Simple and Trustworthy Cluster-Robust GMM Inference

by

Jungbin Hwang
University of Connecticut

Working Paper 2017-19R
August 2017, Revised August 2020

365 Fairfield Way, Unit 1063
Storrs, CT 06269-1063
Phone: (860) 486-3022
Fax: (860) 486-4463
http://www.econ.uconn.edu/

Simple and Trustworthy Cluster-Robust GMM Inference

Jungbin Hwang*

Department of Economics,
University of Connecticut

August 25, 2020

Abstract

This paper develops a new asymptotic theory for GMM estimation and inference in the presence of clustered dependence. The key feature of our alternative asymptotics is that the number of clusters G is regarded as fixed as the sample size increases. Under the fixed-G asymptotics, we show that the Wald and t tests in two-step GMM are asymptotically pivotal only if we recenter the estimated moment process in the 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. The proposed tests are very appealing to practitioners because the test statistics are simple modifications of conventional GMM test statistics, and critical values are readily available from F and t tables. No further simulations or resampling methods are needed. A Monte Carlo study shows that our proposed tests are more accurate than the conventional large-G asymptotic inferences.

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 co-editor Jianqing Fan, an associate editor, and anonymous referees for constructive comments. I am also indebted to Yixiao Sun, Andres Santos, Graham Elliott, Min Seong Kim, Ji Hyung Lee, and Seojeong (Jay) Lee, all of whom helped greatly improve this paper. I would also like to thank seminar partcipants 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 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 less dependent. Failure to control for within-group or cluster dependence 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), and 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 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 or regional blocks of neighboring countries (e.g., Bertrand, Duflo, and Mullainathan, 2004; Ibragimov and Müller, 2016), we cannot convincingly claim that the number of clusters is large enough for the large-G asymptotic approximations to be 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., 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 works well whether or not the number of clusters is large. Our approach works especially well when the chi-squared and normal approximations are poor. Our approximations are obtained from an alternative limiting thought experiment in which the number of clusters G is held fixed. Under the fixed-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. This random limit of the CCE has profound implications for our 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 overidentified. We 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), who employ a CCE-type heteroskedasticity and autocorrelation consistent (HAC) estimator but consider only linear regressions and M-estimators for an exactly identified model.

We next 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. Surprisingly, however, 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 estimator[1] whose weighting matrix is constructed using recentered moment conditions. Invoking centering is not innocuous for an over-identified GMM model, because in this case the empirical moment conditions 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 hence are ignorable, even though they are not identically zero in a finite sample. Under the fixed-G asymptotics, by contrast, the 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" weighting matrix is known. The mixed normality allows us to construct 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, the critical values have to be simulated. The modifications involve only the standard J statistic, and they are very easy to implement because the modified test statistics are scaled versions of the original Wald test statistic 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.

Although the recentering scheme removes the first-order effect of the first-step estimation error, the centered two-step GMM still faces 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, which 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 a variance correction for a two-step GMM estimator, but that correction is valid only in an i.i.d. setting. We show that our finite-sample variance correction does not change the fixed-G limiting distributions of the test statistics but that it can help improve the finite-sample performance of our tests.

Monte Carlo simulations show that our new tests have much more accurate finite-sample sizes than existing 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 in the modification is the usual J statistic. There is also no need to simulate critical values, because the F and t critical values can

[1] Our definition of the centered two-step GMM estimator originates from the recentered (or demeaned) GMM weighting matrix, and it should not be confused with "centering" the estimator itself.

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 the econometrics literature by Kiefer, Vogelsang, and Bunzel (2002), Kiefer and Vogelsang (2005), Müller (2007), Sun, Phillips, and Jin (2008), Sun (2013, 2014), and 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 (co)variance 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 the theory in this paper. However, our asymptotic theory differs from that in Hwang and Sun (2017a), as it delivers the following sophistication in the CCE-type heteroskedasticity and autocorrelation robust (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 yield highly non-standard and non-pivotal fixed-G limits. This sophistication does not appear in the asymptotics of Hwang and Sun (2017a), because their 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 (2017a) in terms of the multiplicative adjustment and the degrees of freedom. Moreover, we propose a finite-sample variance correction in order to adequately capture the uncertainty embodied in the estimated moment process. To our knowledge, the finite-sample variance correction provided in this paper ant its first-order asymptotic validity have not been considered in the literature on 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 propose a group-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 test 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), and Hagemann (2019) establish a theory of randomization inferences in the context of clustered dependence. For other approaches, see Carter, Schnepel, and Steigerwald (2017) which propose a measure of the effective number of clusters under the large-G asymptotics; and 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, focus mainly on exactly identified linear regressions that are special cases of our GMM setting. 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 report simulation evidence and apply our cluster-robust tests to an empirical study in Emran and Hou (2013). The last section concludes. Proofs of the main results are given in Appendix A, and Supplementary Appendix B contains extensions of various 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 proofs that are not given in Appendix A.

2 Basic Setting and the First-Step GMM Estimator

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

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

where $f_{i,n}(\theta) = f(Y_{i,n}, \theta)$ is an $m \times 1$ matrix of twice continuously differentiable functions. For notational simplicity, we suppress the dependence of $Y_{i,n}$ and $f_{i,n}(\theta)$ on n throughout the paper. We assume that $q = m - d \geq 0$ and that the rank of $\Gamma = E\left[\partial f(Y_i, \theta_0)/\partial\theta'\right]$ is d. Thus 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 for 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, symmetric weighting matrix that does not depend on the unknown parameter θ_0 and $\text{plim}_{n\to\infty} W_n = W > 0$. 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 asymptotic properties of $\hat{\theta}_1$, we assume that for any $\sqrt{n}$-consistent estimator $\tilde{\theta}$, $\text{plim}_{n\to\infty}\hat{\Gamma}(\tilde{\theta}) = \Gamma$, and that Γ is of full column rank. Also, under some regularity conditions, we have the following central limit theorem (CLT)

$$\sqrt{n} g_n(\theta_0) \xrightarrow{d} N(0, \Omega),$$

where

$$\Omega = \lim_{n\to\infty} \frac{1}{n} E\left(\sum_{i=1}^n f_i(\theta_0)\right)\left(\sum_{i=1}^n f_i(\theta_0)\right)'.$$

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 this 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 need only estimate Ω to make reliable inference about θ_0. The main issue is how to properly account for the cross-sectional dependence in the moment process $\{f_i(\theta_0)\}_{i=1}^n$. In this paper, we assume that this cross-sectional dependence has a cluster structure, which is popular in many microeconomics applications (Wooldridge, 2003; Cameron and Miller, 2015). Specifically, our data consist of a number of asymptotically mean independent clusters, each of which has an unknown dependence structure. Let G be the total number of mutually exclusive clusters, and let L_g be the size of cluster g such that the total number of observations n is equal to $\sum_{g=1}^G L_g$.

Assumption 1 *(i) The data* $\{Y_i\}_{i=1}^n$ *consist of* G *nonoverlapping clusters. (ii) The number of clusters* G *is fixed, and the size of each cluster* g, L_g, *grows to infinity as* $n \to \infty$ *such that*

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

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

Assumption 1 (i) implies that the set $\{f_i(\theta_0)\}_{i=1}^n$ can be partitioned into G nonoverlapping clusters, that is, $\{f_i(\theta_0)\}_{i=1}^n = \cup_{g=1}^G \{f_i^g(\theta_0)\}_{i=1}^{L_g}$, where the notation in $f_i^g(\theta_0)$ indicates the ith individual in cluster g, which has size L_g. Assumption 1 (ii) assumes that the cluster structure permits the application of the joint CLT for the G-component vector of cluster sums. The joint CLT condition is a crucial device for making an asymptotically valid inference in our fixed-G asymptotics. It also characterizes the choice of clusters to be asymptotically independent, and the within-cluster dependence to be sufficiently weak to apply a suitable cross-sectional CLT. Note that unlike the literature in large-G asymptotics,[2] our Assumption 1 (ii) does not restrict the clustered data to be independent across different clusters.

As one example of the primitive conditions for the joint CLT, Supplementary Appendix B.1 considers a spatial mixing setting, (e.g., Conley (1999); and Jenish and Prucha (2009)), and assumes that the number of observations located on the boundaries between two different clusters is dominated by the average 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.[3]

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

Under Assumption 1, the total (scaled) sum-of-moments 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 from Assumption 1 (iii).

[2] See, for example, Carter, Schnepel, and Steigerwald (2017), Djogbenou, MacKinnon, and Nielsen (2019), and Hansen and Lee (2019).

[3] Bester, Conley, and Hansen (2011, hereinafter BCH), 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 applicable to only the exactly linear regression model, and thus its applicability to our GMM setting with potentially non-linear moment conditions is limited. Also, BCH (2011) assumes that all clusters are of exactly the same size, whereas we allow the cluster sizes to be unbalanced in finite-sample.

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

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

where $\Omega=G^{-1}\sum_{g=1}^{G}\Omega_g$ and Λ_g is the matrix square root of Ω_g. Thus the global covariance matrix of the entire population Ω can be represented as the simple average of Ω_1,..., Ω_G, where the 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 a linear restriction without loss of generality, because the Delta method can be used to convert non-linear restrictions into linear ones in an asymptotic sense. The F test version of the Wald test statistic is given by

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

where

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

When $p=1$ and the alternative is one sided, we can construct the t statistic $t(\hat{\theta}_1)=(R\hat{\theta}_1-r)/(R\,\widehat{var}(\hat{\theta}_1)\,R')^{1/2}$. 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 $\hat{\theta}_1\overset{p}{\to}\theta_0$.

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

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

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 $g=1,...,G$, $\Gamma_g=\Gamma$.*

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

Assumption 2 is made for convenience, and primitive sufficient conditions are available from the standard GMM asymptotic theory in Newey and MacFadden (1994). 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. The homogeneity conditions can be restrictive, depending on empirical contexts of the clustered data. However, one of the advantages of imposing them in our approach is that one can perform asymptotic F and t inferences in a broad class of GMM test statistics such as Wald/LM/LR-type test statistics. Also, our approach allows joint testing for the parameter vector θ as well as J-statistics for testing over-identification. This contrasts with alternative fixed-cluster robust inferences studied in Ibragimov and Müller (2010, 2016), Canay, Romano, and Shaikh (2017), and Canay, Santos, and Shaikh (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

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

Also, let $\mathbb{W}_p(K, \Pi)$ denote a Wishart distribution with K degrees of freedom, and a $p \times p$ 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 $\bar{B}_m$ and $\bar{\mathbb{S}}$ are independent. To present our asymptotic results, we partition $\bar{B}_m$ and $\bar{\mathbb{S}}$ as follows:

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

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) $F(\hat{\theta}_1) \xrightarrow{d} \mathbb{F}_{1\infty} := \frac{G}{p}\cdot \bar{B}_p'\bar{\mathbb{S}}_{pp}^{-1}\bar{B}_p$;

(b) $t(\hat{\theta}_1) \xrightarrow{d} \mathbb{T}_{1\infty} := \frac{N(0,1)}{\sqrt{\chi^2_{G-1}/G}}$, *where* $N(0,1)$ *and* χ^2_{G-1} *are independent.*

Let Λ be 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

$$\Omega_{1\infty} = \Lambda\tilde{\mathbb{D}}\Lambda',$$

where

$$\tilde{\mathbb{D}} = \frac{1}{G}\sum_{g=1}^{G}\tilde{D}_g\tilde{D}_g'$$

and

$$\tilde{D}_g = B_{m,g} - \underbrace{\Gamma_\Lambda(\Gamma_\Lambda' W_\Lambda^{-1}\Gamma_\Lambda)^{-1}\Gamma_\Lambda' W_\Lambda^{-1}}_{\text{Quasi-demeaning factor}}\bar{B}_m, \tag{5}$$

for $\Gamma_\Lambda = \Lambda^{-1}\Gamma$ and $W_\Lambda = \Lambda^{-1}W\left(\Lambda'\right)^{-1}$. $\tilde{D}_g$ is a quasi-demeaned version of $B_{m,g}$ with quasi-demeaning attributable to the estimation error in $\hat{\theta}_1$. 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 $\Gamma'W^{-1}\Omega_{1\infty}W^{-1}\Gamma$. Note that

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

and thus

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

Therefore, in 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 a 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 $Ef_i(\theta_0) = 0$, the model is exactly identified. This setup 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).*[4] *The modified conditional moment condition is*

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

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 an (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. Thus it is straightforward 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 those used in Theorem 6 of Andrews (2005) and are maintained in the rest of this paper.

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

[4] We thank an anonymous referee for pointing this out.

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*

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

Without imposing $\Lambda_g = \Lambda$ and $\Gamma_g = \Gamma$, we cannot further proceed with the above expressions, and we cannot 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 with 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 the large-G asymptotics, which does not require cluster homogeneity.

