mathcal{E}}_n \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) + \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) \hat{\mathcal{E}}'_n + \hat{\mathcal{E}}_n \widehat{var}(\hat{\theta}_1) \hat{\mathcal{E}}'_n, \end{aligned} \quad (28)$$

where the j -th column of $\widehat{\mathcal{E}}_n$ is

$$\widehat{\mathcal{E}}_{n[\cdot, j]} = \left\{ \widehat{\Gamma}' \left[\widehat{\Omega}^c(\widehat{\theta}_1) \right]^{-1} \widehat{\Gamma}' \right\}^{-1} \widehat{\Gamma}' \left\{ \left[\widehat{\Omega}^c(\widehat{\theta}_1) \right]^{-1} \frac{\partial \widehat{\Omega}^c(\theta)}{\partial \theta_j} \Big|_{\theta=\widehat{\theta}_1} \left[\widehat{\Omega}^c(\widehat{\theta}_1) \right]^{-1} \right\} g_n(\widehat{\theta}_2^c),$$

and

$$\Upsilon_j(\theta_0) = \frac{1}{G} \sum_{g=1}^G \left[\frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} \left(f_i^g(\theta_0) - \frac{1}{n} \sum_{i=1}^n f_i(\theta_0) \right) \frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} \left(\frac{\partial f_i^g(\theta_0)}{\partial \theta_j} - \frac{1}{n} \sum_{i=1}^n \frac{\partial f_i(\theta_0)}{\partial \theta_j} \right) \right]'$$

The last three terms in (28), which are of smaller order, serve as a finite sample correction to the original variance estimator.

Besides our asymptotic variance formula in (28) extends Windmeijer (2005)'s one which considers only an i.i.d. setting, there are two principal differences between Windmeijer's approach and ours. First, our asymptotic variance estimator in (28) involves a centered CCE; in contrast, Windmeijer's formula involves only a plain variance estimator. Second, we consider the fixed-G asymptotics; Windmeijer (2005) considers the traditional asymptotics $n \rightarrow \infty$ in an i.i.d setting. To our knowledge, this paper is the first to rigorously prove the asymptotic validity of the Windmeijer type finite sample corrected variance under the fixed-G asymptotic point of view.

With the finite sample corrected variance estimator, we can construct the variance-corrected Wald and t statistics:

$$F_{\widehat{\Omega}^c(\widehat{\theta}_1)}^w(\widehat{\theta}_2^c) = \frac{1}{p} (R\widehat{\theta}_2^c - r)' \left[R \widehat{\text{var}}_{\widehat{\Omega}^c(\widehat{\theta}_1)}^w(\widehat{\theta}_2^c) R' \right]^{-1} (R\widehat{\theta}_2^c - r);$$

$$t_{\widehat{\Omega}^c(\widehat{\theta}_1)}^w(\widehat{\theta}_2^c) = \frac{R\widehat{\theta}_2^c - r}{\sqrt{R \widehat{\text{var}}_{\widehat{\Omega}^c(\widehat{\theta}_1)}^w(\widehat{\theta}_2^c) R'}}.$$

Given that the variance correction terms are of smaller order, the variance-corrected statistic will have the same limiting distribution as the original statistic.

Assumption 6 For each $g = 1, \dots, G$ and $j = 1, \dots, d$, define $Q_j^g(\theta)$ as

$$Q_j^g(\theta) = \lim_{L_g \rightarrow \infty} E \left[\frac{1}{L_g} \sum_{i=1}^{L_g} \frac{\partial}{\partial \theta'} \left(\frac{\partial f_i^g(\theta)}{\partial \theta_j} \right) \right].$$

Then,

$$\sup_{\theta \in \mathcal{N}(\theta_0)} \left\| \frac{1}{L_g} \sum_{i=1}^{L_g} \frac{\partial}{\partial \theta'} \left(\frac{\partial f_i^g(\theta)}{\partial \theta_j} \right) - Q_j^g(\theta) \right\| \xrightarrow{p} 0$$

holds for each $g = 1, \dots, G$ and $j = 1, \dots, d$, where $\mathcal{N}(\theta_0)$ is an open neighborhood of θ_0 , and $\|\cdot\|$ is the Euclidean norm. Also, $Q_j^g(\theta_0) = Q_j(\theta_0)$ for $g = 1, \dots, G$.

This assumption trivially holds if the moment conditions are linear in parameters.

Theorem 13 Let Assumptions 1~6 hold. Then

$$F_{\widehat{\Omega}^c(\widehat{\theta}_1)}^w(\widehat{\theta}_2^c) = F_{\widehat{\Omega}^c(\widehat{\theta}_1)}(\widehat{\theta}_2^c) + o_p(1);$$

$$t_{\widehat{\Omega}^c(\widehat{\theta}_1)}^w(\widehat{\theta}_2^c) = t_{\widehat{\Omega}^c(\widehat{\theta}_1)}(\widehat{\theta}_2^c) + o_p(1).$$

In the proof of Theorem 13, we show that $\widehat{\mathcal{E}}_n = (1 + o_p(1))\mathcal{E}_{2n}$. That is, the high order correction term has been consistently estimated in a relative sense. This guarantees that $\widehat{\mathcal{E}}_n$ is a reasonable estimator for \mathcal{E}_{2n} , which is of order $O_p(n^{-1/2})$.

It is important to point out that our proposed formula captures the higher-order term for linear models. For non-linear models, the order of the remainder term is the same as the correction terms, so the corrections may not necessarily provide finite-sample improvements. This is because the non-linear GMM estimator for the Jacobian matrix $\widehat{\Gamma}(\widehat{\theta}_1)$ also depends on the plugged-in parameter, and corresponding stochastic expansion is not counted on our formula. However, the proof Theorem 13 does not rely on the linearity in moment condition and shows that our proposed finite-sample corrected variance estimator is still consistent. Thus, the robustness property in Theorem 13 holds for both linear and nonlinear models.

As a direct implication of Theorem 13 together with Theorem 10, the Wald and t statistics coupled with the J statistic modification and the finite sample variance correction have the standard F and t limiting distributions found in Theorem 10. That is,

$$\widetilde{F}_{\widehat{\Omega}^c(\widehat{\theta}_1)}^w(\widehat{\theta}_2^c) := \frac{G - p - q}{G} \cdot \frac{F_{\widehat{\Omega}^c(\widehat{\theta}_1)}^w(\widehat{\theta}_2^c)}{1 + \frac{1}{G}J(\widehat{\theta}_2^c)} \xrightarrow{d} \mathcal{F}_{p, G-p-q} \quad (29)$$

and

$$\widetilde{t}_{\widehat{\Omega}^c(\widehat{\theta}_1)}^w(\widehat{\theta}_2^c) := \sqrt{\frac{G - 1 - q}{G}} \frac{t_{\widehat{\Omega}^c(\widehat{\theta}_1)}^w(\widehat{\theta}_2^c)}{\sqrt{1 + \frac{1}{G}J(\widehat{\theta}_2^c)}} \xrightarrow{d} t_{G-1-q}. \quad (30)$$

The multiplicative modification provided in Section 4 can turn the nonstandard distributions into the standard F and t distributions in (29) and (30), respectively. Compared to other types of GMM tests, our simulation evidence demonstrates the size accuracy of the above Wald and t statistics. Also, the tests are very appealing to practitioners because they can implement the standard t and F critical values to the finite-sample corrected test statistics. No further simulations or re-sampling methods are needed. Thus, we recommend practitioners to use these statistics.

Remark 14 *More broadly, we may use the higher-order terms in (26) to develop an Edgeworth expansion. The Edgeworth expansion targets a higher-order expansion of the finite-sample distribution of the cluster robust t and F statistics. However, the Edgeworth expansion usually requires a strong set of assumptions for the true moment process. For example, to derive the Edgeworth expansion for the cluster robust test statistics, Djogbenou et al. (2019) requires to assume a single number of hypothesis test ($p = 1$), an exactly identified linear regression model, a strict independence across clusters, increasing number of clusters G as $n \rightarrow \infty$ with bounded cluster sizes L_g , and a non-degenerate absolutely continuous component of probability distribution (Cramér's condition). The present paper, however, does not require any of these assumptions.*

Remark 15 *In addition to the technical reasons, another reason for avoiding the Edgeworth expansion is that we have already achieved more accurate fixed- G asymptotic approximations of t and F statistics. After capturing the variation of the CCE matrix via the fixed- G asymptotics, we correct a higher-order bias of the asymptotic variance matrix using a stochastic expansion of the feasible two-step GMM estimator in (26).*

Remark 16 *Following Newey and Smith (2004), one can derive a more formal stochastic expansion than (26) and capture additional smaller order errors arising from W_n and $\hat{\Gamma}$ as well as $\hat{\Omega}^c(\hat{\theta}_1)$. A recent paper by Hwang et al. (2019) shows that the additional correction for these terms in i.i.d GMM provides robustness to misspecification in the over-identified moment condition. Technically, it is not difficult to extend our clustered asymptotic variance formula by considering the extra corrections provided in Hwang et al. (2019). However, simulation results in Hwang et al. (2019) show that, under correctly specified moment condition which is the main focus of this paper, there is not much difference between the two formulas. It will be interesting to develop misspecification robust GMM inferences using our fixed- G asymptotics, and we leave it as future research.*

6 Simulation Evidence

6.1 Design

In the presence of both time series and cross-sectional dependence, we compare the finite sample performance of our new tests by focusing on the following linear dynamic panel data model:

$$y_{it} = \gamma y_{it-1} + x_{1,it}\beta_1 + \dots + x_{d-1,it}\beta_{d-1} + \eta_i + u_{it}.$$

The unknown parameter vector is $\theta = (\gamma, \beta_1, \dots, \beta_{d-1})' \in \mathbb{R}^d$, and the corresponding covariates are $w_{it} = (y_{it-1}, x_{it})' \in \mathbb{R}^d$ with $x_{it} = (x_{1,it}, \dots, x_{d-1,it})' \in \mathbb{R}^{d-1}$. We set the number of parameters d as 4 and the true value of θ as $\theta_0 = (0.5, 1, 1, 1)'$. The k -th predetermined regressor $x_{k,it}$ are generated according to the following process:

$$x_{k,it} = \rho x_{k,it-1} + \eta_i + \rho u_{it-1} + e_{k,it},$$

for $k = 1, 2, d-1$, $i = 1, \dots, n$, and $t = 1, \dots, T$. Let $\eta = (\eta_1, \dots, \eta_n)'$, $u_t = (u_{1t}, \dots, u_{nt})$, and $e_{k,t} = (e_{k,1t}, \dots, e_{k,nt})'$. Setting the number of time periods to be $T = 4$, we characterize the cross-sectional dependence in η , e_t , and u_t by spatial locations that are indexed by a one-dimensional lattice. Define Σ_η and Σ_u to be $n \times n$ matrices whose (i, j) -th elements are $\sigma_{ij}^\eta = \lambda^{|i-j|}$ and $\sigma_{ij}^u = \lambda^{|i-j|}$, respectively. Similarly, the (i, j) -th element of $\Sigma_{k,e}$ is $\sigma_{k,ij}^e = \lambda^{|i-j|}$ for $k = 1, \dots, d-1$. The parameter λ governs the degree of spatial dependence. When $\lambda = 0$ with $\lambda^0 = 1$, there is no spatial dependence and our model reduces to that of Windmeijer (2005) which considers a dynamic panel data model with only one regressor. We divide the sample of size n into G equal-sized groups of consecutive observations so that the individuals are correlated across different clusters but asymptotically (mean) independent. The smaller the size of clusters $L = n/G$, the less asymptotic independence presents across different groups. The individual fixed effects and shocks are generated by

$$\begin{aligned} \eta &\sim N(0, \Sigma_\eta), e_{k,t} \stackrel{\text{i.i.d.}}{\sim} N(0, \Sigma_{k,e}) \text{ for } k = 1, \dots, d-1, \\ u_t &:= \tau_t \Sigma_u^{1/2} (\delta_1 \omega_{1t}, \dots, \delta_n \omega_{nt})', \\ \delta_i &\stackrel{\text{i.i.d.}}{\sim} U[0.5, 1.5], \text{ and } \omega_{it} \stackrel{\text{i.i.d.}}{\sim} \chi_1^2 - 1, \end{aligned} \tag{31}$$

over $i = 1, \dots, n$, and $t = 1, \dots, T$, where $\tau_t = 0.5 + 0.1(t-1)$. The DGP of individual component $u_{i,t}$ in u_t features a non-Gaussian process which is heteroskedastic over both time t and individual i . Before we draw an estimation sample for $t = 1, \dots, T$, 50 initial values are generated with