Remark 5 *Under the large-G asymptotics, $G \to \infty$ but $\max_{1 \le g \le G} L_g/n \to 0$, one can show that $\hat{\Omega}(\hat{\theta}_1)$ converges in probability to Ω for*

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

The proof of consistency does not require the homogeneity of Ω_g in Assumption 5 (Hansen and Lee, 2019; Djogbenou, MacKinnon, and Nielsen, 2019). Under this type of asymptotics, the test statistics $F(\hat{\theta}_1)$ and $t(\hat{\theta}_1)$ are asymptotically χ_p^2/p and $N(0,1)$, respectively. 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

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

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

3 Two-Step GMM Estimation and Inference

In an over-identified 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, which lead to consideration of both

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

and

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

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 the inference based on the first-step estimator $\hat{\theta}_1$, it does not matter which asymptotic variance estimator is used, because for any asymptotic variance estimator $\hat{\Omega}(\hat{\theta}_1)$, the Wald statistic depends on $\hat{\Omega}(\hat{\theta}_1)$ via only $\hat{\Gamma}(\hat{\theta}_1)'W_n^{-1}\hat{\Omega}(\hat{\theta}_1)W_n^{-1}\hat{\Gamma}(\hat{\theta}_1)$. It is easy to show 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 regardless of whether the estimated moment process is recentered. In the next two subsections, however, we show that the two asymptotic variance estimators in (7) and (8) are not asymptotically equivalent 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:

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

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

We first 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.

By standard asymptotic arguments and Assumption 1 (iii), we have

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

Using the joint convergences

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

we then 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 as 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

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

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

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

In the above definitions, we use a subscript notation $\hat{\Omega}(\hat{\theta}_1)$ to clarify the choice of CCE in the F and J statistics. Now the natural question is, Is the above F statistic asymptotically pivotal? (And how about the J statistic?) To answer this, we let U be the $m \times m$ matrix of the eigenvectors of $\Gamma_\Lambda'\Gamma_\Lambda = \Gamma'\Omega^{-1}\Gamma$, and we let $U\Sigma V'$ be a singular value decomposition (SVD) of Γ_Λ. By construction, $U'U = UU' = I_m$, $V'V = V'V = I_d$, and $\Sigma' = [\begin{array}{cc} A_{d\times d} & O_{d\times q}\end{array}]$. With this notation, we denote $U'W_\Lambda U$ by $\tilde{W}$ 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:

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

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

Proposition 6 *Let Assumptions 1–5 hold. Then we have*

(a) $F_{\hat{\Omega}(\hat{\theta}_1)}(\hat{\theta}_2) \xrightarrow{d} \left(\frac{G}{p}\right) \cdot \left[\bar{B}_{p+q}'\mathbb{V}_{p+q,p+q}^{-1}\bar{B}_{p+q} - \bar{B}_q'\mathbb{S}_{qq}^{-1}\bar{B}_q\right]$;

(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) hold jointly.

The proofs of Proposition 6 are available in Supplementary Appendix B.2. Because of the presence of the term $\mathbb{V}_{p+q,p+q}$, the result of Proposition 6 (a) indicates that the F statistic is not asymptotically pivotal, and that it depends on several nuisance parameters, including Ω. This is because $\mathbb{V}_{p+q,p+q}$ is a nonconstant function of $\tilde{\beta}_W^p$, which in turn depends on R, Γ, W, and Ω. While the asymptotic distribution 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 formula,[5] we check that the weak limit of $J(\hat{\theta}_2)$ satisfies

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

which implies that we have

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

[5] $(C + ab')^{-1} = C^{-1} - \frac{C^{-1}ab'C^{-1}}{1+b'C^{-1}a}$ for any invertible square matrix C, and any column vectors a, b such that $1 + b'C^{-1}a \neq 0$.

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.

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 size of the shift depends on $\hat{\theta}_1$. A key observation is that the location shift is the same for all groups under the homogeneity conditions in Assumptions 4 and 5. Therefore, if we demean the empirical moment process, we can effectively remove the location shift that is caused by the estimator error in $\hat{\theta}_1$. This motivates us to construct the recentered asymptotic variance estimator in (8).

It is important to point out that the recentering is not innocuous for the 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, however, 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 when the moment condition (1) 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 with respect to the initial estimator $\hat{\theta}_1$. In the following lemma, we prove a more general result by showing that the invariance holds for any $\sqrt{n}$-consistent estimator $\tilde{\theta}$.

Lemma 7 *Let Assumptions 1–5 hold, and let $\tilde{\theta}$ be any $\sqrt{n}$-consistent estimator of θ_0. Then we have*

(a) $\hat{\Omega}^c(\tilde{\theta}) = \hat{\Omega}^c(\theta_0) + o_p(1)$;

(b) $\hat{\Omega}^c(\theta_0) \xrightarrow{d} \Omega^c_\infty$, *where* $\Omega^c_\infty = \Lambda\bar{\mathbb{S}}\Lambda'$.

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

$$\sqrt{n}(\hat{\theta}^c_2 - \theta_0) = -\left(\Gamma'\left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1}\Gamma\right)^{-1}\Gamma'\left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1}\sqrt{n}g_n(\theta_0) + o_p(1) \tag{13}$$

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

Since $(\Omega^c_\infty)^{-1}$ is independent with $\sqrt{G}\Lambda\bar{B}_m \sim N(0,\Omega)$, the limiting distribution of $\hat{\theta}^c_2$ is mixed normal.

On the basis of $\hat{\theta}^c_2$, we can construct the normalized Wald statistic

$$F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) := \frac{1}{p}(R\hat{\theta}^c_2 - r)'\{R\,\widehat{var}_{\hat{\Omega}^c(\hat{\theta}^c_2)}(\hat{\theta}^c_2)\,R'\}^{-1}(R\hat{\theta}^c_2 - r), \tag{15}$$

where

$$\widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) = \frac{1}{n}\left(\hat{\Gamma}(\hat{\theta}^c_2)'\left(\hat{\Omega}^c(\hat{\theta}_1)\right)^{-1}\hat{\Gamma}(\hat{\theta}^c_2)\right)^{-1}.$$

When $p = 1$, the 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 $\hat{\theta}_2^c$,

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

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

$$\begin{aligned}\mathbb{F}_{2\infty} = \Bigg\{ \frac{G}{p} \cdot \left[R \left(\Gamma_\Lambda' \bar{\mathbb{S}}^{-1} \Gamma_\Lambda\right)^{-1} \Gamma_\Lambda' \bar{\mathbb{S}}^{-1} \bar{B}_m \right]' \left[R \left(\Gamma_\Lambda' \bar{\mathbb{S}}^{-1} \Gamma_\Lambda\right)^{-1} R' \right]^{-1} \\ \times \left[R \left(\Gamma_\Lambda' \bar{\mathbb{S}}^{-1} \Gamma_\Lambda\right)^{-1} \Gamma_\Lambda' \bar{\mathbb{S}}^{-1} \bar{B}_m \right] \Bigg\}.\end{aligned} \tag{16}$$

Similarly, it follows that

$$\begin{aligned}J(\hat{\theta}_2^c) \xrightarrow{d} \mathbb{J}_\infty := \Bigg[G \cdot \left\{ \bar{B}_m - \Gamma_\Lambda \left(\Gamma_\Lambda' \bar{\mathbb{S}}^{-1} \Gamma_\Lambda\right)^{-1} \Gamma_\Lambda' \bar{\mathbb{S}}^{-1} \bar{B}_m \right\}' \bar{\mathbb{S}}^{-1} \\ \times \left\{ \bar{B}_m - \Gamma_\Lambda \left(\Gamma_\Lambda' \bar{\mathbb{S}}^{-1} \Gamma_\Lambda\right)^{-1} \Gamma_\Lambda' \bar{\mathbb{S}}^{-1} \bar{B}_m \right\} \Bigg].\end{aligned} \tag{17}$$

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

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*) $F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) \xrightarrow{d} \left(\frac{G}{p}\right) \cdot \left(\bar{B}_p - \bar{\mathbb{S}}_{pq} \bar{\mathbb{S}}_{qq}^{-1} \bar{B}_q\right)' \bar{\mathbb{S}}_{pp\cdot q}^{-1} \left(\bar{B}_p - \bar{\mathbb{S}}_{pq} \bar{\mathbb{S}}_{qq}^{-1} \bar{B}_q\right)' \overset{d}{=} \mathbb{F}_{2\infty}$;

(*b*) $t_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) \xrightarrow{d} \sqrt{G} \left(\bar{B}_p - \bar{\mathbb{S}}_{pq} \bar{\mathbb{S}}_{qq}^{-1} \bar{B}_q\right) / \sqrt{\bar{\mathbb{S}}_{pp\cdot q}} \overset{d}{=} \mathbb{T}_{2\infty}$ *for* $p = 1$;

(*c*) $J(\hat{\theta}_2^c) \xrightarrow{d} G \cdot \bar{B}_q' \bar{\mathbb{S}}_{qq}^{-1} \bar{B}_q \overset{d}{=} \mathbb{J}_\infty$,

where the convergences hold jointly.

The proof of Proposition 8 is provided in Supplementary Appendix B.3. To simplify the representations of $\mathbb{F}_{2\infty}$ and $\mathbb{T}_{2\infty}$ in the Proposition, we note that

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

where $B_{p+q,g} := (B_{p,g}', B_{p,g}')'$. The above random matrix has a standard Wishart distribution $\mathbb{W}_{p+q}(G-1, I_{p+q})$. From the well-known properties of a Wishart distribution, $\bar{\mathbb{S}}_{pp\cdot q} \sim \mathbb{W}_p(G-1-q, I_p)/G$ 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

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

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 (the scaled) $\mathbb{T}_{2\infty}$ can be represented as $t_{G-1-q}(\Delta)$, which is a noncentral t distribution with a noncentrality parameter $t_{p,G-1-q}(\Delta)$. These nonstandard limiting distributions are similar to those in Sun (2014). 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

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

This is consistent with the results of Sun and Kim (2012), except that our scale adjustment and degrees-of-freedom 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 have the same fixed-G limit as the one given in Proposition 8. Similar results were obtained by Hwang and Sun (2017a), who 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, $\mathbb{F}_{2\infty}$ and $\mathbb{T}_{2\infty}$, 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 standard chi-squared and normal distributions. This section shows that a simple modification of the two-step Wald and t statistics enables us to develop standard F and t limiting distributions under the fixed-G asymptotics. The resulting asymptotic F and t tests are very appealing in empirical applications because critical values from the standard F and t distributions are more accessible than those from the nonstandard $\mathbb{F}_{2\infty}$ and $\mathbb{T}_{2\infty}$ distributions.

The modified two-step Wald and t statistics are

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

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 parameter.

It follows from Proposition 8 that

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

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

Thus

$$\tilde{F}_{\hat{\Omega}^c(\hat{\theta}_2^c)}(\hat{\theta}_2^c) \xrightarrow{d} \frac{G-p-q}{G}\cdot\frac{\mathbb{F}_{2\infty}}{1+\frac{1}{G}\mathbb{J}_\infty} \overset{d}{=} \frac{G-p-q}{pG}\cdot\xi_p'\bar{\mathbb{S}}_{pp\cdot q}^{-1}\xi_p, \qquad (21)$$

where

$$\xi_p := \frac{\sqrt{G}(\bar{B}_p - \bar{\mathbb{S}}_{pq}\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q)}{\sqrt{1+\bar{B}_q'\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q}}. \tag{22}$$

Similarly,

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

In the proof of Theorem 10, we show that ξ_p follows a standard normal distribution $N(0, I_p)$ and that 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 $\tilde{F}_{\Omega^c(\hat{\theta}_2^c)}(\hat{\theta}_2^c)$ 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 statistic converges in distribution to $\mathcal{F}_{p,G-p-q}$. Also, the modified t statistic has limiting distribution t_{G-1-q}.*