$\tau_t = 0.5$ for $t = -49, \dots, 0$, $x_{k,-49} \sim N(\eta_i/(1-\rho), (1-\rho)^{-1}\Sigma_{k,e})$ for $k = 1, \dots, d-1$, and $y_{i,-49} = (\sum_{d=1}^3 x_{d,i,-49}\beta_d + \eta_i + u_{i,-49})/(1-\gamma)$. We fix the values of λ and ρ at 0.60; thus each observation is reasonably persistent with respect to both time and spatial dimensions.⁸ The parameters are estimated by the first differenced GMM (Arellano and Bond estimator). In supplemental Appendix B.7, we describe in details how to implement the GMM inference procedures considered in this section. With all possible lagged instruments $z_{it} = (y_{i0}, \dots, y_{it-2}, x'_{i1}, \dots, x'_{it-1})'$, $2 \leq t \leq T$, the number of moment conditions for the Arellano and Bond estimator is $m = dT(T-1)/2$. It could be better to use only a subset of full moment conditions because using this full set of instruments may lead to poor finite sample properties, especially when the number of clusters G is small. Thus, we also employ a reduced set of instruments; that is, we use the most recent lag $z_{it} = (y_{it-2}, x'_{it-1})'$, leading to $d(T-1)$ moment conditions. The initial first-step estimator is chosen by 2SLS with $W_n = n^{-1} \sum_{i=1}^n Z_i' Z_i$, where $Z_i = \text{diag}(z'_{i2}, \dots, z'_{iT})$ is a $(T-1) \times m$ matrix.

6.2 Choice of tests

We examine the empirical size of a variety of testing procedures, all of which are based on the first-step or two-step GMM estimators. For the first-step procedures, we consider the unmodified F statistic $F_1 := F_1(\hat{\theta}_1)$ and the degrees-of-freedom modified F statistic $[(G-p)/G] F_1$, where the associated critical values $\chi_p^{1-\alpha}/p$ justified under the large- G asymptotics, and $\mathcal{F}_{p,G-p}^{1-\alpha}$ under the fixed- G asymptotics, respectively. Note that these two tests have the same size-adjusted power, because the modification only involves a constant multiplier factor.

For the two-step GMM estimation and related tests, we examine the four different procedures that are based on the centered CCE. The first test uses the “plain” F statistic $F_2 := F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)$ in (15), where its critical value $\chi_p^{1-\alpha}/p$ is justified by the large- G asymptotics. The second test uses the modified $\tilde{F}_2 := F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)$ in (18). Compared to the plain two-step GMM F statistic, \tilde{F}_2 has the additional J statistic correction factor $(1 + (q/G)J(\hat{\theta}_2^c))^{-1}$. The third test, $\tilde{F}_2^w = F_{\hat{\Omega}^c(\hat{\theta}_1)}^w(\hat{\theta}_2^c)$, uses the most refined version of the F statistic coupled with the J statistic modification, degrees-of-freedom, and finite sample corrected variance estimator which is defined in (29). The second and third tests employ the new F critical value $\mathcal{F}_{p,G-p-q}^{1-\alpha}$ which is justified under the fixed- G asymptotics. Lastly, we consider a bootstrap procedure of the centered two-step GMM test originally proposed by Hall and Horowitz (1996).⁹ Note that the consistency and the higher-order refinement of Hall-Horowitz bootstrap procedure require the number of cluster G tends to infinity. This is contrast to the previous two tests that are valid under the fixed- G asymptotics.

⁸When the panel data are persistent with ρ being close to one, the lagged instruments are only weakly correlated with the endogenous changes in the first differenced data, and the GMM inferences considered in our paper can suffer a weak identification problem (e.g., Blundell and Bond, 1998; Stock and Wright, 2000; Bun and Windmeijer, 2010). It will be interesting to extend our approach to develop weak identification robust GMM inferences under clustered dependence, and we leave this as a future research.

⁹See supplementary Appendix B.8 for the details about how to implement the bootstrap procedure of Hall and Horowitz (1996) in the presence of clustered dependence.

6.3 Results with balanced and homogeneous cluster size

6.3.1 Size experiment

We first consider the case when we take different values of $G \in \{35, 50, 70, 100\}$, and the equal number of cluster size $L = L_1 = \dots = L_g \in \{50, 100\}$. The null hypotheses of interests are

$$H_{01} : \beta_{10} = 1, \quad H_{02} : \beta_{10} = \beta_{20} = 1, \quad H_{03} : \beta_{10} = \beta_{20} = \beta_{30} = 1,$$

with the corresponding number of joint hypotheses $p = 1, 2$ and 3 , respectively, and the significance level is 5%. All of our simulation results are based on 5,000 times of Monte Carlo repetition, and the number of bootstrap replication is 1,000.

Tables 1 reports the empirical size of the first-step and two-step tests for different values of G 's we consider and $L = 50$. We only report the results when $L = 50$ with $G = 50, 100$, as the qualitative observations for other cases remain quite similar. The results first indicate that both the first-step and two-step tests based on unmodified statistics F_1 and F_2 suffer from severe size distortions, when the conventional chi-squared critical values are used. For example, with $G = 50$ and $p = 3$, the empirical size of the first-step chi-squared test (using the full set of IVs, and $m = 24$) is 24.2%. The empirical sizes of the first-step F test reduce to 19.1% when the F critical values are employed. This finding is consistent with the findings in BCH (2011) and Hansen (2007), which highlight the improved finite sample performance of the fixed- G approximation in the exactly identified models. Tables 1 also indicates that the finite sample size distortion of all tests become less severe as the number of moment conditions decreases or the number of cluster size G increases.

For the two-step tests that employ the plain two-step statistic F_2 with the chi-squared critical values, the empirical sizes are between 23.7%–53.5% for $m = 24$, and $p = 3$. In view of the large size distortion, we can conclude that the two-step chi-squared test suffers more size distortion than the first-step chi-squared test. This relatively large size distortion reflects the additional cost in estimating the weighting matrix, which is not captured by the chi-square approximation. This motivates us to implement additional corrections via degrees of freedom and the J statistic multiplier coupled with the new critical value $\mathcal{F}_{p, G-p-q}^{1-\alpha}$. Tables 1 shows that the additional modifications with the standard F critical value significantly alleviate the distortion. The size distortions in the previous example are reported to be between 6.1% and 6.9% which are much closer to the targeted level 5%. Lastly, we find evidence that the most refined statistic \tilde{F}_2^w , equipped with the finite sample variance correction, results in the empirical sizes between 5.3%–5.9%. This indicates the most refined two-step Wald test successfully captures the higher order estimation uncertainty and yields more accurate finite sample size. We find similar conclusions for other values of L , m , and p .

Tables 1 also shows the empirical rejection probabilities of the two-step GMM bootstrap procedure by Hall and Horowitz (1996), which is denoted HH-Bootstrap in the tables. We find that the HH-Bootstrap is severely undersized when the number of clusters G is small, for example, when G is 50 with $m = 24$ and $p = 1-3$, the empirical sizes are between 0.2% and 2.0%. This fragility of the HH-Bootstrap procedure has been also observed by Bond and Windmeijer (2005) and Windmeijer (2005) in their Monte Carlo analysis of the cross-sectionally independent dynamic panel data estimated by GMM. They point out that the GMM inferences based on the bootstrap procedures become less reliable when there is a problem in estimating the GMM weighting matrix with the sample moment process. Our simulation results extend their findings in the two-step GMM procedures to those in the presence of clustered dependence. We also note

that the empirical rejection probabilities of the GMM bootstrap procedure become close to the nominal size when the reduced set of IV ($m = 12$) is used or the number of cluster G increases.

Lastly, Tables 1 shows that the finite sample size distortions of the (centered) J test, $J^c = J(\hat{\theta}_2^c)$ are also substantially reduced and close to the nominal size of 5% when we employ the F critical values instead of the conventional chi-squared critical values and the GMM bootstrap procedure by Hall and Horowitz (1996).

6.3.2 Power experiment

We investigate the finite sample power performances of the first-step procedure and the two-step procedures. We use the finite sample critical values under the null, so the power is size-adjusted and the power comparison is meaningful. The DGPs are the same as before except that the parameters are generated from the local null alternatives $\beta_1 = \beta_{10} + c/\sqrt{n}$ for $c \in [0, 15]$ and $p = 1$. We simulate power curves for the first-step and two-step tests for $G \in \{35, 50, 70, 100\}$ and $L = 50$. To save space, we only report the case when $G = 35$ and 100 in Figures 1 and 2, respectively. The results first indicate that there is no real difference between power curves of the modified (\tilde{F}_2) and unmodified (F_2) two-step tests. In fact, some simulation results not reported here indicate the modified F test can be slightly more powerful as the number of parameters gets larger. Also, the finite sample corrected test \tilde{F}_2^w does not lead to a loss of power compared with the uncorrected one \tilde{F}_2 .

Figures 1–2 also indicate that the two-step tests are more powerful than the first-step tests in most cases of G, m , and p we consider. The power gain of the two-step GMM procedures becomes more significant as the number of G increases. This can be justified by the asymptotic efficiency of the two-step GMM estimator under the large- G asymptotics. However, under the fixed- G asymptotics, there is a cost in estimating the CCE weighting matrix, and the power of first-step procedures might dominate the power of the two-step ones when the cost of employing CCE weighting matrix outweighs the benefit of estimating it. In fact, Figures 1 shows that the power of the first-step test can be higher than that of two-step tests when G is small and m is large, say, for example, $G = 35$ and $m = 24$. See Hwang and Sun (2018) who compare these two types of tests in a time series GMM framework by employing more accurate fixed-smoothing asymptotics which are in the same spirit of the fixed- G asymptotics.

In sum, our simulation evidence clearly demonstrates the size accuracy of our most refined F test, $\tilde{F}_{\Omega^c(\hat{\theta}_1)}^w(\hat{\theta}_2^c)$, regardless of whether the number of clusters G is small or moderate.

6.4 Results with unbalanced and heterogeneous clusters

6.4.1 Unbalanced Clusters

Although our fixed- G asymptotics is valid as long as the cluster sizes are approximately equal, we remain wary of the effect of the cluster size heterogeneity on the quality of the fixed- G approximation. In this subsection, we turn to simulation designs with heterogeneous cluster sizes. Each simulated data set consists of 5,000. Each simulated data set consists of 5,000 observations that are divided into 50 clusters. The sequence of alternative cluster-size designs starts by assigning 120 individuals to each of first 10 clusters and 95 individuals to each of next 40 clusters. In each succeeding cluster-size design, we subtract 10 individuals from the second group of clusters and add them to the first group of clusters. In this manner, we construct a series of four cluster-size designs, in which the proportion of the samples in the first group of clusters

grows monotonically from 24% to 48%. The design is similar to Carter, Schnepel and Steigerwald (2017) which investigates the behavior of cluster-robust t statistic under cluster heterogeneity. Table 2 describes the unbalanced cluster-size Designs I-IV we consider. All other parameter values are the same as before.

Tables 3–5 report the empirical sizes of the GMM procedures we considered in the previous subsections. The results immediately indicate that the two-step tests suffer from severe size distortion when the conventional chi-squared critical value is employed. For example, under Design III, the empirical size of the “plain” two-step chi-squared test is 58.6% for $m = 24$, and $p = 3$. This size distortions become more severe when the degree of heterogeneity across cluster-size increases, e.g., 62.1% for Design IV. However, our fixed- G asymptotics still performs very well even with unbalanced cluster sizes as they substantially reduce the empirical sizes. For example, under Designs III and IV, the most refined two-step Wald statistic \bar{F}_2^w results in the empirical size 5.8% and 7.4% , respectively, for the above mentioned values of m and p , which is much closer to the nominal size. Similar results for other types of GMM tests are reported in Tables 3~5. The results of J tests are omitted here as they are qualitatively similar to those of the F tests.

6.4.2 Heterogeneous Clusters

In our last experiments, we investigate how violation of the cluster homogeneity conditions in Assumptions 4 and 5 impacts the finite sample performance of cluster-robust tests. There are two important nuisance parameters in the dynamic panel model considered in our simulations—the spatial autoregressive parameter λ for the innovations $\{u_{it}, e_{k,it}\}$, and the autoregressive parameter ρ for the regressors $x_{k,it}$. Both of these parameter choices affect the variance and Jacobian of each cluster. To consider the impact of pronounced heterogeneity across clusters, we allow these parameters to be group specific, i.e. (ρ_g, λ_g) for $g = 1, \dots, G$. The alternative DGPs consider randomly drawn ρ_g and λ_g at each simulations. The probability distributions for (ρ_g, λ_g) have two supports at $(0.35, 0.85)^2 \in \mathbb{R}^2$. Figure 8 shows four specific designs of probability distributions considered in our simulation. Design 1 describes a mild degree of heterogeneity where the values for (ρ_g, λ_g) slightly deviates from the benchmark value $(0.60, 0.60)$. Designs II-IV illustrate more substantial levels of heterogeneities, each of which represents a left, right-skewed, and symmetric distribution. To reduce the computational burden, we assume observations are independent across different clusters. All other settings are same as our original model in subsection 6.3.

With $p = 1$, we consider the following three types of t-tests. The one-step GMM test, with its asymptotic critical value t_{G-1} , is referred to BCH. The two-step GMM approach, together with J-statistics modification and finite-sample corrected variance formula, has asymptotic critical value t_{G-q-1} and is referred to Hwang. Note that both of these tests estimate parameters using the entire $n = \sum_{g=1}^G L_g$ observations. The last method is the Fama-Macbeth type procedure in Ibragimov and Müller (2010, 2016, IM hereafter). The IM’s t-test is formed by a cluster-specific estimator which only use observations within the cluster. Letting $R\hat{\theta}_g$ be the cluster-specific estimator for each $g = 1, \dots, G$, the IM’s t-statistic is

$$t_{\text{IM}} = \frac{\sqrt{G}R(\bar{\theta}_G - \theta_0)}{\sqrt{\sum_{g=1}^G (R\hat{\theta}_g - R\bar{\theta}_G)^2 / (G - 1)}},$$

where $\bar{\theta}_G = G^{-1} \sum_{g=1}^G \hat{\theta}_g$. IM (2010, 2016) shows that using t_{IM} with critical value t_{G-1} yields asymptotically valid inference when our Assumptions 4-5 are violated in the presence of cluster

heterogeneity.

Setting $d = 3$, $T = 3$, $G = 35$, and $L = L_1 = \dots = L_g \in \{50, 100\}$, the parameters are estimated by the Arellano-Bond moment conditions with the reduced set of instruments, i.e. $m = 6$. It is important to point out that the subsample estimation of $\hat{\theta}_g$ uses only $L_g \in \{50, 100\}$ number of cluster-specific observations. Compared to the full-sample GMM approaches that uses the entire $n \in \{1750, 3500\}$ observations, we conjecture that the IM's sub-sample estimators are exposed to more finite-sample bias, especially when the model is over-identified. This motivates us to also consider an exactly identified moment conditions, i.e. $m = d = 3$, $E[w'_{it-1}\Delta u_{it}] = 0$ with $w_{it} = (y_{it-1}, x'_{it})'$ in Anderson and Hsiao (1981). The Anderson-Hsiao's exactly identified moment condition is used for both one-step GMM (BCH) and sub-sample based t-tests (IM).

Tables 6–7 present empirical sizes, bias, and root mean squared errors (RMSE) of the t-tests we consider at various degree of heterogeneities in Design I–IV. We first note that, when the model is over-identified, IM's sub-sample based approach suffers from severe size distortions in all simulation designs. This is consistent with our previous conjecture—the large finite sample biases in the cluster-specific estimators have negative impacts on corresponding t-test's performances. The full-sample GMM approaches in BCH and Hwang also suffer from size distortions as we move away from the mild degree of heterogeneity in Design I. Table 6 shows that the amount of size distortions in Hwang is much smaller than that of IM. Also, compared to the one-step GMM test in BCH, Hwang's two-step GMM approaches results in smaller size distortions. For example, in Design IV with $L = 50$, the empirical size of Hwang is 10% while those of BCH and IM with over-identifications are 13.2% and 72%, respectively. When the model is exactly identified, however, the size distortions no longer present for IM. For example, Table 6 indicates that the empirical sizes of the exactly identified IM's test are below 5% at all degrees of heterogeneities in Design I–IV. However, we note that this reduction in size distortions comes at the sacrifice of using the exactly-identified instruments (Anderson-Hsiao estimator) instead of the over-identified instruments (Arellano-Bond estimator), which results in a huge loss of efficiency. For example, Table 6 indicates that in Design IV, IM has 3.0% empirical size with corresponding 1.774 of RMSE, while Hwang has 9.3% of empirical size with 0.386 of RMSE. Lastly, we find that the performances of all of the considered t-tests are improved when the size of the clusters L increases.

Overall, the results of our simulations indicate that neither our full-sample GMM tests nor the sub-sample based tests in IM dominate the other one, and they are best described as complementary approaches. If one mainly concerns about the size accuracy of the GMM parameter test under pronounced cluster heterogeneity, the sub-sample based IM's approach is preferred. However, when there is a mild degree of heterogeneities, or we want to increase the accuracy of point-estimators, our full-sample based efficient GMM estimation and corresponding finite-sample corrected t-tests are preferred.

7 Empirical Application

In this section, we employ the proposed procedures to revisit the study of Emran and Hou (2013, The Review of Economics and Statistics) in development economics. The study investigates the causal effects of access to domestic and international markets on rural household consumption. They use a survey data of 7998 rural households from the Chinese Household Income Project (ICPSR 3012) in 1995. The data set is downloadable from the journal website.¹⁰

¹⁰<https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/CWFOFN>

7.1 Model and choice of clusters

The regression equation for per capita consumption for household i , C_i , in 1995 (yuan) is specified as

$$C_i = \beta_0 + \beta_d A_i^d + \beta_s A_i^s + \beta_{ds} (A_i^d \times A_i^s) + X_i' \beta_h + X_i^{v'} \beta_v + X_i^{c'} \beta_c + \alpha_p + \epsilon_i, \quad (32)$$

where A_i^d and A_i^s are the log distances of access to domestic (km) and international markets (km), respectively. X_i is the vector of household characteristics that may affect consumption choice, and X_i^v , X_i^c are village, county level controls, respectively, which capture the heterogeneity in economic environments across different regions, and α_p is the province level fixed effect.

Among the unknown parameters in vector $\theta = (\beta_0, \beta'_m, \beta'_h, \beta'_v, \beta'_c, \alpha)' \in \mathbb{R}^d$, our focus of interest is $\beta_m = (\beta_d, \beta_s, \beta_{ds})'$ which measures the casual effect of access to domestic and international markets on household consumption in the rural areas. To identify these parameters, Emran and Hou (2013) employs geographic instrumental variables that capture exogenous variations in access to markets, e.g., straight-line distances to the nearest navigable river and coastline, along with the topographic and agroclimatic features of the counties.¹¹ There are 21 instrument variables and 31 control variables in their IV regressions, including province dummy variables, so that the number of moment conditions m is 52, and the number of estimated parameters d is 34. The corresponding degree of over-identification q is 18.

Because of the close economic and cultural ties within the same county in rural Chinese areas, the study clusters the data by the county level and estimates the model using 2SLS and two-step GMM with uncentered cluster-robust weighting matrix. The data set consists of $n = 7462$ observations divided into $G = 86$ clusters, where the number of households vary across from a low of 49 to a high of 270 with the average size of cluster $\bar{L} = 87$. Instead of the county-level clustering, one might want to check whether a finer clustering at the level of individuals is innocuous. As for a diagnostic test, we implement a testing for the appropriate level of clustering, which is recently proposed by Ibragimov and Müller (2016). The test considers a practically very plausible scenario that empirical researchers may face – a choice between a small number of coarse clusters, e.g., county level cluster, and a large number of the finer level of clusters, e.g., individual level. The null hypothesis is that a finer level of clustering is appropriate with consistent CCE estimators, against the alternative that the only fewer clusters provide valid information. The suggested test obtains the critical value by a simulation algorithm provided in Ibragimov and Müller (2016). The detailed algorithms for Ibragimov and Müller (2016)'s test are described in supplementary Appendix B.9.

Table 8 reports the results of significance tests of the validity of fine level (individual level) of clustering. Since Ibragimov and Muller (2016)'s test depends on a variable of interest in regression equation (32), we provide corresponding test statistics and critical values for the three key variables A^d , A^s , and A^{ds} . The results in Table 8 indicate that at all three variables of interests, we reject the validity of clustering at the level of the individual at 5% significance, against the coarser clustering at county level. These results might not be too surprising, as rural individuals are closely connected within each county with economic and cultural ties, which, in turn, might well lead to non-trivial interactions.

¹¹For the detailed description of the control variables and instrument variables, see the appendix in Emran and Hou (2013).

7.2 Results

Since the statistical inferences in Emran and Hou (2013) are conducted using the large- G asymptotics only, we apply our more accurate fixed- G asymptotics to their study. The 2SLS (one-step GMM) test together with degrees of freedom correction uses critical value t_{G-1} . For the two-step GMM procedures, we implement the recommended fixed- G test in our paper, which includes the degree-of-freedom correction, the J correction, and the finite sample corrected variance corrections and use the t_{G-1-q} critical value. Table 9 shows the point estimation results, standard error estimates, and associated confidence intervals (CIs) for each of 2SLS and the uncentered and centered two-step GMM estimators. Similar to Emran and Hou (2013), our results show that the better access to domestic and international markets has a substantial positive effect on household consumption, and that the domestic market effect is significantly higher. For the 2SLS method, there are no much differences in confidence interval and standard error between the large- G and fixed- G results. This is well expected because the number of clusters $G = 86$ is large enough so that the large- G and fixed- G approximations are close to each other.

The uncentered two-step GMM estimate of the effect of access to domestic market is $\beta_d = -2722.22$. The reported standard error 400.5 is about 40% smaller than that of 2SLS. However, the plain two-step standard error estimate might be downward biased because the variation of the cluster-robust weighting matrix is not considered. The centered two-step GMM estimator gives a smaller effect of market access $\beta_d = -2670.0$ with the modified standard error of 519.2, which is 25% larger than the plain two-step standard error. However, the modified standard error is still smaller than that based on the 2SLS method. So the two-step estimator still enjoys the benefit of using the cluster-robust weighting matrix. The results for other parameters deliver similar qualitative messages. Table 9 also provides the finite sample corrected standard error estimates of two-step estimators that capture the extra variation of feasible CCE, leading to slightly larger standard errors and wider CIs than the uncorrected ones. Overall, our results suggest that the effect of access to markets may be lower than the previous finding after we take into account the randomness of the estimated optimal GMM weighting matrix.