The limiting F and t results in Theorem 10 are consistent with those in Hwang and Sun (2017a), which establish a similar F and t limit theory of the two-step GMM in a time series setting. However, the F and t limits in Theorem 10 are different from those in 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

$$\sqrt{n}(\hat{\theta}_2^c - \theta_0) \xrightarrow{d} MN\left(0, \left(\Gamma'\Omega^{-1}\Gamma\right)^{-1} \cdot (1+\bar{B}_q'\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q)\right) \tag{23}$$

and

$$J(\hat{\theta}_2^c) \xrightarrow{d} G \cdot \bar{B}_q'\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q$$

hold jointly under the fixed-G asymptotics. Here, $MN(0, A)$ denotes a random variable that follows a mixed normal distribution with conditional variance A. The random multiplication factor $(1+\bar{B}_q'\tilde{S}_{qq}^{-1}\bar{B}_q)$ in (23) reflects the effect of the estimation uncertainty in the CCE weighting matrix on the limiting distribution of $\sqrt{n}(\hat{\theta}_2^c - \theta_0)$. The fixed-$G$ limiting distribution in (23) is in sharp contrast to that under the conventional large-G asymptotics, as the latter completely ignores the variability in the cluster-robust GMM weighting matrix. By the continuous mapping theorem, we obtain

$$\frac{\sqrt{n}(\hat{\theta}_2^c - \theta_0)}{\sqrt{1+\frac{1}{G}J(\hat{\theta}_2^c)}} \xrightarrow{d} N\left(0, \left(\Gamma'\Omega^{-1}\Gamma\right)^{-1}\right), \tag{24}$$

and this shows that the modification factor using the J statistic in the denominator effectively cancels out the uncertainty in the CCE and helps $\sqrt{n}(\hat{\theta}_2^c - \theta_0)$ to recover the standard normal limit. In view of (24), 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 $MN(0, \widetilde{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c))$, where

$$\widetilde{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) := \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) \cdot \left(1+\frac{1}{G}J(\hat{\theta}_2^c)\right). \tag{25}$$

The modification factor $(1+(1/G)J(\hat{\theta}_2^c))^{-1}$ degenerates to 1 as G increases, hence the two variance estimates in (25) become close to each other. Thus the multiplicative factor $(1+(1/G)J(\hat{\theta}_2^c))^{-1}$ in (18) can be regarded as a finite-sample modification to the standard variance estimate $\widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)$.[6]

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

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

Clearly, the submatrices $\bar{\mathbb{M}}_{pp}$ and $\bar{\mathbb{M}}_{qq}$ do not follow (scaled) Wishart distributions unless $\lambda_1 = \ldots = \lambda_G = 1/G$. As a result, $\bar{\mathbb{M}}_{pp\cdot q} = \bar{\mathbb{M}}_{pp} - \bar{\mathbb{M}}_{pq}\bar{\mathbb{M}}_{qq}^{-1}\bar{\mathbb{M}}_{qp}$ does not follow a (scaled) Wishart distribution and is not independent of $\bar{\mathbb{M}}_{pq}$ and $\bar{\mathbb{M}}_{qq}$. In addition, the distribution of ξ_p in (22) is not $N(0, I_p)$. Since these are the key conditions that drive the F and t limit theory in Theorem 10, we conclude that the exact asymptotic F and t theory cannot be developed without the approximately equal cluster size in Assumption 1 (iii).

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 considers 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 Section 3 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 arising from $\hat{\Omega}^c(\hat{\theta}_1)$. This is because the unobserved moment process in $\hat{\Omega}^c(\hat{\theta}_1)$ 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 stochastic approximation of the GMM estimator. The 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] 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 the corrected variance formula of Windmeijer (2005) 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

[6] For further discussion of the role of the J statistic modification, see Hwang and Sun (2017b), which casts the two-step GMM problems as OLS estimation and inference problems in classical normal linear regression.

[7] More than 5,000 citations of Windmeijer (2005), according to Google Scholar in March 2020. See also Roodman (2009) for its practical implementation in popular statistical software such as STATA.

the linear GMM two-step estimator which uses the CCE weighting matrix. Specifically, we 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 the 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

$$\sqrt{n}(\hat{\theta}_2^c - \theta_0) = -\left[\hat{\Gamma}'\left(\hat{\Omega}^c(\hat{\theta}_1)\right)^{-1}\hat{\Gamma}\right]^{-1}\hat{\Gamma}'\left(\hat{\Omega}^c(\hat{\theta}_1)\right)^{-1}\sqrt{n}g_n(\theta_0). \tag{26}$$

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

$$\sqrt{n}(\hat{\theta}_2^c - \theta_0) = \sqrt{n}(\tilde{\theta}_2^c - \theta_0) + o_p(1) \tag{27}$$

as $n \to \infty$ if G is held 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 $\hat{\Omega}^c(\hat{\theta}_1)$. Focusing on the leading-order term of $o_p(1)$ in (27), we can be more explicit in regard to the source of the extra variation as follows:

$$\begin{aligned}\sqrt{n}(\hat{\theta}_2^c - \theta_0) = &-\left[\hat{\Gamma}'\left(\hat{\Omega}^c(\hat{\theta}_1)\right)^{-1}\hat{\Gamma}\right]^{-1}\hat{\Gamma}'\left(\hat{\Omega}^c(\hat{\theta}_1)\right)^{-1}\sqrt{n}g_n(\theta_0)\\ &+(\mathcal{E}_{1n}+\mathcal{E}_{1n})\sqrt{n}(\hat{\theta}_1-\theta_0)+o_p\left(\frac{1}{\sqrt{n}}\right),\end{aligned} \tag{28}$$

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

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

and

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

respectively. It is easy to check that both $\mathcal{E}_{1n}$ and $\mathcal{E}_{2n}$ are $O_p(n^{-1/2})$, so that the term $(\mathcal{E}_{1n} + \mathcal{E}_{1n})\sqrt{n}(\hat{\theta}_1 - \theta_0)$ in (28) exactly captures the leading order of the errors in (27).

Considering the leading order term in (28), our goal is to come up with a finite-sample-corrected variance formula. To do so, we seek a refined distributional approximation of $\sqrt{n}(\hat{\theta}_2^c - \theta_0)$ based on (28). Note that, from the first-order correction (FOC) of $\hat{\theta}_2^c$, if we try to estimate the factor $\hat{\Gamma}'[\hat{\Omega}^c(\theta_0)]^{-1}g_n(\theta_0)$ 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 drop $\mathcal{E}_{1n}$ and consider only $\mathcal{E}_{2n}$ in the distributional approximation

$$\sqrt{n}(\hat{\theta}_2^c - \theta_0) \overset{a}{\sim} -\begin{pmatrix}\left[\Gamma'\left(\Omega_\infty^c\right)^{-1}\Gamma\right]^{-1} & \tilde{\mathcal{E}}_{2n}(\Gamma'W^{-1}\Gamma)^{-1}\end{pmatrix}\begin{pmatrix}\Gamma'\left(\Omega_\infty^c\right)^{-1}\Lambda Z\\ \Gamma'W\Lambda Z\end{pmatrix}, \tag{29}$$

where $Z \sim N(0, I_d)$ and Z is independent of Ω^c_∞. $\tilde{\mathcal{E}}_{2n}$ is a $d \times d$ matrix which has the same marginal distribution as $\mathcal{E}_{2n}$ and is independent of Z and Ω^c_∞. Here, the notation $\overset{a}{\sim}$ in (29) denotes that the stochastically bounded sequences of random vectors ξ_n and η_n converge in distribution to the same limit.

Since $\tilde{\mathcal{E}}_{2n}$ converges to zero in probability, the multiplicative factor $\tilde{\mathcal{E}}_{2n}(\Gamma' W^{-1}\Gamma)^{-1}$ on the right-hand side of (29) has no effect on the characterization of the first-order asymptotic distribution of $\sqrt{n}(\hat{\theta}^c_2 - \theta_0)$. Nevertheless, the smaller-order factor $\tilde{\mathcal{E}}_{2n}$ motivates us to develop a finite-sample correction to the asymptotic variance estimator. From (29), it follows that $\sqrt{n}(\hat{\theta}^c_2 - \theta_0)$ is asymptotically equivalent in distribution to the mixed normal distribution whose conditional variance is given by

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

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

$$\begin{aligned}
&\widehat{var}^{\mathrm{w}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) = \frac{1}{n}\hat{\Xi}_n \\
&= \frac{1}{n}\begin{pmatrix} \left[\hat{\Gamma}' \left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} \hat{\Gamma}\right]^{-1} & \widehat{\mathcal{E}}_n(\hat{\Gamma}' W_n^{-1}\hat{\Gamma})^{-1} \end{pmatrix} \times \begin{pmatrix} \hat{\Gamma}' \left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} \hat{\Gamma} & \hat{\Gamma}' W_n^{-1}\hat{\Gamma} \\ \hat{\Gamma}' W_n^{-1}\hat{\Gamma} & \hat{\Gamma}' W_n^{-1}\hat{\Omega}^c(\hat{\theta}_1) W_n^{-1}\hat{\Gamma} \end{pmatrix} \\
&\quad \times \begin{pmatrix} \left[\hat{\Gamma}' \left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} \hat{\Gamma}'\right]^{-1} \\ (\hat{\Gamma}' W_n^{-1}\hat{\Gamma})^{-1}\widehat{\mathcal{E}}'_n \end{pmatrix} \\
&= \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) + \widehat{\mathcal{E}}_n\, \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) + \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2)\, \widehat{\mathcal{E}}'_n + \widehat{\mathcal{E}}_n\, \widehat{var}(\hat{\theta}_1)\, \widehat{\mathcal{E}}'_n, \qquad (30)
\end{aligned}$$

where the jth column of $\widehat{\mathcal{E}}_n$ is

$$\widehat{\mathcal{E}}_n[\cdot, j] = \left\{\hat{\Gamma}' \left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} \hat{\Gamma}'\right\}^{-1} \hat{\Gamma}' \left\{\left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} \left.\frac{\partial \hat{\Omega}^c(\theta)}{\partial \theta_j}\right|_{\theta = \hat{\theta}_1} \left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1}\right\} g_n(\hat{\theta}^c_2)$$

with

$$\frac{\partial \hat{\Omega}^c(\theta)}{\partial \theta_j} = \Upsilon_j(\theta) + \Upsilon'_j(\theta)$$

and

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

The last three terms in (30), which are of smaller order, serve as a finite-sample correction to the original variance estimator.

Besides our asymptotic variance formula in (30), which extends the formula used by Windmeijer (2005), who considered an i.i.d. setting, there are two principal differences between Windmeijer's approach and ours. First, our asymptotic variance estimator in (30) involves a centered CCE; Windmeijer's formula, in contrast, involves only a plain variance estimator. Second, we consider the fixed-G asymptotics, while Windmeijer (2005) considers the traditional asymptotics $n \to \infty$ 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 from the fixed-G asymptotics point of view.

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

$$
\begin{aligned}
F^{\mathrm{w}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) &= \frac{1}{p}(R\hat{\theta}_2^c - r)' \left[R\, \widehat{var}^{\mathrm{w}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)\, R'\right]^{-1} (R\hat{\theta}_2^c - r); \\
t^{\mathrm{w}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) &= \frac{R\hat{\theta}_2^c - r}{\sqrt{R\, \widehat{var}^{\mathrm{w}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c))\, R'}}.
\end{aligned}
$$

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 $g = 1, ..., G$ and $j = 1, ..., d$, define $Q_j^g(\theta)$ as*

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

Then there exists $\epsilon > 0$ such that

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

holds for $g = 1, ..., G$ and $j = 1, ..., d$. Also, $Q_j^g(\theta_0) = Q_j(\theta_0)$ for $g = 1, ..., G$.

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

Theorem 13 *Let Assumptions 1–6 hold. Then we have*

$$
\begin{aligned}
F^{w}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) &= F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) + o_p(1)\ ; \\
t^{w}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) &= t_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) + o_p(1).
\end{aligned}
$$

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 in (28) is the same as that of the correction term, so the corrected variance formula may not necessarily provide finite-sample improvements. This is because the Jacobian matrix $\hat{\Gamma}(\hat{\theta}_1)$ for the non-linear two-step GMM estimator also depends on the plugged-in parameter $\hat{\theta}_1$, but its stochastic expansion is not taken into account in our formula. However, the proof of Theorem 13 does not rely on the linearity-in-moments and shows that our proposed finite-sample-corrected variance estimator is still consistent for both linear and non-linear models. Thus the robustness property in Theorem 13 holds for both linear and non-linear models.

A direct implication of Theorem 13 is that 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 in Theorem 10. That is,

$$
\tilde{F}^{\mathrm{w}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) := \frac{G-p-q}{G} \cdot \frac{F^{\mathrm{w}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)}{1 + \frac{1}{G}J(\hat{\theta}_2^c)} \xrightarrow{d} \mathcal{F}_{p,G-p-q} \tag{31}
$$

and

$$\tilde{t}^{\mathrm{w}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) := \sqrt{\frac{G-1-q}{G}} \cdot \frac{t^{\mathrm{w}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2)}{\sqrt{1+\frac{1}{G}J(\hat{\theta}^c_2)}} \xrightarrow{d} t_{G-1-q}. \tag{32}$$

The multiplicative modification provided in Section 4 can turn the nonstandard distributions into the standard F and t distributions in (31) and (32), 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 using (31) and (32) are very appealing to practitioners because the standard F and t critical values can be implemented in the finite-sample-corrected test statistics. No further simulations or resampling methods are needed. Thus, we recommend that practitioners to use these statistics.

Remark 14 *More broadly, we could use the higher-order terms in (28) to develop an Edgeworth expansion. This approach targets a higher-order expansion of the finite-sample distribution of the cluster-robust Wald and t 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, MacKinnon, and Nielsen (2019) requires assuming a single number for the hypothesis test* (*namely,* $p = 1$), *an exactly identified linear regression model, strict independence across clusters, an increasing number of clusters G* (*with bounded cluster sizes L_g*) *as $n \to \infty$, and a non-degenerate absolutely continuous component of probability distribution (Cramér's condition). In contrast, the approach introduced in this section 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 for the Wald and t statistics. After capturing the variation of the CCE matrix via the fixed-G asymptotics, our approach corrects a higher-order bias of the asymptotic variance matrix by using the stochastic expansion in (28).*

Remark 16 *Following Newey and Smith (2004), one can derive a more formal stochastic expansion than (28) and capture additional, smaller-order arising from W_n and $\hat{\Gamma}$. A recent paper by Hwang, Kang, and Lee (2020) shows that the additional correction for these terms in an i.i.d GMM setting provides robustness to a misspecified moment condition. Technically, it is not difficult to extend our clustered asymptotic variance formula by considering the extra corrections as in Hwang, Kang, and Lee (2020). However, simulation results in Hwang, Kang, and Lee (2020) show that, under a correctly specified moment condition, which is the main focus of this paper, there is not much numerical difference between the two formulas. It would be interesting to develop a misspecification-robust GMM inference using our fixed-G asymptotics, and we leave this for future research.*

6 Simulation Evidence

6.1 Design

We compare the finite-sample performance of our new tests by focusing on the following linear dynamic panel data model:

$$y_{it} = \gamma y_{it-1} + x_{1,it}\beta_1 + \ldots + x_{d-1,it}\beta_{d-1} + \eta_i + u_{it}. \tag{33}$$

The unknown parameter vector is $\theta = (\gamma, \beta_1, ..., \beta_{d-1})' \in \mathbb{R}^d$, and the corresponding covariates are $w_{it} = (y_{it-1}, x_{it})' \in \mathbb{R}^d$ with $x_{it} = (x_{1,it}, ..., x_{d-1,it})' \in \mathbb{R}^{d-1}$. We set the number of parameters d to 4, and the true value of θ to $\theta_0 = (0.5, 1, 1, 1)'$. The kth predetermined regressor $x_{k,it}$ for $k = 1, \ldots, d-1$, $i = 1, \ldots, n$, and $t = 1, \ldots, T$ is generated according to the following process:

$$x_{k,it} = \rho x_{k,it-1} + \eta_i + \rho u_{it-1} + e_{k,it}. \tag{34}$$

We set the number of periods T to 4. There are two key features of the panel regression model in (33) and (34). First, the presence of the common fixed effect η_i introduces a cross-sectional (intra-temporal) endogeneity between the regressor x_{it} and the outcome variable y_{it}. Second, the lagged regressor y_{it-1} in (33) creates an inter-temporal correlation with y_{it} through a dynamic feedback.

Let $\eta = (\eta_1, \ldots, \eta_n)'$, $u_t = (u_{1t}, \ldots, u_{nt})$, and $e_{k,t} = (e_{k,1t}, \ldots, e_{k,nt})'$. We characterize the cross-sectional dependence in η, e_t, and u_t using 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 $\sigma^\eta_{ij} = \lambda^{|i-j|}$ and $\sigma^u_{ij} = \lambda^{|i-j|}$, respectively. Similarly, the (i, j)th element of $\Sigma_{k,e}$ is $\sigma^e_{k,ij} = \lambda^{|i-j|}$ for $k = 1, ..., d-1$. 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 the clusters, $L = n/G$, the less the extent to which asymptotic independence presents across different groups. The individual fixed effects and shocks are generated by

$$\begin{gathered} \eta \sim N(0, \Sigma_\eta),\ e_{k,t} \overset{\text{i.i.d.}}{\sim} N(0, \Sigma_{k,e}) \text{ for } k = 1, \ldots, d-1, \\ u_t := \tau_t \Sigma_u^{1/2} (\delta_1 \omega_{1t}, , .., \delta_n \omega_{nt})', \\ \delta_i \overset{\text{i.i.d.}}{\sim} U[0.5, 1.5], \text{ and } \omega_{it} \overset{\text{i.i.d.}}{\sim} \chi^2_1 - 1, \end{gathered} \tag{35}$$

for $i = 1, ..., n$, and $t = 1, ..., T$, where $\tau_t = 0.5 + 0.1(t-1)$. The data-generating process (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 $t = 1, ..., T$, 50 initial values are generated with $\tau_t = 0.5$ for $t = -49, ..., 0$, $x_{k,-49} \sim N(\eta_i/(1-\rho), (1-\rho)^{-1}\Sigma_{k,e})$ for $k = 1, ..., d-1$, and $y_{i,-49} = (\sum_{k=1}^{d-1} x_{k,i,-49}\beta_d + \eta_i + u_{i,-49})/(1-\gamma)$. 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]

We estimate the parameter θ by the first-differenced GMM (the Arellano and Bond (1991) estimator).[9] With all possible lagged instruments $z_{it} = (y_{i0}, ..., y_{it-2}, x'_{i1}, ..., x'_{it-1})', 2 \leq t \leq T$, 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 the full set of 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 $Z_i = diag(z'_{i2}, ..., z'_{iT})$ is a $(T-1) \times m$ matrix.

[8] When the panel data are persistent with ρ being close to 1, the lagged instruments are only weakly correlated with the endogenous changes in the first-differenced data, and the GMM inferences considered in our paper could suffer from a weak identification problem (e.g., Blundell and Bond, 1998; Stock and Wright, 2000; Bun and Windmeijer, 2010). It would be interesting to extend our approach to develop weak identification-robust GMM inferences under clustered dependence, and we leave this for future research.

[9] In Supplemental Appendix B.7, we describe in detail how to implement our cluster-robust procedures in the context of the panel GMM model.

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 $\chi_p^{1-\alpha}/p$ and $\mathcal{F}_{p,G-p}^{1-\alpha}$ justified under the large-G asymptotics and the fixed-G asymptotics, respectively. Note that these two tests have the same size-adjusted power, because the modification involves only a constant multiplicative 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 $\tilde{F}_2 := F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)$ with the degrees of freedom correction 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, $\tilde{F}^{\mathrm{w}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)$, uses the most refined version of the F statistic coupled with the J-statistic modification, the degrees of freedom correction, and the finite-sample-corrected variance estimator which is defined in (31). 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).[10] Note that the consistency and the higher-order refinement of the Hall-Horowitz bootstrap procedure require the number of clusters G to tend to infinity. This is in contrast to the previous tests, which are valid under the fixed-G asymptotics.

6.3 Results with balanced and homogeneous cluster size

6.3.1 Size experiment

We first consider the case where we take different values of $G \in \{35, 50, 70, 100\}$ and the equal cluster size $L = L_1 = \ldots = L_G \in \{50, 100\}$. The null hypotheses of interest are

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

with the corresponding numbers of joint hypotheses $p = 1, 2$, and 3, respectively, and a significance level of 5%. All of our simulation results are based on 5,000 Monte Carlo repetitions, and the number of bootstrap replications is 999.

Tables 1 reports the empirical size of the first-step and two-step tests for different values of G and $L = 50$. We report only the results when $L = 50$ with $G \in \{50, 100\}$, as the qualitative observations for the other cases are quite similar. The results indicate that both the first-step and two-step tests based on the unmodified statistics F_1 and F_2 suffer from severe size distortion, when the conventional chi-squared critical values are used. $G = 50$ and $p = 3$, for example, the empirical size of the first-step chi-squared test (using the full set of IVs and $m = 24$) is 24.2%. The empirical size of the first-step F test decreases 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 in all the tests becomes less severe as the number of moment conditions decreases or the number of clusters 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% and 53.5% for $m = 24$ and $p = 3$. In view of the

[10] See Supplementary Appendix B.8 for details on how to implement the bootstrap procedure of Hall and Horowitz (1996) in the presence of clustered dependence.

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 of estimating the weighting matrix, which is not captured by the chi-squared approximation. This motivates us to implement additional corrections via degrees of freedom and the J-statistic multiplier coupled with the new critical value $\mathcal{F}^{1-\alpha}_{p,G-p-q}$. Table 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 of 5%. Lastly, we find evidence that the most refined statistic $\tilde{F}_2^{\mathrm{w}}$, equipped with the finite-sample variance correction, results in the empirical sizes lying between 5.3% and 5.9%. This indicates that the most refined two-step Wald test successfully captures the higher-order estimation uncertainty and yields more accurate tests on a finite sample. We find similar conclusions for other values of L, m, and p.

Table 1 also shows the empirical rejection probabilities of the two-step GMM bootstrap procedure by Hall and Horowitz (1996), which is indicated by HH-Bootstrap. 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 $p \in \{1, 2, 3\}$, the empirical sizes are between 0.2% and 2.0%. This fragility of the HH-Bootstrap procedure has also been 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 procedure 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 clusters G increases.

Lastly, Table 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 report only cases where $G \in \{35, 100\}$, which are shown in Figures 1 and 2, respectively. The results indicate that there is no real difference between the power curves for the modified $(\tilde{F}_2)$ and unmodified (F_2) two-step tests. However, some simulation results not reported here indicate that the modified F test can be slightly more powerful as the number of parameters increases. Also, the finite-sample-corrected test $\tilde{F}_2^{\mathrm{w}}$ does not lead to a loss of power compared with the uncorrected one, $\tilde{F}_2$.

Figures 1 and 2 also indicate that the two-step tests are more powerful than the first-step tests for most combinations of G, m, and p that we considered. The power gain of the two-step GMM procedures becomes more significant as the number of clusters G increases. This can be explained 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 of estimating the CCE weighting matrix, and the power of the first-step procedures might dominate the power of the two-step ones when the cost of employing the CCE weighting matrix outweighs the benefit of estimating it. In fact, Figure 1 shows that the power of the first-step test can be higher than that of two-step tests when $G = 35$ and $m = 24$. See Hwang and Sun (2018), which compares these two types of tests in a time series GMM framework by employing more accurate fixed-smoothing asymptotics that are in the same spirit as the fixed-G asymptotics.

In sum, our simulation evidence clearly demonstrates the size accuracy of our most refined F test, $\tilde{F}^{\mathrm{w}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2)$, 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 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 $5,000$ observations that are divided into 50 clusters. The sequence of alternative cluster-size designs begins by assigning 120 individuals to each of the first 10 clusters and 95 individuals to each of the 40 clusters. In each succeeding cluster-size design, we subtract 10 individuals from each cluster in the second group and add them to the clusters in the first group. 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 that in Carter, Schnepel and Steigerwald (2017), which investigates the behavior of the cluster-robust t statistic under cluster heterogeneity. Table 2 describes the unbalanced cluster-size designs we considered (Designs I–IV). 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 distortion becomes more severe when the degree of heterogeneity across cluster sizes increases, for example, 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 $\tilde{F}^{\mathrm{w}}_2$ yields empirical sizes of 5.8% and 7.4% , respectively, for the aforementioned values of m and p, and these empirical sizes are much closer to the nominal size. Similar results for other types of GMM tests are reported in Tables 3–5. The results of the J tests are omitted here, as they are qualitatively similar to those of the F tests.

6.4.2 Heterogeneous clusters

As our last numerical experiment, 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, that is, (ρ_g, λ_g) for $g = 1, \ldots, G$. The alternative DGPs consider randomly drawn ρ_g and λ_g for each simulation. The probability distributions for (ρ_g, λ_g) have two supports in $(0.35, 0.85)^2$ is an element of $\mathbb{R}^2$. Figure 3 shows four specific designs of probability distributions considered in our simulations. Design I describes a mild degree of heterogeneity where the values for (ρ_g, λ_g) deviate slightly from the benchmark value $(0.60, 0.60)$. Designs II–IV illustrate more substantial levels of heterogeneity, each of which represents a left-skewed, right-skewed, or symmetric distribution. To reduce the computational burden, we assume that observations are independent across different clusters. All other settings are the same as in our original model in Subsection 6.3.

With $p = 1$ in H_{01}, we consider three types of t-tests. The one-step GMM test, with its asymptotic critical value $t_{G-1}^{1-\alpha/2}$, is referred to as BCH. The two-step GMM approach, together with the J-statistic modification and the finite-sample-corrected variance formula, has asymptotic critical value $t_{G-q-1}^{1-\alpha/2}$ and is referred to as Hwang. Note that both of these tests estimate parameters using all $n = \sum_{g=1}^{G} L_g$ observations. The last method is the Fama–MacBeth-type procedure in Ibragimov and Müller (2010, 2016, hereinafter IM). IM's t-test is formed by a cluster-specific estimator which uses only observations within the cluster. Letting $R\hat{\theta}_g$ be the cluster-specific estimator for $g = 1, \ldots, G$, IM's t-statistic is

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

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 even when Assumptions 4 and 5 are violated because of the cluster heterogeneity.