8 Conclusion

This paper studies GMM estimation and inference under clustered dependence. To obtain more accurate asymptotic approximations, we utilize an alternative asymptotics under which the sample size of each cluster is growing, but the number of cluster size G is fixed. The paper is comprehensive in that it covers the second-step GMM as well as the first-step GMM estimators. For the two-step GMM estimator, we show that only if centered moment processes are used in constructing the weighting matrix can we obtain asymptotically pivotal Wald statistic and t statistic. With the help of the standard J statistic, the Wald statistic and t statistic based on these estimators can be modified to have to standard F and t limiting distributions. A finite sample variance correction is suggested to further improve the performance of the asymptotic t and F tests.

Our simulation evidence and empirical application demonstrate the size accuracy of the two-step GMM Wald and t test statistics coupled with the J statistic modification, degrees-of-freedom correction, and finite-sample corrected variance estimator. The tests are very appealing to practitioners because they can implement the standard t and F critical values to the finite-sample corrected test statistics. No further simulations or re-sampling methods are needed. We recommend practitioners to use these statistics.

References

- [1] Andrews, D. W. (2005). Cross-section regression with common shocks. *Econometrica*, 73(5), 1551-1585.
- [2] Arellano, M. (1987): "Computing Robust Standard Errors for Within-Group Estimators." *Oxford Bulletin of Economics and Statistics*, 49, 431-434.
- [3] Arellano, M. and Bond, S. (1991): "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations." *The Review of Economic Studies*, 58(2), 277-297.
- [4] Bertrand, M., Duflo, E., and Mullainathan, S. (2004): "How Much Should We Trust Differences-in-Differences Estimates?" *The Quarterly Journal of Economics*, 119(1), 249-275.
- [5] Bester, C. A., Conley, T. G., and Hansen, C. B. (2011): "Inference with Dependent Data Using Cluster Covariance Estimators." *Journal of Econometrics* 165(2), 137-151.
- [6] Bilodeau, M., and Brenner, D. (2008): "Theory of multivariate statistics." Springer Science & Business Media.
- [7] Bond, S., and Windmeijer, F. (2005). Reliable inference for GMM estimators? Finite sample properties of alternative test procedures in linear panel data models. *Econometric Reviews*, 24(1), 1-37.
- [8] Blundell, R., and Bond, S. (1998): "Initial conditions and moment restrictions in dynamic panel data models." *Journal of Econometrics*, 87(1), 115-143.
- [9] Bun, M. J., and Windmeijer, F. (2010): "The weak instrument problem of the system GMM estimator in dynamic panel data models." *The Econometrics Journal*, 13(1), 95-126.
- [10] Brown, J. R., Fazzari, S. M., & Petersen, B. C. (2009). "Financing innovation and growth: Cash flow, external equity, and the 1990s R&D boom." *The Journal of Finance*, 64(1), 151-185.
- [11] Canay, I. A., Romano, J. P., and Shaikh, A. M. (2017). Randomization tests under an approximate symmetry assumption. *Econometrica*, 85(3), 1013-1030.
- [12] Canay, I.A., Santos, A. and Shaikh, A.M., (2018). "The wild bootstrap with a 'small' number of 'large' clusters". *Review of Economics and Statistics*, pp.1-45.
- [13] Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008): "Bootstrap-based improvements for inference with clustered errors." *Review of Economics and Statistics*, 90(3), 414-427.
- [14] Cameron, A. C. and Miller, D. L. (2015): "A practitioner's guide to cluster-robust inference." *Journal of Human Resources*, 50(2), 317-372.
- [15] Carter, A.V., Schnepel, K.T. and Steigerwald, D.G., (2017). "Asymptotic behavior of at-test robust to cluster heterogeneity." *Review of Economics and Statistics*, 99(4), pp.698-709.
- [16] Conley, T.G., 1999. "GMM estimation with cross sectional dependence". *Journal of econometrics*, 92(1), pp.1-45.

- [17] Djogbenou, A.A., MacKinnon, J.G. and Nielsen, M.Ø., (2019). “Asymptotic theory and wild bootstrap inference with clustered errors.” *Journal of Econometrics*, 212(2), pp.393-412.
- [18] Emran, M. S., and Hou, Z. (2013): “Access to markets and rural poverty: evidence from household consumption in China.” *Review of Economics and Statistics*, 95(2), 682-697.
- [19] Hagemann, A. (2019). “Placebo inference on treatment effects when the number of clusters is small.” *Journal of Econometrics*, 213(1), 190-209.
- [20] Hall, A. R. (2000): “Covariance matrix estimation and the power of the overidentifying restrictions test.” *Econometrica*, 68(6), 1517-1527.
- [21] Hall, P., and Horowitz, J. L. (1996): “Bootstrap critical values for tests based on generalized-method-of-moments estimators.” *Econometrica: Journal of the Econometric Society*, 891-916.
- [22] Hansen, B.E. and Lee, S., 2019. “Asymptotic theory for clustered samples”. *Journal of econometrics*, 210(2), pp.268-290.
- [23] Hansen, C. B. (2007): “Asymptotic properties of a robust variance matrix estimator for panel data when T is large.” *Journal of Econometrics*, 141(2), 597-620.
- [24] Hansen, L. P. (1982): “Large Sample Properties of Generalized Method of Moments Estimators.” *Econometrica* 50, 1029-1054.
- [25] Hansen, L. P., Heaton, J. and Yaron, A: (1996): “Finite-sample properties of some alternative GMM estimators.” *Journal of Business & Economic Statistics*, 14(3), 262-280.
- [26] Hwang, J., Kang, B. and Lee, S., 2019. A Doubly Corrected Robust Variance Estimator for Linear GMM.
- [27] Hwang, J., and Sun, Y. (2017a): “Asymptotic F and t Tests in an Efficient GMM Setting.” *Journal of Econometrics*, 198(2), 277-295.
- [28] Hwang, J., and Sun, Y. (2017b): “Supplementary Appendix to ‘Asymptotic F and t Tests in an Efficient GMM Setting’.” *Journal of Econometrics Online Supplementary Appendix*, <http://dx.doi.org/10.1016/j.jeconom.2017.02.003>.
- [29] Hwang, J. and Sun, Y., (2018). “Should we go one step further? An accurate comparison of one-step and two-step procedures in a generalized method of moments framework”. *Journal of econometrics*, 207(2), pp.381-405.
- [30] Ibragimov, R., and Müller, U. K. (2010): “t-Statistic based correlation and heterogeneity robust inference.” *Journal of Business & Economic Statistics*, 28(4), 453-468.
- [31] Ibragimov, R. and Müller, U.K. (2016): “Inference with few heterogeneous clusters.” *Review of Economics and Statistics*, 98(1), 83-96.
- [32] Imbens, G. W., and Kolesar, M. (2016): “Robust standard errors in small samples: Some practical advice.” *Review of Economics and Statistics*, 98(4), 701-712.
- [33] Jenish, N., and Prucha, I. R. (2009): “Central limit theorems and uniform laws of large numbers for arrays of random fields.” *Journal of econometrics*, 150(1), 86-98.

- [34] Jenish, N., and Prucha, I. R. (2012): “On spatial processes and asymptotic inference under near-epoch dependence.” *Journal of econometrics*, 170(1), 178-190.
- [35] Kiefer, N. M., Vogelsang, T. J. and Bunzel, H. (2002): “Simple Robust Testing of Regression Hypotheses” *Econometrica* 68 (3), 695-714.
- [36] Kiefer, N. M. and Vogelsang, T. J. (2005): “A New Asymptotic Theory for Heteroskedasticity-Autocorrelation Robust Tests.” *Econometric Theory* 21, 1130-1164.
- [37] Liang, K.-Y. and Zeger, S. (1986): “Longitudinal Data Analysis Using Generalized Linear Models.” *Biometrika* 73(1):13-22.
- [38] MacKinnon, J. G., and Webb M. D. (2017): “Wild bootstrap inference for wildly different cluster sizes.” *Journal of Applied Econometrics* 32, 233-254.
- [39] Müller, U. K. (2007): “A theory of robust long-run variance estimation.” *Journal of Econometrics*, 141(2), 1331-1352.
- [40] Newey, W. K., and Smith, R. J. (2004): “Higher order properties of GMM and generalized empirical likelihood estimators.” *Econometrica* 72, no. 1 : 219-255.
- [41] Oberholzer-Gee, F., & Strumpf, K. (2007). “The effect of file sharing on record sales: An empirical analysis.” *Journal of political economy*, 115(1), 1-42.
- [42] Roodman, D. (2009). “How to do xtabond2: An introduction to difference and system GMM in Stata.” *The stata journal*, 9(1), 86-136.
- [43] Stock, J. H., and Wright, J. H. (2000): “GMM with weak identification. *Econometrica*.”, 68(5), 1055-1096.
- [44] Stock, J. H. and Watson, M. W. (2008): “Heteroskedasticity-Robust Standard Errors for Fixed Effects Panel Data Regression.” *Econometrica*, 76: 155-174.
- [45] Sun, Y. (2013): “A Heteroskedasticity and Autocorrelation Robust F Test Using Orthonormal Series Variance Estimator.” *Econometrics Journal* 16(1), 1–26.
- [46] Sun, Y. (2014): “Fixed-smoothing Asymptotics in a Two-step GMM Framework.” *Econometrica* 82(6), 2327-2370.
- [47] Sun, Y., and Kim, M. S. (2012): “Simple and powerful GMM over-identification tests with accurate size.” *Journal of Econometrics*, 166(2), 267-281.
- [48] Sun, Y., and Kim, M. S. (2015): “Asymptotic F-Test in a GMM Framework with Cross-Sectional Dependence.” *Review of Economics and Statistics*, 97(1), 210-223.
- [49] Sun, Y., Phillips, P. C. B. and Jin, S. (2008): “Optimal Bandwidth Selection in Heteroskedasticity-Autocorrelation Robust Testing.” *Econometrica* 76(1), 175–94.
- [50] White, H. (1984), “Asymptotic Theory for Econometricians” (San Diego: Academic Press).
- [51] Windmeijer, F. (2005): “A finite sample correction for the variance of linear efficient two-step GMM estimators.” *Journal of Econometrics*, 126(1), 25-51.
- [52] Wooldridge, J. M. (2003): “Cluster-sample methods in applied econometrics.” *American Economic Review*, 133-138.

Table 1: Empirical size of GMM tests based on the centered CCE when the number of clusters $G = 50, 100$, the number of population within cluster $L = 50$, the number of joint hypothesis $p = 1 \sim 3$, and the number of moment conditions $m = 12, 24$, with $T = 4$.