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

Tables 6 present the empirical sizes, biases, and root-mean-squared errors (RMSEs) of the t-tests considered for various degrees of heterogeneity (Designs I–IV). We first note that, when the model is over-identified, IM's group-based approach suffers from severe size distortion in all simulation designs. This is consistent with our previous conjecture: that large finite-sample in the cluster-specific estimators have a negative impact on the performance of the corresponding t-tests. 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 extent of size distortion in Hwang is much smaller than that for IM. Also, compared to the one-step GMM test in BCH, Hwang's two-step GMM approach yields results in smaller size distortions. For example, in Design IV with $L = 100$, the empirical size of Hwang is 8.7%, while those of BCH and IM with over-identification are 11.5% and 78.3%, respectively. When the model is exactly identified, however, the size distortions are no longer present for IM. For example, the empirical size of the exactly identified IM test is below 5% for all degrees of heterogeneity (Designs I–IV). However, we note that this reduction in size distortion comes at the sacrifice of using the the exactly identified instrument (Anderson–Hsiao's estimator) instead of the over-identified instrument (the Arellano–Bond estimator), which results in a huge loss of efficiency. For example, Table 6 indicates that in Design IV, the IM test with exact identification has an empirical size of 2.8% with an RMSE of 1.765, while Hwang has 8.7% an empirical size of 8.7% with an RMSE of 0.303.

All in all, the results of our simulations indicate that neither our full-sample GMM tests nor the group-based tests in IM strictly dominate the other one, and that they are best described as complementary approaches. If the main concern is about the size accuracy of the parameter test under pronounced cluster heterogeneity, the IM's group-based approach is preferred. However, when there is only a mild degree of heterogeneity, or we want to increase the accuracy of the 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, in *Review of Economics and Statistics*) in development economics. Their study investigates the causal effects of access to domestic and international markets on rural household consumption. They use survey data for 7998 rural households from the Chinese Household Income Project (ICPSR 3012) in 1995. The data set is downloadable from the journal's website.[11]

7.1 Model and choice of clusters

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

$$\begin{aligned} C_i = \beta_0 + \beta_d A_i^d + \beta_s A_i^s + \beta_{ds}(A_i^d \times A_i^s) + X_i'\beta_h \\ + X_i^{v\prime}\beta_v + X_i^{c'}\beta_c + \alpha_p + \epsilon_i, \end{aligned} \tag{36}$$

where A_i^d and A_i^s are the log distances (with distances in km) of access to domestic and international markets, respectively. X_i is the vector of household characteristics that may affect consumption choice, X_i^v, X_i^c are village- and 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 the 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 rural areas. To identify these parameters, Emran and Hou (2013) employ geographic instrumental variables that capture exogenous variations in access to markets, for example, straight-line distances to the nearest navigable river and coastline, along with the topographic and agroclimatic features of the counties.[12] 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 at the county level and estimates the model using 2SLS and two-step GMM with an uncentered cluster-robust weighting matrix. The data set consists of $n = 7462$ observations divided into $G = 86$ clusters, where the number of households varies across from a low of 49 to a high of 270; the average cluster size is $\bar{L} = 87$. 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 test 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., for example, county-level clusters, and a large number of finer level of clusters, for example, at the

[11] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/CWFOFN

[12] For a detailed description of the control variables and instrument variables, see the appendix in Emran and Hou (2013).

individual level. The null hypothesis is that a finer level of clustering is appropriate with consistent CCE estimators, against the alternative that only the coarser 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 their test are described in Supplementary Appendix B.9.

Table 7 reports the results of the significance tests of the validity of the fine level (individual level) of clustering. Since the test of Ibragimov and Müller (2016) depends on a variable of interest in regression equation (36), we provide corresponding test statistics and critical values for the key variables A^d, A^s, and their interaction term $A^d \times A^s$. The results in Table 7 indicate that for all three of these variables, we reject the validity of clustering at the level of the individual at 5% significance, against the coarser clustering at the county level. These results might not be too surprising, as rural individuals within a county are closely connected through economic and cultural ties, which in turn, might well lead to non-trivial interactions.

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 the degrees-of-freedom correction uses the critical value $t_{G-1}^{1-\alpha/2}$. For the two-step GMM procedures, we implement the recommended fixed-G test in our paper, which includes the degrees-of-freedom correction, the modification factor using the J-statistic, and the finite-sample variance correction and use the critical value $t_{G-1-q}^{1-\alpha/2}$. Table 8 shows that the point estimation results, standard error estimates, and associated confidence intervals (CIs) for 2SLS and for the uncentered and centered two-step GMM estimators. Similar to Emran and Hou (2013), our results show that better access to domestic and international markets has a substantial positive effect on household consumption, and that the domestic-market effect is higher than the international-market effect. For the 2SLS method, the differences in the confidence interval and standard error between the large-G and fixed-G results are very small. This is definitely as 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 the domestic market is $\beta_d = -2722.8$. 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 a modified standard error of 520.7, which is about 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 the other parameters deliver similar qualitative messages. Table 8 also provides the finite-sample-corrected standard error estimates of two-step estimators that capture the extra variation of the 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 clusters 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 and t statistics. With the help of the standard J statistic, the Wald statistic and t statistic based on these estimators can be modified to have standard F and t limiting distributions. A finite-sample variance correction is suggested to further improve the performance of the asymptotic F and t tests.

Our simulation evidence and empirical application demonstrate the size accuracy of the two-step GMM Wald and t 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 the standard F and t critical values can be implemented for the finite-sample-corrected test statistics. No further simulations or resampling methods are needed. We recommend that practitioners 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] Blundell, R., and Bond, S. (1998): "Initial conditions and moment restrictions in dynamic panel data models." *Journal of Econometrics*, 87(1), 115–143.

[8] 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.

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

[11] Canay, I. A., Santos, A., and Shaikh, A.M. (2019): "The wild bootstrap with a 'small' number of 'large' clusters". *Review of Economics and Statistics*, pp.1–45.

[12] 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.

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

[14] 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.

[15] Conley, T. G. (1999): "GMM estimation with cross sectional dependence*". Journal of econometrics*, 92(1), pp.1–45.

[16] 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.

[17] 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.

[18] Hagemann, A. (2019): "Placebo inference on treatment effects when the number of clusters is small." *Journal of Econometrics*, 213(1), 190–209.

[19] Hall, A. R. (2000): "Covariance matrix estimation and the power of the overidentifying restrictions test." *Econometrica*, 68(6), 1517–1527.

[20] 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.

[21] Hansen, B. E. and Lee, S. (2019): "Asymptotic theory for clustered samples". *Journal of econometrics*, 210(2), pp.268–290.

[22] 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.

[23] Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica* 50, 1029–1054.

[24] 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.

[25] Hwang, J., Kang, B. and Lee, S. (2020): A Doubly Corrected Robust Variance Estimator for Linear GMM, Working paper.

[26] Hwang, J. and Sun, Y. (2017a): "Asymptotic F and t Tests in an Efficient GMM Setting." *Journal of Econometrics*, 198(2), 277–295.

[27] 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.

[28] 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.

[29] Ibragimov, R. and Müller, U. K. (2010): "t-Statistic based correlation and heterogeneity robust inference." *Journal of Business & Economic Statistics*, 28(4), 453–468.

[30] Ibragimov, R. and Müller, U.K. (2016): "Inference with few heterogeneous clusters." *Review of Economics and Statistics*, 98(1), 83–96.

[31] 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.

[32] 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.

[33] Kiefer, N. M., Vogelsang, T. J. and Bunzel, H. (2002): "Simple Robust Testing of Regression Hypotheses" *Econometrica* 68 (3), 695–714.

[34] Kiefer, N. M. and Vogelsang, T. J. (2005): "A New Asymptotic Theory for Heteroskedasticity-Autocorrelation Robust Tests." *Econometric Theory* 21, 1130–1164.

[35] Liang, K.-Y. and Zeger, S. (1986): "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73(1):13–22.

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

[37] Müller, U. K. (2007): "A theory of robust long-run variance estimation." *Journal of Econometrics*, 141(2), 1331–1352.

[38] Newey, W. K. and MacFadden, D. (1994): Large Sample Estimation and Hypothesis Testing, Chapter 36. *Handbook of Econometrics* Vol, 4.

[39] Newey, W. K. and Smith, R. J. (2004): "Higher order properties of GMM and generalized empirical likelihood estimators." *Econometrica* 72, no. 1 : 219–255.

[40] Roodman, D. (2009): "How to do xtabond2: An introduction to difference and system GMM in Stata." *The stata journal*, 9(1), 86–136.

[41] Stock, J. H. and Wright, J. H. (2000): "GMM with weak identification. Econometrica.", 68(5), 1055–1096.

[42] Stock, J. H. and Watson, M. W. (2008): "Heteroskedasticity-Robust Standard Errors for Fixed Effects Panel Data Regression." *Econometrica*, 76: 155–174.

[43] Sun, Y. (2013): "A Heteroskedasticity and Autocorrelation Robust F Test Using Orthonormal Series Variance Estimator." *Econometrics Journal* 16(1), 1–26.

[44] Sun, Y. (2014): "Fixed-smoothing Asymptotics in a Two-step GMM Framework." *Econometrica* 82(6), 2327–2370.

[45] Sun, Y. and Kim, M. S. (2012): "Simple and powerful GMM over-identification tests with accurate size." *Journal of Econometrics*, 166(2), 267–281.

[46] 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.

[47] Sun, Y., Phillips, P. C. B. and Jin, S. (2008): "Optimal Bandwidth Selection in Heteroskedasticity-Autocorrelation Robust Testing." *Econometrica* 76(1), 175–94.

[48] White, H. (1984): "Asymptotic Theory for Econometricians" (San Diego: Academic Press).

[49] Windmeijer, F. (2005): "A finite sample correction for the variance of linear efficient two-step GMM estimators." *Journal of Econometrics*, 126(1), 25–51.

[50] 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 $G \in \{50, 100\}$, the number of individuals within a cluster is $L = 50$, the number of joint hypotheses is $p \in \{1, 2, 3\}$, the number of moment conditions $m \in \{12, 24\}$, and the number of periods is 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
	F_2	$\chi_p^{1-\alpha}/p$	0.320	0.439	0.535	0.137	0.171	0.202
Two-step	$\tilde{F}_2$	$F_{p,G-p-q}^{1-\alpha}$	0.073	0.063	0.061	0.064	0.059	0.058
	$\tilde{F}_2^{\mathrm{W}}$	$F_{p,G-p-q}^{1-\alpha}$	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	–
	$\frac{G-q}{Gq} J^c$	$F_{q,G-q}^{1-\alpha}$	–	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
	F_2	$\chi_p^{1-\alpha}/p$	0.167	0.201	0.237	0.094	0.105	0.120
Two-step	$\tilde{F}_2$	$F_{p,G-p-q}^{1-\alpha}$	0.074	0.067	0.066	0.066	0.059	0.060
	$\tilde{F}_2^{\mathrm{W}}$	$F_{p,G-p-q}^{1-\alpha}$	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 statistic $J^c = J(\hat{\theta}_2^c)$ with or without degrees-of-freedom (d.f.) correction. All two-step tests are based on the centered two-step GMM estimator $\hat{\theta}_2^c$ but use different test statistics: the unmodified $F_2 = F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)$, the J-statistic-modified and d.f.-corrected $\tilde{F}_2 = \tilde{F}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)$; and the J-statistic-modified, d.f.-corrected, and finite-sample-corrected $\tilde{F}_2^{\mathrm{W}} = \tilde{F}^{\mathrm{W}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)$.