$G = 50$	Test statistic	Critical values	$m = 24$			$m = 12$		
			$p = 1$	$p = 2$	$p = 3$	$p = 1$	$p = 2$	$p = 3$
First-step	F_1	$\chi_p^{1-\alpha}/p$	0.240	0.247	0.242	0.188	0.178	0.176
	$\frac{G-p}{G}F_1$	$F_{p,G-p}^{1-\alpha}$	0.220	0.212	0.191	0.171	0.145	0.133
Two-step	F_2	$\chi_p^{1-\alpha}/p$	0.320	0.439	0.535	0.137	0.171	0.202
	\tilde{F}_2	$F_{p,G-p-q}^{1-\alpha}$	0.073	0.063	0.061	0.064	0.059	0.058
	\tilde{F}_2^w	$F_{p,G-p-q}^{1-\alpha}$	0.063	0.054	0.053	0.053	0.050	0.049
	F_2	HH-Bootstrap	0.020	0.004	0.002	0.048	0.043	0.033
J test	J^c	$\chi_q^{1-\alpha}/q$	—	0.566	—	—	0.150	—
	$\frac{G-q}{G}J^c$	$F_{q,G-q}^{1-\alpha}$	—	0.059	—	—	0.055	—
	J^c	HH-Bootstrap	—	0.296	—	—	0.116	—

$G = 100$	Test statistics	Critical values	$m = 24$			$m = 12$		
			$p = 1$	$p = 2$	$p = 3$	$p = 1$	$p = 2$	$p = 3$
First-step	F_1	$\chi_p^{1-\alpha}/p$	0.167	0.158	0.154	0.150	0.130	0.134
	$\frac{G-p}{G}F_1$	$F_{p,G-p}^{1-\alpha}$	0.159	0.145	0.132	0.142	0.118	0.114
Two-step	F_2	$\chi_p^{1-\alpha}/p$	0.167	0.201	0.237	0.094	0.105	0.120
	\tilde{F}_2	$F_{p,G-p-q}^{1-\alpha}$	0.074	0.067	0.066	0.066	0.059	0.060
	\tilde{F}_2^w	$F_{p,G-p-q}^{1-\alpha}$	0.070	0.061	0.059	0.058	0.051	0.055
	F_2	HH-Bootstrap	0.051	0.042	0.035	0.053	0.048	0.051
J test	J^c	$\chi_q^{1-\alpha}/q$	—	0.256	—	—	0.098	—
	$\frac{G-q}{G}J^c$	$F_{q,G-q}^{1-\alpha}$	—	0.057	—	—	0.056	—
	J^c	HH-Bootstrap	—	0.197	—	—	0.091	—

Notes: The first-step tests are based on the first-step GMM estimator $\hat{\theta}_1$ with the associated F statistic $F_1 = F_1(\hat{\theta}_1)$. The J tests employ the statistics $J^c = J(\hat{\theta}_2^c)$ with or without degree of freedom (d.f.) correction. All two-step tests are based on the centered two-step GMM estimator $\hat{\theta}_2$ but use different test statistics: the unmodified $F_2 = F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)$, J statistic and d.f. corrected $\tilde{F}_2 = \tilde{F}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)$, and J statistic, d.f., and finite-sample-variance corrected $\tilde{F}_2^w = \tilde{F}_{\hat{\Omega}^c(\hat{\theta}_1)}^w(\hat{\theta}_2^c)$.

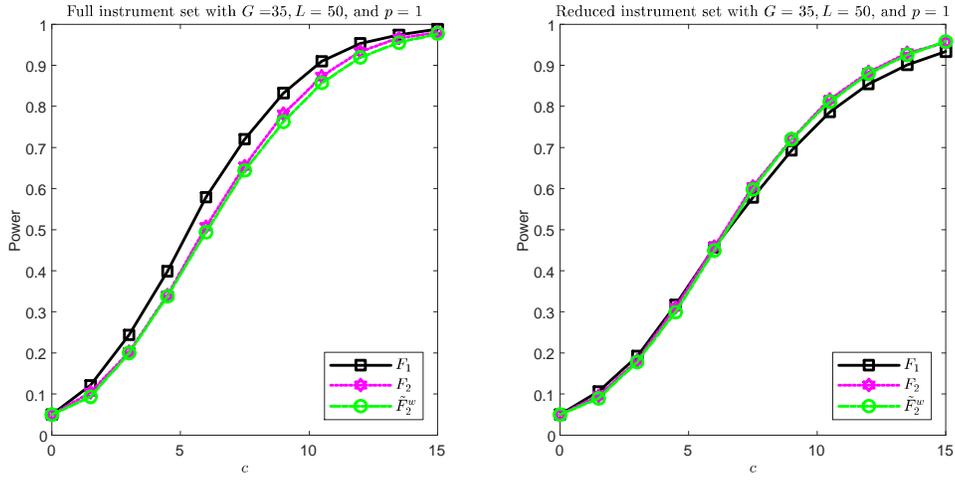


Figure 1: Size-adjusted power of the first-step (2SLS) and two-step tests with $G=35$ and $L=50$.

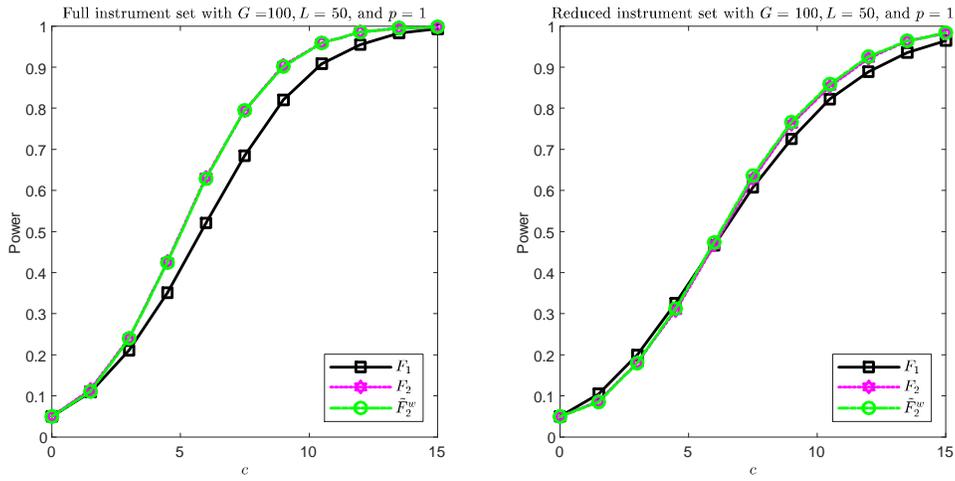


Figure 2: Size-adjusted power of the first-step (2SLS) and two-step tests with $G = 100$ and $L = 50$.

Table 2: Design of unbalanced cluster size

$G = 50$	$L_1 = \dots = L_{10}$	$L_{11} = \dots = L_{50}$	n
Design I	120	95	5000
Design II	160	85	5000
Design III	200	75	5000
Design IV	240	65	5000

Table 3: Empirical size of first-step and two-step tests based on the centered CCE when there is a heterogeneity in cluster size: Design I

Unbalanced sizes in clusters: Design I								
	Test statistic	Critical values	$m = 24$			$m = 12$		
			$p = 1$	$p = 2$	$p = 3$	$p = 1$	$p = 2$	$p = 3$
First-step	F_1	$\chi_p^{1-\alpha}/p$	0.169	0.170	0.170	0.152	0.142	0.152
	$\frac{G-p}{G} F_1$	$\mathcal{F}_{p,G-p}^{1-\alpha}$	0.155	0.140	0.125	0.137	0.117	0.109
Two-step	F_2	$\chi_p^{1-\alpha}/p$	0.307	0.441	0.545	0.139	0.173	0.210
	\tilde{F}_2	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.071	0.069	0.070	0.065	0.062	0.066
	\tilde{F}_2^w	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.054	0.051	0.051	0.049	0.051	0.051
	F_2	HH-Bootstrap	0.015	0.005	0.000	0.042	0.033	0.032
J test	J^c	$\chi_q^{1-\alpha}/q$	–	0.571	–	–	0.157	–
	$\frac{G-q}{G_q} J^c$	$\mathcal{F}_{q,G-q}^{1-\alpha}$	–	0.053	–	–	0.056	–
	J^c	HH-Bootstrap	–	0.275	–	–	0.102	–

See footnote to Table 1.

Table 4: Empirical size of first-step and two-step tests based on the centered CCE when there is a heterogeneity in cluster size: Designs II and III

Unbalanced sizes in clusters: Design II								
	Test statistic	Critical values	$m = 24$			$m = 12$		
			$p = 1$	$p = 2$	$p = 3$	$p = 1$	$p = 2$	$p = 3$
First-step	F_1	$\chi_p^{1-\alpha}/p$	0.172	0.168	0.172	0.154	0.144	0.151
	$\frac{G-p}{G}F_1$	$\mathcal{F}_{p,G-p}^{1-\alpha}$	0.157	0.137	0.124	0.138	0.116	0.111
Two-step	F_2	$\chi_p^{1-\alpha}/p$	0.304	0.394	0.482	0.147	0.173	0.211
	\tilde{F}_2	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.076	0.075	0.071	0.069	0.069	0.068
	\tilde{F}_2^w	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.057	0.054	0.049	0.056	0.052	0.054
	F_2	HH-Bootstrap	0.015	0.003	0.000	0.040	0.034	0.030
J test	J^c	$\chi_q^{1-\alpha}/q$	–	0.593	–	–	0.163	–
	$\frac{G-q}{Gq}J^c$	$\mathcal{F}_{q,G-q}^{1-\alpha}$	–	0.068	–	–	0.058	–
	J^c	HH-Bootstrap	–	0.292	–	–	0.102	–

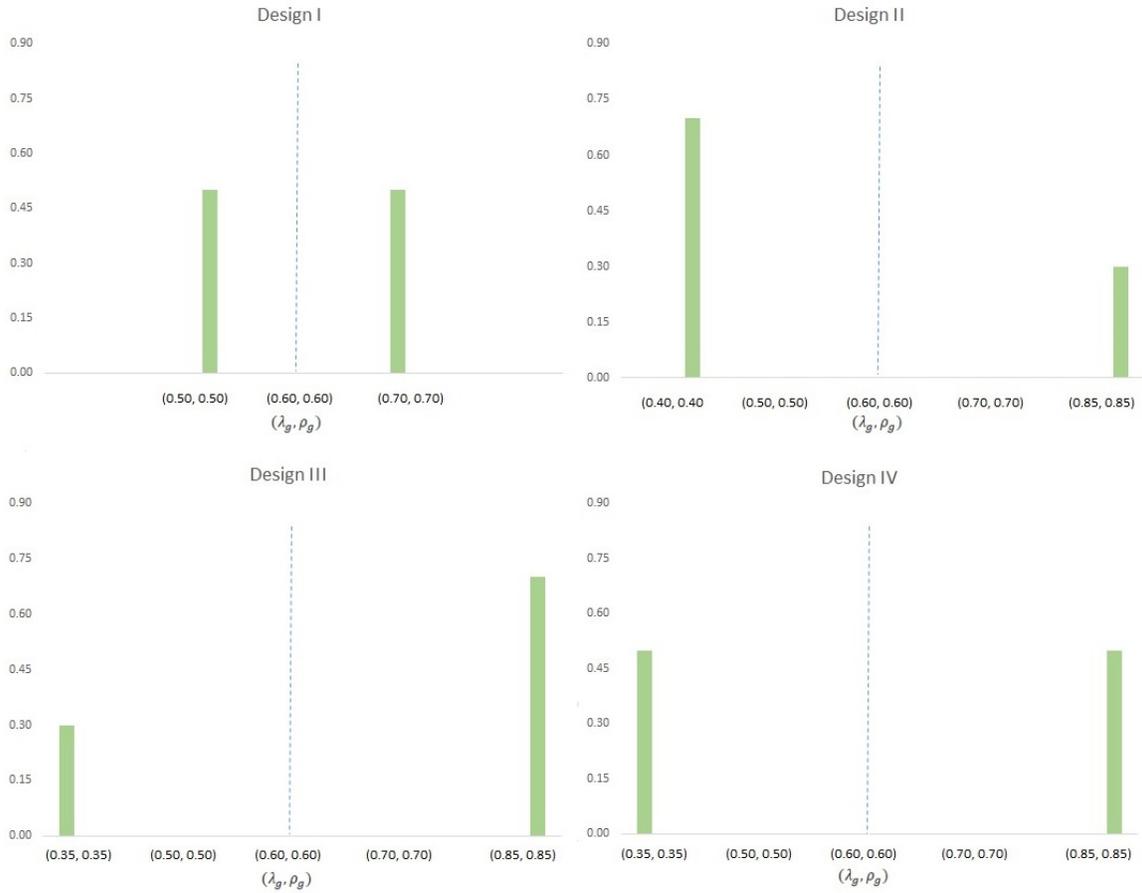
Unbalanced sizes in clusters: Design III								
	Test statistic	Critical values	$m = 24$			$m = 12$		
			$p = 1$	$p = 2$	$p = 3$	$p = 1$	$p = 2$	$p = 3$
First-step	F_1	$\chi_p^{1-\alpha}/p$	0.173	0.174	0.180	0.154	0.148	0.158
	$\frac{G-p}{G}F_1$	$\mathcal{F}_{p,G-p}^{1-\alpha}$	0.160	0.140	0.133	0.141	0.122	0.117
Two-step	F_2	$\chi_p^{1-\alpha}/p$	0.336	0.469	0.586	0.153	0.193	0.242
	\tilde{F}_2	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.084	0.081	0.084	0.072	0.071	0.075
	\tilde{F}_2^w	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.060	0.057	0.058	0.057	0.054	0.057
	F_2	HH-Bootstrap	0.013	0.003	0.000	0.040	0.033	0.026
J test	J^c	$\chi_q^{1-\alpha}/q$	–	0.640	–	–	0.182	–
	$\frac{G-q}{Gq}J^c$	$\mathcal{F}_{q,G-q}^{1-\alpha}$	–	0.085	–	–	0.070	–
	J^c	HH-Bootstrap	–	0.328	–	–	0.110	–

See footnote to Table 1.