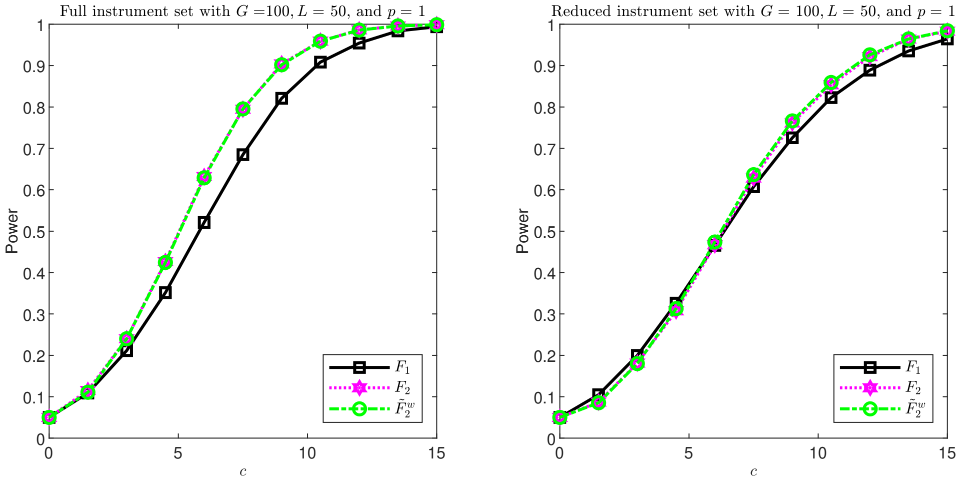

Figure 1: Size-adjusted power of the first-step (2SLS) and two-step tests with $G = 35$ and $L = 50$

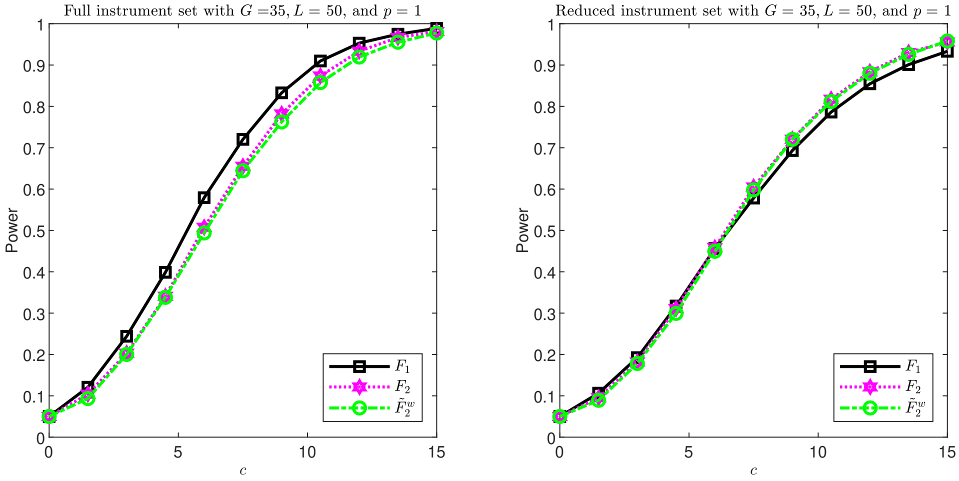

Figure 2: Size-adjusted power of the first-step (2SLS) and two-step tests with $G = 100$ and $L = 50$

Table 2: Design of unbalanced cluster size

$G=50$	$L_1 = ... = L_{10}$	$L_{11} = ... = L_{50}$	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 heterogeneity in the 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
	F_2	$\chi_p^{1-\alpha}/p$	0.307	0.441	0.545	0.139	0.173	0.210
Two-step	$\tilde{F}_2$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.071	0.069	0.070	0.065	0.062	0.066
	$\tilde{F}_2^{\text{w}}$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	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	–
	$\frac{G-q}{Gq}J^c$	$\mathcal{F}_{q,G-q}^{1-\alpha}$	–	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 heterogeneity in the 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
	F_2	$\chi_p^{1-\alpha}/p$	0.304	0.394	0.482	0.147	0.173	0.211
Two-step	$\tilde{F}_2$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.076	0.075	0.071	0.069	0.069	0.068
	$\tilde{F}_2^{\mathrm{w}}$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	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
	F_2	$\chi_p^{1-\alpha}/p$	0.336	0.469	0.586	0.153	0.193	0.242
Two-step	$\tilde{F}_2$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.084	0.081	0.084	0.072	0.071	0.075
	$\tilde{F}_2^{\mathrm{w}}$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	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 heterogeneity in the 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
	F_2	$\chi_p^{1-\alpha}/p$	0.360	0.511	0.621	0.165	0.218	0.271
Two-step	$\tilde{F}_2$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.091	0.095	0.100	0.082	0.086	0.099
	$\tilde{F}_2^{\mathrm{w}}$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	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.

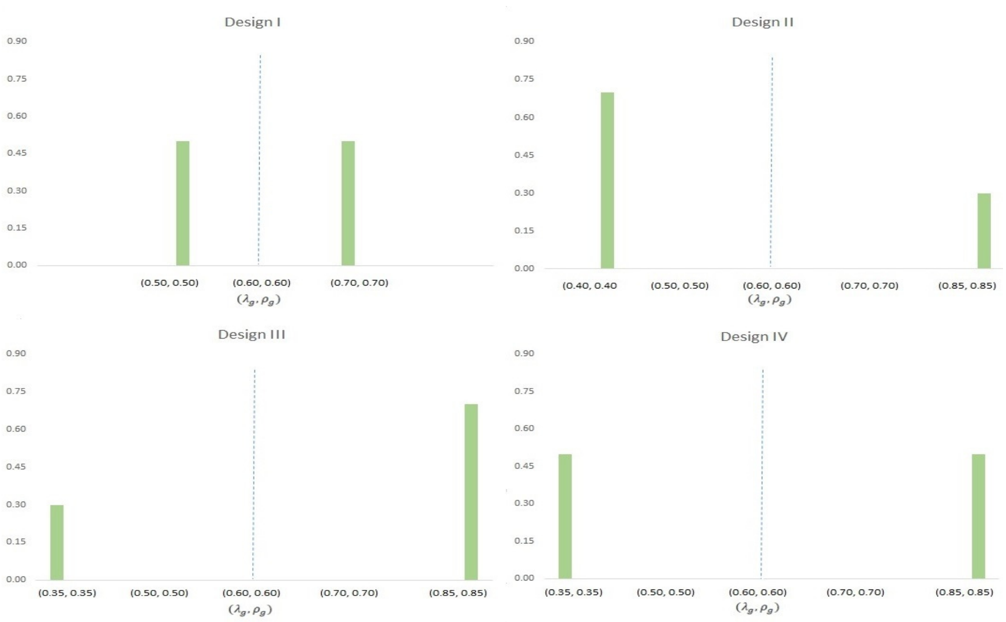

Figure 3: Design of heterogeneous clusters: Probability distributions for heteregeneous (ρ_g, λ_g)

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

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 the 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-corrected $\tilde{t}_2^{\mathrm{w}} = \tilde{t}^{w}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2)$ and critical value t_{G-q-1}. The exactly identified moment condition is constructed using the instrument variables of Anderson and Hsiao (1981). The over-identified moment condition is constructed from the instruments of Arellano and Bond (1991) using the most recent lag.

Table 7: 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.			
	Key variables		
·	A^d	A^s	$A^d \times A^s$
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 interest in (36). In all cases, the test statistics and critical values are calculated by the method of Ibragimov and Müller (2016).

Table 8: Results for Emran and Hou (2013) data

	2SLS	
Variables	Large-G asymptotics	Fixed-G asymptotics
Domestic market (A_i^d)	-2713.2 (712.1)	-2713.2 (716.8)
	$[-4109.9, -1316.4]$	$[-4138.0, -1288.0]$
International market (A_i^s)	-1993.5 (514.8)	-1993.5 (517.9)
	$[-3002.5, -984.4]$	$[-3023.10, -963.8]$
Interaction ($A_i^d \times A_i^s$)	345.8 (105.0)	345.8 (105.6)
	$[140.0, 551.7]$	$[135.8, 555.9]$
$H_0 \ : \beta_d = \beta_s$	-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)	-2670.0 (520.7)
	$[-3507.7, -1937.9]$	$[-3709.2, -1630.7]$
International market (A_i^s)	-2000.2 (344.3)	-1981.3 (447.7)
	$[-2675.0, -1325.5]$	$[-2874.9, -1087.7]$
Interaction ($A_i^d \times A_i^s$)	362.7 (68.7)	364.1 (89.4)
	$[228.0, 497.3]$	$[187.5, 542.4]$
$H_0 \ : \beta_d = \beta_s$	-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 the (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 A: Proofs of Main Results

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

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

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

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

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

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

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

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

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

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

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

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

Therefore,

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

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:

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

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

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

Using this, we get

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

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:

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

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

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

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

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

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

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

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

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

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:

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

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

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

■

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

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

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

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

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

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

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

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

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 $\widehat{\mathcal{E}}_n = \mathcal{E}_{2n}\left(1+o_p\left(1\right)\right).$ For each $j = 1, ..., d$, we have

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

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

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

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

$$\left.\frac{\partial \hat{\Omega}^c(\theta)}{\partial \theta_j}\right|_{\theta=\hat{\theta}_1} = \Upsilon_j(\hat{\theta}_1) + \Upsilon_j'(\hat{\theta}_1),$$

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

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

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

for each $g = 1, ..., G$. By Assumptions 3, 6, and a Taylor expansion, we have

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

for $j = 1, ..., d$ and $g = 1, ..., G$. 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

$$\widehat{\mathcal{E}}_n = \mathcal{E}_{2n}\left(1 + o_p\left(1\right)\right). \tag{37}$$

Now, define the infeasible corrected variance

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

and the corresponding infeasible Wald statistic

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

The result in (37) implies that

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

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

$$\begin{aligned}\widehat{var}^{\mathrm{w,inf}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) &= \widehat{var}^{\mathrm{w}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)(1 + o_p(1)) \\ &= \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)(1 + o_p(1)),\end{aligned}$$

and so

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

as desired. ■

Online Supplementary Appendix
"Simple and Trustworthy Cluster-Robust GMM Inference"

by Jungbin Hwang[1]
University of Connecticut

This version: August 2020

Appendix B: Supplemental Material

B.1 Clustered (grouped) dependence in spatial setting

We provide a set of primitive conditions and prove the key conditions used in the main body of the paper to establish the GMM theory under fixed-G asymptotics. The two main results that we investigate are the joint central limit theorem (CLT) condition in Assumption 1 (ii) and the uniform law of large numbers (ULLN) for the Jacobian process in Assumption 3 (i). Relating our work to existing literature on spatial econometrics, Conley (1999) considers weakly dependent spatial random fields and develops an asymptotic theory for GMM estimation, but does not have cluster or group structure. Bester, Conley, and Hansen (2011, hereinafter BCH), which is closely related to our clustered structure, considers a spatial setting with group structure and makes an essential contribution toward providing a set of regularity conditions that are sufficient to obtain their fixed-G limiting distributions. However, the asymptotic theory developed in BCH (2011) is applicable only to the exactly linear regression model, and it thus has limited applicability to our GMM setting with potentially non-linear moment conditions. Also, BCH (2011) assumes exactly equal cluster size, whereas we allow the cluster sizes to be unbalanced.

We follow Jenish and Prucha (2009, hereinafter JP) and consider a generic spatial random field $\{Y_{i,n}\}_{i\in B_n}$ which is located on a sampling region B_n. We assume that the sampling region B_n observations and that B_n is a finite subset of $\mathbb{Z}_v$, where $\mathbb{Z}_v$ is a (possibly unevenly spaced) v-dimensional integer lattice which grows uniformly in v non-opposing directions as $n\to\infty$. The clustered structure in a spatial setting means that the spatial process $\{f(Y_{i,n},\theta), i\in B_n\}$ can be partitioned into G sub-sampling regions in $B_n=\cup_{g=1}^{G}\mathcal{G}_{g,n}$, that is,

$$\{f(Y_{i,n},\theta), i\in B_n\}=\cup_{g=1}^{G}\{f_{i,n}(\theta): i\in\mathcal{G}_{g,n}\}.^{2}$$

For a given sub-sampling region $\mathcal{G}\subseteq B_n$, we use $|\mathcal{G}|$ to refer to the number of samples in the region, for example, $|B_n|=n$ and $|\mathcal{G}_{g,n}|=L_g$. We also assume that $\mathbb{Z}_v$ is equipped with the matrix function $\operatorname{dist}(i,j)=\max\{|i_1-j_1|,\ldots,|i_v-j_v|\}$ for any two $i=(i_1,\ldots,i_v)'$ and $j=(j_1,\ldots,j_v)'$ in $\mathbb{Z}_v$. The distance between any subsets $U,V\subset\mathbb{R}^v$ is then defined as $\operatorname{dist}(U,V)=\min\{\operatorname{dist}(i,j),\ i\in U$ and $j\in V\}$. All elements in B_n are assumed to be located at a minimum distance of $d_0\geq 1$ uniformly over the sample size n. Without loss of generality, we assume that $d_0=1$. The following definition is the standard notion of a strong mixing coefficient of a spatial process on a generic random field $\{W_{i,n}\}_{i\in B_n}$ in JP (2009).

Definition B.1 (Spatial mixing coefficients) *For given sub-sampling regions $U\subseteq B_n$ and $V\subseteq B_n$, let $\sigma_n(U)=\sigma(\{W_{i,n}\}_{i\in U})$ be the smallest σ-field generated by $\{W_{i,n}\}_{i\in U}$, and define $\sigma_n(V)$*

[1] Address: 365 Fairfield Way U-1063, Storrs, CT 06269-1063. Email: jungbin.hwang@uconn.edu.

[2] Note that $f_{i,n}(\theta)$ with $i\in\mathcal{G}_{g,n}$ is equivalent to the $f_i^g(\theta)$ which is used in the main body of our paper.

similarly using V. Then the α- and ϕ- mixing coefficients between $\sigma_n(U)$ and $\sigma_n(V)$ are defined as

$$\alpha_n(U,V) = \sup_{A\in\sigma_n(U),\ B\in\sigma_n(V)} \{P(A\cap B) - P(A)P(B)\}; \tag{B.1}$$

$$\phi_n(U,V) = \sup_{A\in\sigma_n(U),\ B\in\sigma_n(V), P(B)>0} \{P(A|B) - P(A)\}. \tag{B.2}$$

Also, the generalized α- and ϕ- mixing coefficients, with $k, l, d \in \mathbb{N}$, for a spatial random field $\{W_{i,n}\}_{i\in B_n}$ are defined as

$$\alpha_{k,l}(d) = \sup_{n\in\mathbb{N}} \sup_{U,V\subseteq B_n} \{\alpha_n(U,V) :\ |U| \le k, |V| \le l,\ dist(U,V) \ge d\}; \tag{B.3}$$

$$\phi_{k,l}(d) = \sup_{n\in\mathbb{N}} \sup_{U,V\subseteq B_n} \{\phi_n(U,V) :\ |U| \le k, |V| \le l,\ dist(U,V) \ge d\}. \tag{B.4}$$

Given k and l, the α- and ϕ- mixing conditions require $\alpha_{k,l}(d)$ and $\phi_{k,l}(d)$, respectively to decay to zero as $d \to \infty$.

B.1.1 Primitive conditions for ULLN of Jacobian process

We begin by imposing the following set of mixing conditions.

Assumption B.1 (α- and ϕ- mixing) *The observations on $B_n, \{Y_{i,n}\}_{i\in B_n}$, satisfy the following α- and ϕ- mixing coefficients:*
i) $\sum_{d=1}^{\infty} d^{\nu-1}\alpha_{1,1}(d) < \infty$.
ii) $\sum_{d=1}^{\infty} d^{\nu-1}\phi_{1,1}(d) < \infty$.

The mixing condition in Assumption B.1 allows for a general form of weak dependence, together with heteroskedasticity and non-stationarity, among the cross-sectional units $Y_{i,n}$. See, for example, Bolthausen (1982) and JP (2011). Since the mixing conditions are characterized by the entire sampling region B_n, Assumption B.1 allows the dependency across different sub-sampling regions or clusters. The mixing conditions are also provided in BCH (2011) to give the spatial LLN and CLT, but only the α-mixing condition is discussed. Given our m-dimensional moment process $f(Y_{i,n},\theta)$ on $\Theta \subseteq \mathbb{R}^d$, let $q_{(j)}(Y_{i,n},\theta)$ be a dm-component vector of stochastic functions $vec(\partial f(Y_{i,n},\theta)/\theta')$. In addition to Assumption B.1, we further impose the following conditions on $\{q_{(j)}(Y_{i,n},\theta)\}_{i\in B_n}$.

Assumption B.2 *For $j = 1,\dots,dm$, the following conditions hold*:
i) $\sup_{n\in\mathbb{N}} \sup_{i\in B_n} E[d_{(j),i,n}^{1+\delta}] < \infty$ *for some $\delta > 0$, where* $d_{(j),i,n} := \sup_{\theta\in\Theta} |q_{(j)}(Y_{i,n},\theta)|$.
ii) For all $\theta, \theta' \in \Theta$, the stochastic function $q_{(j)}(Y_{i,n},\theta)$ satisfies

$$|q_{(j)}(Y_{i,n},\theta) - q_{(j)}(Y_{i,n},\theta')| \le R_{i,n} \cdot h(\theta,\theta') \text{ almost surely,}$$

where $h(\theta,\theta')$ is a non-random function such that $h(\theta,\theta') \to 0$ as $\theta \to \theta'$, and $\{R_{i,n}\}_{i\in B_n}$ are random variables which do not depend on θ such that

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{i\in B_n} E[R_{i,n}^p] < \infty \text{ for some } p > 0.$$

Recalling that $f(Y_{i,n},\theta)$ is Borel-measurable and continuously differentiable function at each $\theta \in \Theta$, for each j it is easy to check that $\{q_{(j)}(\cdot,\theta) : i \in B_n, n \in \mathbb{N}\}$ is also a well-defined family of Borel measurable functions for a given $\theta \in \Theta$. Since the α-mixing and ϕ-mixing coefficient conditions in Assumption B.1 are preserved under any measurable transformation, the random field $\{q_{(j)}(Y_{i,n},\theta)\}_{i\in B_n}$ also satisfies Assumption B.1 at each $\theta\in\Theta$. Assumption B.2 (i) implies the uniform $L_{1+\delta}$-boundedness condition of $q_{(j)}(Y_{i,n},\theta)$. In the context of a linear IV regression model with $f_{i,n}(\theta_0)=Z_{i,n}(y_{i,n}-x_{i,n}'\theta_0)$, Assumption B.2 (i) is guaranteed once we assume that for some $r>1$,

$$\sup_{n\in\mathbb{N}}\sup_{i\in B_n} E[Z^{2r}_{(s),i,n}]<\infty \text{ for } s=1,\ldots,m; \tag{B.5}$$

$$\sup_{n\in\mathbb{N}}\sup_{i\in B_n} E[x^{2r}_{(s),i,n}]<\infty \text{ for } s=1,\ldots,d. \tag{B.6}$$

Assumption B.2 (ii) imposes the smoothness condition on the moment process. It is trivially satisfied under the linear moment condition.

Proposition B.2 *Under Assumption B.1 (i) or Assumption B.1 (ii), and Assumption B.2,*

$$\sup_{\theta\in\Theta}\left\|\frac{1}{L_g}\sum_{i\in\mathcal{G}_{g,n}}\frac{\partial f_i(\theta)}{\partial\theta'}-\Gamma_g(\theta)\right\|\xrightarrow{p}0$$

as $n\to\infty$ holding G fixed, for $g=1,\ldots,G$.

The proof of the Proposition is given in Subsection B.1.3.

B.1.2 Primitive conditions for CLT for group means

We provide a joint CLT for the following G-array of (scaled) group means,

$$\left(\frac{1}{\sqrt{L_1}}\sum_{i\in\mathcal{G}_{1,n}}f_{i,n}(\theta_0)',\ldots,\frac{1}{\sqrt{L_G}}\sum_{i\in\mathcal{G}_{G,n}}f_{i,n}(\theta_0)'\right)'\in\mathbb{R}^{Gm}, \tag{B.7}$$

which is established under asymptotically unbalanced cluster sizes. The joint CLT focuses on the α- and ϕ- mixing conditions introduced in the previous section. We impose the following set of mixing and moment conditions.

Assumption B.3 (α-mixing) *The observations on B_n, $\{Y_{i,n}\}_{i\in B_n}$, satisfy the following α-mixing coefficients for some $\delta>0$:*
i) $\sum_{d=1}^{\infty}d^{\nu-1}[\alpha_{1,1}(d)]^{\frac{\delta}{2+\delta}}<\infty$.
ii) $\sum_{d=1}^{\infty}d^{\nu-1}\alpha_{k,l}(d)<\infty$ *for* $k+l\le 4$.
iii) $\alpha_{1,\infty}(d)=O(d^{-\nu-\delta})$.

Assumption B.4 (ϕ-mixing) *The observations on B_n, $\{Y_{i,n}\}_{i\in B_n}$, satisfy the following ϕ-mixing coefficients:*
i) $\sum_{d=1}^{\infty}d^{\nu-1}[\phi_{1,1}(d)]^{1/2}<\infty$.
ii) $\sum_{d=1}^{\infty}d^{\nu-1}\phi_{k,l}(d)<\infty$ *for* $k+l\le 4$.
iii) $\phi_{1,\infty}(d)=O(d^{-\nu-\delta})$ *for some* $\delta>0$.

Assumption B.5 $\sup_{n\in\mathbb{N}}\sup_{i\in B_n}E[||f_{i,n}(\theta_0)||^{2+\delta}]<\infty$ *for some* $\delta>0$.

Assumptions B.3 and B.4 are mixing conditions for the spatial random field $\{Y_{i,n}\}_{i \in B_n}$. It is important to point out again that those conditions allow the weak dependence to be presented between different clusters as well as within a cluster. Assumption B.5 is a moment restriction that guarantees the uniform L_2-integrability condition for $\{f_{i,n}(\theta_0)\}_{i \in B_n}$. In the linear IV model, the condition holds if (B.5) and

$$\sup_{n \in \mathbb{N}} \sup_{i \in B_n} E[u_{i,n}^{2r}] < \infty \text{ holds with } u_{i,n} = y_{i,n} - x_{i,n}'\theta_0$$

hold for some $r > 2$. Our Assumptions B.3–B.5 are necessary for establishing the CLT for the spatial random field $\{f_{i,n}(\theta_0)\}_{i \in B_n}$; see, for example, Dedecker (1998) and JP (2009).

Assuming that the size of each cluster grows to infinity but the number of clusters G is fixed, BCH (2011) develops a CLT for a linear regression model. Following BCH (2011), we use the notation $\partial \mathcal{G}_{g,n}$ to refer to the boundary of region $\mathcal{G}_g$ and define it as

$$\partial \mathcal{G}_{g,n} = \{i \in \mathcal{G}_{g,n} : \text{there exists } j \neq \mathcal{G}_{h,n} \text{ such that } \mathrm{dist}(i,j) = d_0 \text{ and } g \neq h\}.$$

Assumption B.6 *i) Groups are mutually exclusive and exhaustive. ii) Groups are contiguous in the metric distance function dist$(\cdot,\cdot)$. iii) For all $g = 1, \ldots, G$, $|\partial \mathcal{G}_{g,n}| < C\bar{L}^{\frac{\nu-1}{\nu}}$, where $\bar{L} = G^{-1}\sum_{g=1}^{G} L_g$.*

Assumption B.7 *For $g = 1, ..., G$,*

$$\lim_{L_g \to \infty} var\left(\frac{1}{\sqrt{L_g}} \sum_{i \in \mathcal{G}_{g,n}} f_{i,n}(\theta_0)\right) = \Omega_g > 0.$$

Assumption B.6 is about geographical restrictions on clustered sampling regions $\{\mathcal{G}_1, \ldots, \mathcal{G}_G\}$. Assumption B.7 guarantees that each group has asymptotically non-negligible variations in the limits. Parts (i)–(iii) of Assumption B.6 are similar to conditions (i)–(iv) of Assumption 2 in BCH (2011), but our condition (iii) substantially relaxes conditions (iii) and (iv) in BCH (2011) because it does not require equal cluster sizes, $L_1 = L_2 = \ldots = L_G = \bar{L}$. In fact, what we need to establish the joint CLT for the random array in (B.7) different cluster sizes grow. The corresponding cluster sizes are allowed to be unbalanced, even asymptotically. The following proposition formally provides the joint asymptotic normality and independence of cluster means.

Proposition B.3 *Under Assumption B.3 or Assumption B.4, and Assumptions B.6–B.7, and $n \to \infty$ such that G fixed with $L_g/n \to \lambda_g > 0$, we have*

$$\begin{pmatrix} \frac{1}{\sqrt{L_1}} \sum_{i \in \mathcal{G}_{1,n}} f_{i,n}(\theta_0) \\ \vdots \\ \frac{1}{\sqrt{L_G}} \sum_{i \in \mathcal{G}_{G,n}} f_{i,n}(\theta_0) \end{pmatrix} \xrightarrow{d} N\left(0, \begin{pmatrix} \Omega_1 & & 0 \\ & \ddots & \\ 0 & & \Omega_G \end{pmatrix}\right). \tag{B.8}$$

The proof of the Proposition is given in Subsection B.1.3. It proceeds by showing that the G-random array in (B.7) after rescaling converges to a multivariate normal distribution, and the corresponding variance–covariance matrix is block diagonal in the limit. Our proof of the Proposition follows largely from the proofs in BCH (2011), the major difference being that we allow for asymptotically unbalanced cluster sizes. Also, we explicitly investigate the joint asymptotic normality of the G-random array in (B.7). This makes our proof more straightforward than the proofs of BCH (2011), who explain the joint convergence only through the marginal convergences of the individual summations in (B.7).

For $g_n(\theta) = n^{-1}\sum_{i\in B_n} f_{i,n}(\theta)$, the total sum-of-moments process can be decomposed into the following form of cluster sums:

$$\sqrt{n}g_n(\theta_0) = \sum_{g=1}^{G}\sqrt{\frac{L_g}{n}}\frac{1}{\sqrt{L_g}}\sum_{i\in\mathcal{G}_{g,n}} f_{i,n}(\theta_0).$$

Then the result in Proposition B.3 together with the continuous mapping theorem implies the following "fixed" cluster central limit theorem (CLT):

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

where $\Omega = \sum_{g=1}^{G}\lambda_g\Omega_g$.