Table 5: Empirical size of first-step and two-step tests based on the centered CCE when there is a heterogeneity in cluster size: Design IV

			Unbalanced sizes in clusters: Design IV					
			$p = 1$	$p = 2$	$p = 3$	$p = 1$	$p = 2$	$p = 3$
First-step	F_1	$\chi_p^{1-\alpha}/p$	0.178	0.180	0.188	0.158	0.152	0.169
	$\frac{G-p}{G} F_1$	$\mathcal{F}_{p,G-p}^{1-\alpha}$	0.160	0.148	0.142	0.142	0.128	0.123
Two-step	F_2	$\chi_p^{1-\alpha}/p$	0.360	0.511	0.621	0.165	0.218	0.271
	\tilde{F}_2	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.091	0.095	0.100	0.082	0.086	0.099
	\tilde{F}_2^w	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.071	0.072	0.074	0.067	0.069	0.076
	F_2	HH-Bootstrap	0.011	0.003	0.001	0.041	0.034	0.030
J test	J^c	$\chi_q^{1-\alpha}/q$	—	0.709	—	—	0.206	—
	$\frac{G-q}{Gq} J^c$	$\mathcal{F}_{q,G-q}^{1-\alpha}$	—	0.131	—	—	0.084	—
	J^c	HH-Bootstrap	—	0.380	—	—	0.111	—

See footnote to Table 1.



Design of heterogenous clusters: Probability distributions for heterogeneous (ρ_g, λ_g)

Table 6: Empirical size, bias, RMSE of GMM t-tests for heterogeneous clusters with $G = 35$ and $L = 50$

$d = 3$ and $T = 3$				
Heterogeneous Clusters: Design I				
	Tests	Bias	RMSE	Size
Exactly-identified GMM ($m = 3$)	BCH	-0.080	1.283	0.075
	IM	-0.091	1.542	0.035
Over-identified GMM ($m = 6$)	BCH	-0.075	0.180	0.084
	IM	-0.236	0.249	0.829
	Hwang	-0.037	0.164	0.046
Heterogeneous Clusters: Design II				
	Tests	Bias	RMSE	Size
Exactly-identified GMM ($m = 3$)	BCH	-0.096	1.149	0.059
	IM	-0.139	1.567	0.031
Over-identified GMM ($m = 6$)	BCH	-0.194	0.352	0.130
	IM	-0.217	0.233	0.846
	Hwang	-0.146	0.346	0.102
Heterogeneous Clusters: Design III				
	Tests	Bias	RMSE	Size
Exactly-identified GMM ($m = 3$)	BCH	-0.075	1.175	0.039
	IM	-0.089	1.774	0.030
Over-identified GMM ($m = 6$)	BCH	-0.231	0.392	0.123
	IM	-0.246	0.265	0.713
	Hwang	-0.186	0.386	0.093
Heterogeneous Clusters: Design IV				
	Tests	Bias	RMSE	Size
Exactly-identified GMM ($m = 3$)	BCH	-0.080	1.217	0.042
	IM	-0.120	1.690	0.031
Over-identified GMM ($m = 6$)	BCH	-0.224	0.385	0.132
	IM	-0.229	0.246	0.717
	Hwang	-0.177	0.380	0.100

Notes: BCH's tests are based on the first-step GMM estimator $\hat{\theta}_1$ with the associated t statistic $t_1 = t_1(\hat{\theta}_1)$ and critical value t_{G-1} . IM's tests are based on t -statistics of cluster specific estimators $\hat{\theta}_g$ and critical value t_{G-1} . Hwang's tests are based on the centered two-step GMM estimator $\hat{\theta}_2^c$ with finite-sample-variance corrected $\tilde{t}_2^v = \tilde{t}_{\Omega^c(\hat{\theta}_1)}^w(\hat{\theta}_2^c)$ and critical value t_{G-q-1} . The exactly-identified moment condition is constructed using Anderson and Hsiao (1981)'s instrument variables. The over-identified moment condition is constructed from Arellano-Bond (1991)'s instruments using the most recent lag.

Table 7: Empirical size, bias, RMSE of GMM t-tests for heterogeneous clusters with $G = 35$ and $L = 100$

$d = 3$ and $T = 3$				
Heterogeneous Clusters: Design I				
	Tests	Bias	RMSE	Size
Exactly-identified GMM ($m = 3$)	BCH	-0.131	1.357	0.098
	IM	-0.107	1.380	0.032
Over-identified GMM ($m = 6$)	BCH	-0.038	0.117	0.073
	IM	-0.228	0.238	0.892
	Hwang	-0.014	0.109	0.044
Heterogeneous Clusters: Design II				
	Tests	Bias	RMSE	Size
Exactly-identified GMM ($m = 3$)	BCH	-0.067	1.170	0.069
	IM	-0.150	1.603	0.034
Over-identified GMM ($m = 6$)	BCH	-0.134	0.276	0.122
	IM	-0.206	0.219	0.797
	Hwang	-0.086	0.268	0.084
Heterogeneous Clusters: Design III				
	Tests	Bias	RMSE	Size
Exactly-identified GMM ($m = 3$)	BCH	-0.093	1.222	0.049
	IM	-0.116	1.835	0.025
Over-identified GMM ($m = 6$)	BCH	-0.179	0.322	0.113
	IM	-0.245	0.261	0.793
	Hwang	-0.130	0.312	0.085
Heterogeneous Clusters: Design IV				
	Tests	Bias	RMSE	Size
Exactly-identified GMM ($m = 3$)	BCH	-0.099	1.213	0.055
	IM	-0.116	1.765	0.028
Over-identified GMM ($m = 6$)	BCH	-0.165	0.314	0.115
	IM	-0.222	0.237	0.783
	Hwang	-0.118	0.303	0.087

Notes: See footnote to Table 6.

Table 8: Test of the appropriate level of clustering

Test of validity of fine clustering			
H_0 : Both individual (fine) and county (coarse) clusterings are valid.			
H_1 : Only county (coarse) clustering is valid.			
	Variable of interests		
	A^d	A^s	A^{ds}
Test statistic	504,156.2	429,620.7	12,225.1
5% Critical value	138,547.3	109,723.0	1,288.7
p-value	0.00%	0.00%	0.00%

Notes: The test statistics, corresponding critical values, and p-values are constructed for each variable of interests in (32). For all considered tests, the test statistics and critical values are calculated by the Ibragimov and Müller (2016)'s method.

Table 9: Results for Emran and Hou (2013) data

2SLS		
Variables	Large- G asymptotics	fixed- G asymptotics
Domestic market (A_i^d)	-2713.2 (712.1) [-4109.9, -1316.4]	-2713.2 (716.8) [-4138.0, -1288.0]
International market (A_i^s)	-1993.5 (514.8) [-3002.5, -984.4]	-1993.5 (517.9) [-3023.10, -963.8]
Interaction ($A_i^d \times A_i^s$)	345.8 (105.0) [140.0, 551.7]	345.8 (105.6) [135.8, 555.9]
$H_0 : \beta_d = \beta_s$	-2.3218 (2.02%)	-2.771 (2.26%)
Two-step GMM		
Variables	Large- G asymptotics	fixed- G asymptotics
Domestic market (A_i^d)	-2722.8 (400.5) [-3507.7, -1937.9]	-2670.0 (520.7) [-3709.2, -1630.7]
International market (A_i^s)	-2000.2 (344.3) [-2675.0, -1325.5]	-1981.3 (447.7) [-2874.9, -1087.7]
Interaction ($A_i^d \times A_i^s$)	362.7 (68.7) [228.0, 497.3]	364.1 (89.4) [187.5, 542.4]
$H_0 : \beta_d = \beta_s$	-5.239 (0%)	-3.3217 (0%)
J statistic ($q = 18$)	1.1708 (99.8%)	0.3096 (45.83%)

Notes: Standard errors for 2SLS and the weighting matrix for (centered) two-step GMM estimators are clustered at the county level. Numbers in parentheses are standard errors and intervals are 95% confidence intervals. For hypothesis testing, the numbers in parentheses are p-values.

Appendix of A: Proofs of main results

Proof of Proposition 1. Part (a). For each $g = 1, \dots, G$,

$$\frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} f_i^g(\hat{\theta}_1) = \frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} \left\{ f_i^g(\theta_0) + \frac{\partial f_i^g(\tilde{\theta}^*)}{\partial \theta'} (\hat{\theta}_1 - \theta_0) \right\},$$

where $\tilde{\theta}^*$ is between $\hat{\theta}_1$ and θ_0 . Here, $\tilde{\theta}^*$ may be different for different rows of $\partial f_i^g(\tilde{\theta}^*)/\partial \theta'$. For notational simplicity, we do not make this explicit. Also, standard asymptotic arguments give

$$\begin{aligned} \hat{\theta}_1 - \theta_0 &= (\Gamma'W^{-1}\Gamma)^{-1}\Gamma'W^{-1}g_n(\theta_0) + o_p\left(\frac{1}{\sqrt{n}}\right) \\ &= (\Gamma'W^{-1}\Gamma)^{-1}\Gamma'W^{-1}\frac{1}{G}\sum_{g=1}^G\frac{1}{L}\left(\sum_{i=1}^{L_g}f_i^g(\theta_0)\right) + o_p\left(\frac{1}{\sqrt{L}}\right), \end{aligned}$$

as $n \rightarrow \infty$ holding G fixed. Combining these results together, we have

$$\begin{aligned} \frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} f_i^g(\hat{\theta}) &= \frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} f_i^g(\theta_0) - \frac{1}{L} \sum_{k=1}^{L_g} \frac{\partial f_k^g(\tilde{\theta}^*)}{\partial \theta'} (\Gamma'W^{-1}\Gamma)^{-1}\Gamma'W^{-1} \frac{1}{G} \sum_{g=1}^G \left(\frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} f_i^g(\theta_0) \right) + o_p(1) \\ &= \sqrt{\frac{L_g}{L}} \cdot \frac{1}{\sqrt{L_g}} \sum_{i=1}^{L_g} f_i^g(\theta_0) - \frac{L_g}{L} \left(\frac{1}{L_g} \sum_{i=1}^{L_g} \frac{\partial f_i^g(\tilde{\theta}^*)}{\partial \theta'} \right) (\Gamma'W^{-1}\Gamma)^{-1}\Gamma'W^{-1} \\ &\quad \times \frac{1}{G} \sum_{\tilde{g}=1}^G \left(\sqrt{\frac{L_g}{L}} \cdot \frac{1}{\sqrt{L_g}} \sum_{i=1}^{L_g} f_i^g(\theta_0) \right) + o_p(1) \\ &= \frac{1}{\sqrt{L_g}} \sum_{i=1}^{L_g} f_i^g(\theta_0) - \Gamma_g(\Gamma'W^{-1}\Gamma)^{-1}\Gamma'W^{-1} \frac{1}{G} \sum_{g=1}^G \left(\frac{1}{\sqrt{L_g}} \sum_{i=1}^{L_g} f_i^g(\theta_0) \right) + o_p(1), \end{aligned} \tag{33}$$

where the last equality follows by Assumption 1-iii) and Assumption 3. Using Assumption 1-ii) and Assumption 5, we then obtain

$$\begin{aligned} \frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} f_i^g(\hat{\theta}_1) &\xrightarrow{d} \Lambda B_{m,g} - \Gamma_g(\Gamma'W^{-1}\Gamma)^{-1}\Gamma'W^{-1}\Lambda \bar{B}_m \\ &= \Lambda B_{m,g} - \Gamma(\Gamma'W^{-1}\Gamma)^{-1}\Gamma'W^{-1}\Lambda \bar{B}_m, \end{aligned}$$

where $\bar{B}_m := G^{-1} \sum_{g=1}^G B_{m,g}$. It follows that

$$\begin{aligned} \hat{\Gamma}(\hat{\theta}_1)'W_n^{-1} \frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} f_i^g(\hat{\theta}_1) &\xrightarrow{d} \Gamma'W^{-1} [\Lambda B_{m,g} - \Gamma(\Gamma'W^{-1}\Gamma)^{-1}\Gamma'W^{-1}\Lambda \bar{B}_m] \\ &= \Gamma'W^{-1}\Lambda B_{m,g} - \Gamma'W^{-1}\Lambda \bar{B}_m = \Gamma'W^{-1}\Lambda (B_{m,g} - \bar{B}_m). \end{aligned}$$

Thus, the scaled CCE matrix converges in distribution to a random matrix:

$$\begin{aligned}
& \hat{\Gamma}(\hat{\theta}_1)' W_n^{-1} \hat{\Omega}(\hat{\theta}_1) W_n^{-1} \hat{\Gamma}(\hat{\theta}_1) \\
&= \frac{1}{G} \sum_{g=1}^G \left[\hat{\Gamma}(\hat{\theta}_1)' W_n^{-1} \left(\frac{1}{\sqrt{L}} \sum_{k=1}^{L_g} f_k^g(\hat{\theta}_1) \right) \left(\frac{1}{\sqrt{L}} \sum_{k=1}^{L_g} f_k^g(\hat{\theta}_1) \right)' W_n^{-1} \hat{\Gamma}(\hat{\theta}_1) \right] \\
&\xrightarrow{d} \Gamma' W^{-1} \Lambda \left\{ \frac{1}{G} \sum_{g=1}^G (B_{m,g} - \bar{B}_m) (B_{m,g} - \bar{B}_m)' \right\} (\Gamma' W^{-1} \Lambda)'.
\end{aligned}$$

Therefore,

$$\begin{aligned}
n \cdot R \widehat{\text{var}}(\hat{\theta}_1) R' &= R \left[\hat{\Gamma}(\hat{\theta}_1)' W_n^{-1} \hat{\Gamma}(\hat{\theta}_1) \right]^{-1} \left[\hat{\Gamma}(\hat{\theta}_1)' W_n^{-1} \hat{\Omega}(\hat{\theta}_1) W_n^{-1} \hat{\Gamma}(\hat{\theta}_1) \right] \left[\hat{\Gamma}(\hat{\theta}_1)' W_n^{-1} \hat{\Gamma}(\hat{\theta}_1) \right]^{-1} R' \\
&= R \left[\Gamma' W^{-1} \Gamma \right]^{-1} \Gamma' W^{-1} \Lambda \left\{ \frac{1}{G} \sum_{g=1}^G (B_{m,g} - \bar{B}_m) (B_{m,g} - \bar{B}_m)' \right\} \Lambda W^{-1} \Gamma \left[\Gamma' W^{-1} \Gamma \right]^{-1} R' + o_p(1) \\
&= \tilde{R} \left\{ \frac{1}{G} \sum_{g=1}^G (B_{m,g} - \bar{B}_m) (B_{m,g} - \bar{B}_m)' \right\} \tilde{R}' + o_p(1),
\end{aligned}$$

where $\tilde{R} := R \left[\Gamma' W^{-1} \Gamma \right]^{-1} \Gamma' W^{-1} \Lambda$. Also, it follows by Assumption 1-ii) and iii) that

$$\begin{aligned}
\sqrt{n}(R\hat{\theta}_1 - r) &= -R(\Gamma' W^{-1} \Gamma)^{-1} \Gamma' W^{-1} \sqrt{n} g_n(\theta_0) + o_p(1) \\
&= -R(\Gamma' W^{-1} \Gamma)^{-1} \Gamma' W^{-1} \frac{1}{\sqrt{G}} \sum_{g=1}^G \left(\frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} f_i^g(\theta_0) \right) + o_p(1) \\
&\xrightarrow{d} -\tilde{R} \frac{1}{\sqrt{G}} \sum_{g=1}^G B_{m,g} = -\tilde{R} \sqrt{G} \bar{B}_m.
\end{aligned}$$

Combining the results so far yields:

$$F(\hat{\theta}_1) \xrightarrow{d} \frac{1}{p} \left(\tilde{R} \sqrt{G} \bar{B}_m \right)' \left\{ \tilde{R} \frac{1}{G} \sum_{g=1}^G (B_{m,g} - \bar{B}_m) (B_{m,g} - \bar{B}_m)' \tilde{R}' \right\}^{-1} \tilde{R} \sqrt{G} \bar{B}_m = \mathbb{F}_{1\infty}.$$

Define the $p \times p$ matrix $\tilde{\Lambda}$ such that $\tilde{\Lambda} \tilde{\Lambda}' = \tilde{R} \tilde{R}'$. Then we have the following distributional equivalence

$$\left[\tilde{R} \sqrt{G} \bar{B}_m, \tilde{R} G^{-1} \sum_{g=1}^G (B_{m,g} - \bar{B}_m) (B_{m,g} - \bar{B}_m)' \tilde{R}' \right] \stackrel{d}{=} \left[\sqrt{G} \tilde{\Lambda} \bar{B}_p, \tilde{\Lambda} \tilde{\mathbb{S}}_{pp} \tilde{\Lambda}' \right].$$

Using this, we get

$$\mathbb{F}_{1\infty} \stackrel{d}{=} \frac{G}{p} \cdot \bar{B}_p' \tilde{\mathbb{S}}_{pp}^{-1} \bar{B}_p$$

as desired for Part (a). Part (b) can be similarly proved. ■

Proof of Lemma 7. The centered CCE $\hat{\Omega}^c(\tilde{\theta})$ can be represented as:

$$\begin{aligned} \hat{\Omega}^c(\tilde{\theta}) &= \frac{1}{G} \sum_{g=1}^G \left\{ \frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} \left(f_i^g(\tilde{\theta}) - \frac{1}{n} \sum_{g=1}^G \sum_{i=1}^{L_g} f_i^g(\tilde{\theta}) \right) \right. \\ &\quad \left. \times \frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} \left(f_i^g(\tilde{\theta}) - \frac{1}{n} \sum_{g=1}^G \sum_{i=1}^{L_g} f_i^g(\tilde{\theta}) \right)' \right\}. \end{aligned}$$

To prove Part (a), it suffices to show that

$$\frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} \left(f_i^g(\tilde{\theta}) - \frac{1}{n} \sum_{g=1}^G \sum_{i=1}^{L_g} f_i^g(\tilde{\theta}) \right) = \frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} \left(f_i^g(\theta_0) - \frac{1}{n} \sum_{g=1}^G \sum_{i=1}^{L_g} f_i^g(\theta_0) \right) (1 + o_p(1)) \quad (34)$$

holds for each $g = 1, \dots, G$. By Assumption 3 and Taylor expansion, we have

$$\frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} f_i^g(\tilde{\theta}) = (1 + o_p(1)) \left(\frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} f_i^g(\theta_0) + \frac{1}{L} \sum_{i=1}^{L_g} \frac{\partial f_i^g(\tilde{\theta})}{\partial \theta'} \sqrt{L}(\tilde{\theta} - \theta_0) \right).$$

Using $\sqrt{n}(\tilde{\theta} - \theta_0) = O_p(1)$, and Assumptions 1-iii) and 4, we have

$$\frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} f_i^g(\tilde{\theta}) = (1 + o_p(1)) \left(\frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} f_i^g(\theta_0) + \Gamma \sqrt{L}(\tilde{\theta} - \theta_0) \right)$$

for each $g = 1, \dots, G$. It then follows that

$$\begin{aligned} \frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} \left(f_i^g(\tilde{\theta}) - \frac{1}{n} \sum_{g=1}^G \sum_{i=1}^{L_g} f_i^g(\tilde{\theta}) \right) &= \frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} f_i^g(\tilde{\theta}) - \frac{L_g}{n} \sum_{g=1}^G \left(\frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} f_i^g(\tilde{\theta}) \right) \\ &= \frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} f_i^g(\tilde{\theta}) - \frac{1}{G} \sum_{g=1}^G \left(\frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} f_i^g(\tilde{\theta}) \right) (1 + o_p(1)) \\ &= \left(\frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} f_i^g(\theta_0) - \frac{1}{G} \sum_{g=1}^G \frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} f_i^g(\theta_0) \right) (1 + o_p(1)), \end{aligned}$$

which completes the proof of part (a).

To prove Part (b), we apply the CLT in Assumption 1-ii) together with Assumptions Assumption 1-iii) and 5 to obtain:

$$\begin{aligned} &\frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} f_i^g(\theta_0) - \frac{1}{G} \sum_{g=1}^G \frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} f_i^g(\theta_0) \\ &= \left(\frac{1}{\sqrt{L_g}} \sum_{i=1}^{L_g} f_i^g(\theta_0) - \frac{1}{G} \sum_{g=1}^G \frac{1}{\sqrt{L_g}} \sum_{i=1}^{L_g} f_i^g(\theta_0) \right) (1 + o_p(1)) \\ &\xrightarrow{d} \Lambda(B_{m,g} - \bar{B}_m), \end{aligned}$$

where the convergence holds jointly for $g = 1, \dots, G$. As a result,

$$\hat{\Omega}^c(\theta_0) \xrightarrow{d} \frac{1}{G} \Lambda \left(\sum_{g=1}^G (B_{m,g} - \bar{B}_m) (B_{m,g} - \bar{B}_m)' \right) \Lambda'.$$

■

Proof of Theorem 10. Define $\mathbf{B}'_q = (B'_{q,1}, \dots, B'_{q,G})'$ and denote

$$v_g = (B_{q,g} - \bar{B}_q)' \left[\sum_{g=1}^G (B_{q,g} - \bar{B}_q) (B_{q,g} - \bar{B}_q)' \right]^{-1} \bar{B}_q.$$

Then, the distribution of $\sqrt{G} \bar{\mathbb{S}}_{pq} \bar{\mathbb{S}}_{qq}^{-1} \bar{B}_q$ conditional on \mathbf{B}_q can be represented as

$$\begin{aligned} & \sqrt{G} \left(\sum_{g=1}^G (B_{p,g} - \bar{B}_p) (B_{q,g} - \bar{B}_q)' \right) \left(\sum_{g=1}^G (B_{q,g} - \bar{B}_q) (B_{q,g} - \bar{B}_q)' \right)^{-1} \bar{B}_q \\ &= \sqrt{G} \sum_{g=1}^G (B_{p,g} - \bar{B}_p) v_g = \sqrt{G} \sum_{g=1}^G B_{p,g} v_g - \sqrt{G} \bar{B}_p \sum_{g=1}^G v_g \\ &\stackrel{d}{=} N \left(0, G \sum_{g=1}^G v_g^2 \cdot I_p \right), \end{aligned}$$

where the last line holds because $\sum_{g=1}^G v_g = 0$. Note that

$$\begin{aligned} G \sum_{g=1}^G v_g^2 &= G \sum_{g=1}^G \left\{ (B_{q,g} - \bar{B}_q)' \left[\sum_{g=1}^G (B_{q,g} - \bar{B}_q) (B_{q,g} - \bar{B}_q)' \right]^{-1} \bar{B}_q \right. \\ &\quad \left. \times \bar{B}'_q \left[\sum_{g=1}^G (B_{q,g} - \bar{B}_q) (B_{q,g} - \bar{B}_q)' \right]^{-1} (B_{q,g} - \bar{B}_q) \right\} \\ &= G \bar{B}'_q \left[\sum_{g=1}^G (B_{q,g} - \bar{B}_q) (B_{q,g} - \bar{B}_q)' \right]^{-1} \left[\sum_{g=1}^G (B_{q,g} - \bar{B}_q) \right. \\ &\quad \left. \times (B_{q,g} - \bar{B}_q)' \right] \left[\sum_{g=1}^G (B_{q,g} - \bar{B}_q) (B_{q,g} - \bar{B}_q)' \right] \bar{B}_q \\ &= \bar{B}'_q \left[\sum_{g=1}^G (B_{q,g} - \bar{B}_q) (B_{q,g} - \bar{B}_q)' / G \right]^{-1} \bar{B}_q \\ &= \bar{B}'_q \bar{\mathbb{S}}_{qq}^{-1} \bar{B}_q. \end{aligned}$$

So conditional on \mathbf{B}_q , $\sqrt{G} \bar{\mathbb{S}}_{pq} \bar{\mathbb{S}}_{qq}^{-1} \bar{B}_q$ is distributed as $N(0, \bar{B}'_q \bar{\mathbb{S}}_{qq}^{-1} \bar{B}_q \cdot I_p)$. It then follows that the distribution of $\sqrt{G}(\bar{B}_p - \bar{\mathbb{S}}_{pq} \bar{\mathbb{S}}_{qq}^{-1} \bar{B}_q)$ conditional on \mathbf{B}_q is

$$\sqrt{G}(\bar{B}_p - \bar{\mathbb{S}}_{pq} \bar{\mathbb{S}}_{qq}^{-1} \bar{B}_q) \sim N(0, (1 + \bar{B}'_q \bar{\mathbb{S}}_{qq}^{-1} \bar{B}_q) \cdot I_p),$$

using the independence of \bar{B}_p from $\bar{\mathbb{S}}_{pq}\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q$ conditional on \mathbf{B}_q . Therefore the conditional distribution of ξ_p is

$$\xi_p := \frac{\sqrt{G}(\bar{B}_p - \bar{\mathbb{S}}_{pq}\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q)}{\sqrt{1 + \bar{B}'_q\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q}} \sim N(0, I_p).$$

Given that the conditional distribution of ξ_p does not depend on \mathbf{B}_q , the unconditional distribution of ξ_p is also $N(0, I_p)$.

Using $\xi_p \sim N(0, I_p)$, $\bar{\mathbb{S}}_{pp\cdot q} \sim G^{-1}\mathbb{W}_p(G - q - 1, I_p)$, and ξ_p which is independent of $\bar{\mathbb{S}}_{pp\cdot q}$, we have

$$\xi'_p \left(\frac{G\bar{\mathbb{S}}_{pp\cdot q}}{G - q - 1} \right)^{-1} \xi_p \sim \text{Hotelling's } T^2 \text{ distribution } T^2_{p, G-q-1}.$$

It then follows that

$$\frac{G - p - q}{p(G - q - 1)} \xi'_p \left(\frac{G\bar{\mathbb{S}}_{pp\cdot q}}{G - q - 1} \right)^{-1} \xi_p \sim \mathcal{F}_{p, G-p-q}.$$

That is,

$$\frac{G - p - q}{pG} \xi'_p \bar{\mathbb{S}}_{pp\cdot q}^{-1} \xi_p \sim \mathcal{F}_{p, G-p-q}.$$

Together with Proposition 8(a) and (c), this completes the proof of the F limit theory. The proof of the t limit theory is similar and is omitted here. ■

Proof of Theorem 13.

We first show that $\hat{\mathcal{E}}_n = \mathcal{E}_{2n}(1 + o_p(1))$. For each $j = 1, \dots, d$, we have

$$\begin{aligned} \hat{\mathcal{E}}_n[., j] &= \left\{ \hat{\Gamma}' \left[\hat{\Omega}^c(\hat{\theta}_1) \right]^{-1} \hat{\Gamma} \right\}^{-1} \hat{\Gamma}' \left[\hat{\Omega}^c(\hat{\theta}_1) \right]^{-1} \frac{\partial \hat{\Omega}^c(\theta)}{\partial \theta_j} \Bigg|_{\theta=\hat{\theta}_1} \left[\hat{\Omega}^c(\hat{\theta}_1) \right]^{-1} g_n(\hat{\theta}_2^c) \\ &= \left\{ \Gamma' \left[\hat{\Omega}^c(\theta_0) \right]^{-1} \Gamma \right\}^{-1} \Gamma' \left[\hat{\Omega}^c(\theta_0) \right]^{-1} \frac{\partial \hat{\Omega}^c(\theta)}{\partial \theta_j} \Bigg|_{\theta=\hat{\theta}_1} \left[\hat{\Omega}^c(\hat{\theta}_1) \right]^{-1} g_n(\hat{\theta}_2^c)(1 + o_p(1)), \end{aligned}$$

where the second equality holds by Assumption 3, 4 and Lemma 7. Using a Taylor expansion, we have

$$g_n(\hat{\theta}_2^c) = g_n(\theta_0) - \Gamma \left\{ \Gamma' \left[\hat{\Omega}^c(\theta_0) \right]^{-1} \Gamma \right\}^{-1} \Gamma' \left[\hat{\Omega}^c(\theta_0) \right]^{-1} g_n(\theta_0)(1 + o_p(1)).$$

Thus,

$$\begin{aligned} \hat{\mathcal{E}}_n[., j] &= \left\{ \Gamma' \left[\hat{\Omega}^c(\theta_0) \right]^{-1} \Gamma \right\}^{-1} \Gamma' \left[\hat{\Omega}^c(\theta_0) \right]^{-1} \frac{\partial \hat{\Omega}^c(\theta)}{\partial \theta_j} \Bigg|_{\theta=\hat{\theta}_1} \left[\hat{\Omega}^c(\theta_0) \right]^{-1} g_n(\theta_0)(1 + o_p(1)) \\ &\quad - \left\{ \Gamma' \left[\hat{\Omega}^c(\theta_0) \right]^{-1} \Gamma \right\}^{-1} \Gamma' \left[\hat{\Omega}^c(\theta_0) \right]^{-1} \frac{\partial \hat{\Omega}^c(\theta)}{\partial \theta_j} \Bigg|_{\theta=\hat{\theta}_1} \left[\hat{\Omega}^c(\theta_0) \right]^{-1} \Gamma \\ &\quad \times \left\{ \Gamma' \left[\hat{\Omega}^c(\theta_0) \right]^{-1} \Gamma \right\}^{-1} \Gamma' \left[\hat{\Omega}^c(\theta_0) \right]^{-1} g_n(\theta_0) \right\} (1 + o_p(1)), \end{aligned}$$

for each $j = 1, \dots, d$. For the term, $\left. \frac{\partial \hat{\Omega}^c(\theta)}{\partial \theta_j} \right|_{\theta = \hat{\theta}_1}$, recall that

$$\begin{aligned} \left. \frac{\partial \hat{\Omega}^c(\theta)}{\partial \theta_j} \right|_{\theta = \hat{\theta}_1} &= \Upsilon_j(\hat{\theta}_1) + \Upsilon_j'(\hat{\theta}_1), \\ \Upsilon_j(\theta) &= \frac{1}{n} \sum_{g=1}^G \left[\sum_{i=1}^{L_g} \left(f_i^g(\theta) - \frac{1}{n} \sum_{i=1}^n f_i(\theta) \right) \left(\sum_{i=1}^{L_g} \left(\frac{\partial f_k^g(\theta)}{\partial \theta_j} - \frac{1}{n} \sum_{i=1}^n \frac{\partial f_i(\theta)}{\partial \theta_j} \right) \right) \right]'. \end{aligned}$$

It remains to show that $\Upsilon_j(\hat{\theta}_1) = \Upsilon_j(\theta_0)(1 + o_p(1))$. From the proof of Lemma 7, we have

$$\begin{aligned} &\frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} \left(f_i^g(\hat{\theta}_1) - \frac{1}{n} \sum_{i=1}^n f_i(\hat{\theta}_1) \right) \\ &= \frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} \left(f_i^g(\theta_0) - \frac{1}{n} \sum_{i=1}^n f_i(\theta_0) \right) (1 + o_p(1)), \end{aligned} \quad (35)$$

for each $g = 1, \dots, G$. By Assumption 3, 6 and a Taylor expansion, we have:

$$\begin{aligned} \frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} \frac{\partial f_i^g(\hat{\theta}_1)}{\partial \theta_j} &= \left(\frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} \frac{\partial f_i^g(\theta_0)}{\partial \theta_j} + \frac{1}{L} \sum_{i=1}^{L_g} \frac{\partial}{\partial \theta'} \left(\frac{\partial f_i^g(\theta_0)}{\partial \theta_j} \right) \sqrt{L}(\hat{\theta}_1 - \theta_0) \right) (1 + o_p(1)) \\ &:= \left(\frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} \frac{\partial f_k^g(\theta_0)}{\partial \theta_j} + Q(\theta_0) \sqrt{L}(\hat{\theta}_1 - \theta_0) \right) (1 + o_p(1)), \end{aligned}$$

for $j = 1, \dots, d$ and $g = 1, \dots, G$. This implies that

$$\frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} \left(\frac{\partial f_i^g(\hat{\theta}_1)}{\partial \theta_j} - \frac{1}{n} \sum_{i=1}^n \frac{\partial f_i(\hat{\theta}_1)}{\partial \theta_j} \right) = \frac{1}{\sqrt{L}} \sum_{i=1}^{L_g} \left(\frac{\partial f_k^g(\theta_0)}{\partial \theta_j} - \frac{1}{n} \sum_{i=1}^n \frac{\partial f_i(\theta_0)}{\partial \theta_j} \right) (1 + o_p(1)).$$

Combining these together, we have $\Upsilon(\hat{\theta}_1) = \Upsilon(\theta_0)(1 + o_p(1))$ from which we obtain the desired result

$$\hat{\mathcal{E}}_n = \mathcal{E}_{2n} (1 + o_p(1)). \quad (36)$$

Now, define the infeasible corrected variance

$$\begin{aligned} &\widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}^{\text{w,inf}}(\hat{\theta}_2^c) \\ &= \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) + \mathcal{E}_{2n} \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) + \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) \mathcal{E}_{2n}' + \mathcal{E}_{2n} \widehat{var}(\hat{\theta}_1) \mathcal{E}_{2n}', \end{aligned}$$

and the corresponding infeasible Wald statistic

$$F_{\hat{\Omega}^c(\hat{\theta}_1)}^{\text{w,inf}}(\hat{\theta}_2^c) = \frac{1}{p} (R\hat{\theta}_2^c - r)' \left[R \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}^{\text{w,inf}}(\hat{\theta}_2^c) R' \right]^{-1} (R\hat{\theta}_2^c - r).$$

The result in (36) implies

$$F_{\hat{\Omega}^c(\hat{\theta}_1)}^{\text{w,inf}}(\hat{\theta}_2^c) = F_{\hat{\Omega}^c(\hat{\theta}_1)}^{\text{w}}(\hat{\theta}_2^c)(1 + o_p(1)).$$

Also, $\mathcal{E}_{2n} = o_p(1)$ and we have

$$\widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}^{w,\text{inf}}(\hat{\theta}_2^c) = \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}^w(\hat{\theta}_2^c)(1 + o_p(1)) = \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)(1 + o_p(1)),$$

and so

$$F_{\hat{\Omega}^c(\hat{\theta}_1)}^{w,\text{inf}}(\hat{\theta}_2^c) = F_{\hat{\Omega}^c(\hat{\theta}_1)}^w(\hat{\theta}_2^c) + o_p(1) = F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) + o_p(1),$$

as desired. ■