B.1.3 Proof of Propositions B.2 and B.3

Proof of Proposition B.2. Let $\Gamma_g^{(j)}(\theta)$ be the jth element in $vec(\Gamma_g(\theta))$. It suffices to show that the following ULLN holds for $j = 1,\ldots,dm$:

$$\sup_{\theta\in\Theta}\left|\frac{1}{L_g}\sum_{i\in\mathcal{G}_{g,n}} q_{(j)}(Y_{i,n},\theta) - \Gamma_g^{(j)}(\theta)\right| \xrightarrow{p} 0. \tag{B.9}$$

We prove (B.9) using Theorem 2(a) in JP (2009). Setting the non-random constants $c_{i,n}$ to 1 in JP (2009), we check that (B.9) holds if the following are true:

$$\lim_{M\to\infty}\limsup_{L_g\to\infty}\left(\frac{1}{L_g}\sum_{i\in\mathcal{G}_{g,n}} E[d_{(j),i,n}^{p}1(d_{(j),i,n} > M)]\right) = 0 \text{ for some } p \geq 1 \text{ (domination);} \tag{B.10}$$

$$\left|\frac{1}{L_g}\sum_{i\in\mathcal{G}_{g,n}} q_{(j)}(Y_{i,n},\theta) - \Gamma_g^{(j)}(\theta)\right| \xrightarrow{p} 0 \text{ for each } \theta\in\Theta \text{ (pointwise LLN);} \tag{B.11}$$

and the spatial random process $\{q_{(j)}(Y_{i,n},\theta) : i \in \mathcal{G}_{g,n}, n \in \mathbb{N}\}$ satisfies

$$\limsup_{n}\frac{1}{L_g}\sum_{i\in\mathcal{G}_{g,n}} P\left(\sup_{\theta'\in\Theta}\sup_{\theta\in B(\theta',\delta)}\left|q_{(j)}(Y_{i,n},\theta) - q_{(j)}(Y_{i,n},\theta') > \epsilon\right.\right) \to 0, \text{ as } \delta \to 0. \tag{B.12}$$

(L_0 stochastic equicontinuity on Θ)

We begin by showing the domination condition in (B.10). It is not difficult to check that (B.10) is implied by

$$\lim_{M\to\infty}\sup_{n\in\mathbb{N}}\sup_{i\in\mathcal{G}_{g,n}} E\left[d_{(j),i,n}^{p}1(d_{(j),i,n} > M)\right] = 0 \text{ for some } p \geq 1. \tag{B.13}$$

Using Assumption B.2(i), we can choose $p = 1$ such that

$$\begin{aligned}
&\sup_{n\in\mathbb{N}} \sup_{i\in\mathcal{G}_{g,n}} E\left[d_{(j),i,n} 1(d_{(j),i,n} > M)\right] \\
&\leq \sup_{n\in\mathbb{N}} \sup_{i\in\mathcal{G}_{g,n}} E\left[M^{-\delta} d_{(j),i,n}^{1+\delta} 1(d_{(j),i,n} > M)\right] \\
&\leq \sup_{n\in\mathbb{N}} \sup_{i\in B_n} E\left[d_{(j),i,n}^{1+\delta}\right] \cdot M^{-\delta} \to 0 \text{ as } M \to \infty,
\end{aligned}$$

which gives us the desired result in (B.10).

To show the pointwise LLN in (B.11), we note that the generalized α- and ϕ- mixing coefficients for the sub-sampling region $\{Y_{i,n}\}_{i\in\mathcal{G}_{g,n}}$ can easily be specified by the definitions in (B.1)–(B.3), and we check that the mixing conditions in Assumption B.1 are also satisfied for cluster $\{Y_{i,n}\}_{i\in\mathcal{G}_{g,n}}$. Moreover, the mixing conditions in Assumption B.1 for $\{Y_{i,n}\}_{i\in\mathcal{G}_{g,n}}$ are preserved in its measurable transformation $\{q_{(j)}(Y_i, \theta)\}_{i\in\mathcal{G}_{g,n}}$, which indicates that condition (a) or (b) in Theorem 3 in JP (2009) holds. Also, it is easy to check that our Assumption B.2(i) directly satisfies Assumption 2^* in JP (2009), and that both of these assumptions state the pointwise uniform L_1 integrability of $q_{(j)}(Y_{i,n}, \theta)$. Therefore, we can apply Theorem 3 in JP (2009) and obtain the pointwise LLN in (B.11).

Lastly, it follows directly from Proposition 1 in JP (2009) that our Assumption B.2(ii) implies the L_0 stochastic equicontinuity condition in (B.12). ■

Proof of Proposition B.3. We want to show the joint CLT

$$\begin{pmatrix} \frac{1}{\sqrt{n}} \sum_{i\in\mathcal{G}_{1,n}} f_{i,n}(\theta_0) \\ \vdots \\ \frac{1}{\sqrt{n}} \sum_{i\in\mathcal{G}_{G,n}} f_{i,n}(\theta_0) \end{pmatrix} \xrightarrow{d} N\left(0, \begin{pmatrix} \lambda_1\Omega_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_G\Omega_G \end{pmatrix}\right), \tag{B.14}$$

which, together with $L_g/n \to \lambda_g > 0$, implies the joint CLT result in (B.8). For any non-zero $t = (t_1', \ldots, t_G') \in \mathbb{R}^{Gm}$ with $t_g \in \mathbb{R}^m$ for $g = 1, \ldots, G$, define

$$\begin{aligned}
t'S_n(\theta_0) &= \sum_{i\in\mathcal{G}_{1,n}} t_1' f_{i,n}(\theta_0) + ... + \sum_{i\in\mathcal{G}_{G,n}} t_G' f_{i,n}(\theta_0); \\
\sigma_{n,t}^2 &= var(t'S_n(\theta_0)).
\end{aligned}$$

Also, for $g, h = 1, \ldots, G$, let

$$V_{gh,n} = E\left[\left(\sum_{i\in\mathcal{G}_{g,n}} f_{i,n}(\theta_0)\right)\left(\sum_{i\in\mathcal{G}_{h,n}} f_{i,n}(\theta_0)\right)'\right].$$

Then we have

$$\frac{\sigma_{n,t}^2}{n} = \sum_{g=1}^{G} \left(\frac{L_g}{n}\right) \cdot t_g' \left(\frac{V_{gg,n}}{L_g}\right) t_g + \frac{1}{n} \sum_{g\neq h} t_g' V_{gh,n} t_h. \tag{B.15}$$

By the Cramer–Wold device and Slutsky's theorem, (B.14) is implied by

$$\frac{\sigma_{n,t}^2}{n} \to \sum_{g=1}^{G} \lambda_g (t_g' \Omega_g t_g) > 0; \tag{B.16}$$

$$\frac{t'S_n(\theta_0)}{\sigma_{n,t}} \xrightarrow{d} N(0,1). \tag{B.17}$$

Without loss of generality, we consider the case where $\{f_{i,n}(\theta_0)\}_{i\in B_n}$ is a scalar random field. The vector case can be dealt with by constructing an arbitrary linear combination of $f_{i,n}(\theta_0)$ and repeating the Cramer–Wold device.

We first prove (B.16) for an α-mixing random field. In view of (B.15), Assumption B.7 and $L_g/n \to \lambda_g > 0$, (B.16) holds if

$$\frac{1}{n}|V_{gh,n}| \leq \frac{1}{n}\sum_{i\in\mathcal{G}_{g,n}}\sum_{j\in\mathcal{G}_{h,n}} |Ef_{i,n}(\theta_0)f_{j,n}(\theta_0)| \to 0 \tag{B.18}$$

for any $g \neq h$. Define the set of dth-order neighbors located on two clusters $g \neq h$ as

$$N_{g,h}(d) = \{(i,j) \in B_n \times B_n : \text{dist}(i,j) = d,\ i \in \mathcal{G}_{g,n} \text{ and } j \in \mathcal{G}_{h,n}\}.$$

Under Assumption B.5, we can use a standard α-inequality (e.g., Lemma 1 in Bolthausen (1982)), and obtain

$$|Ef_{i,n}(\theta_0)f_{j,n}(\theta_0)| \leq \Delta \cdot [\alpha_{1,1}(d)]^{\frac{\delta}{2+\delta}} \tag{B.19}$$

for any $(i,j) \in N_{g,h}(d)$, where Δ is a positive constant. Also, using the same arguments as in the proof of Lemma 1 in BCH (2011, p. 149), we can check that the geographic restrictions in Assumption B.6 imply that

$$|N_{g,h}(d)| \leq C_\nu \cdot \bar{L}^{\frac{\nu-1}{\nu}} \cdot d \tag{B.20}$$

for some C_v which only depends on the dimension of index set ν. That is, the maximum number of pairs in the dth order neighbor set is bounded by $C_\nu \cdot \bar{L}^{\frac{\nu-1}{\nu}} \cdot d$. Combining the results in (B.19) and (B.20), we obtain

$$\begin{aligned}
&\sum_{i\in\mathcal{G}_{g,n}}\sum_{j\in\mathcal{G}_{h,n}} |Ef_{i,n}(\theta_0)f_{j,n}(\theta_0)'| \\
&= \sum_{d=1,(i,j)\in N_{g,h}(d)}^{\infty} |Ef_{i,n}(\theta_0)f_{j,n}(\theta_0)'| \\
&\leq O\left(\bar{L}^{\frac{\nu-1}{\nu}}\right) \cdot \sum_{d=1}^{\infty} d[\alpha_{1,1}(d)]^{\frac{\delta}{2+\delta}}.
\end{aligned}$$

Together with Assumption B.3(i), the above inequality leads to

$$\begin{aligned}
&\frac{1}{n}\sum_{i\in\mathcal{G}_{g,n}}\sum_{j\in\mathcal{G}_{h,n}} E|f_{i,n}(\theta_0)f_{j,n}(\theta_0)| \\
&\leq \frac{1}{G}O\left(\frac{1}{\bar{L}^\nu}\right) \cdot \sum_{d=1}^{\infty} d[\alpha_{1,1}(d)]^{\frac{\delta}{2+\delta}} = o(1),
\end{aligned}$$

which is the desired result in (B.18). The proof in the case of a ϕ-mixing random field follows by the same arguments as above, but replacing (B.19) by the following ϕ-mixing inequality (e.g., Hall and Heyde (1980, p. 277)):

$$|Ef_{i,n}(\theta_0)f_{j,n}(\theta_0)| \leq \Delta \cdot [\phi_{1,1}(d)]^{1/2}.$$

Next, we prove the result in (B.17). It is straightforward to check that $\cup_{g=1}^{G}\{\lambda_g t_g f_{i,n}^{g}(\theta_0) : i \in \mathcal{G}_{g,n}\}$ is a measurable transformation of the original process $\{f(Y_{i,n},\theta_0), i \in B_n\}$, so it also satisfies the α-mixing [ϕ-mixing] conditions in Assumption B.3 [Assumption B.4]. Then the (joint) CLT for the α-mixing fields $\cup_{g=1}^{G}\{\lambda_g t_g f_{i,n}^{g}(\theta_0) : i \in \mathcal{G}_{g,n}\}$ follows from Corollary 1 in JP (2009) by setting their non-random constants $c_{i,n}$ to 1 if the following conditions are satisfied for some $\delta > 0$:

$$\lim_{M\to\infty} \sup_{n} \sup_{i\in B_n} E[|f_{i,n}(\theta_0)|^{2+\delta} 1(|f_{i,n}(\theta_0)| > M)] = 0; \tag{B.21}$$

$$\lim\inf_{n\to\infty} \frac{\sigma_{n,t}^2}{n} > 0; \tag{B.22}$$

$$\sum_{d=1}^{\infty} \alpha_{1,1}(d) d^{\frac{\nu(2+\delta)}{\delta}-1} < \infty \tag{B.23}$$

It is not difficult to check that (B.21) is implied by our Assumption B.5. Also, (B.22) is proved by (B.16). To show that our Assumption B.3(i) implies (B.23), note that the ratio of the summands in Assumption B.3(i) and (B.23) is equal to

$$\frac{\alpha_{1,1}(d) d^{\frac{\nu(2+\delta)}{\delta}-1}}{d^{\nu-1}\alpha_{1,1}(d)^{\frac{\delta}{2+\delta}}} = \left(d^{\nu}\alpha_{1,1}(d)^{\frac{\delta}{2+\delta}}\right)^{\frac{2}{\delta}}. \tag{B.24}$$

If Assumption B.3(i) holds, we check that

$$d^{\nu}\alpha_{1,1}(d)^{\frac{\delta}{2+\delta}} \to 0,$$

so the ratio in (B.24) converges to zero as well. Thus given $\epsilon > 0$, our Assumption B.3(i) implies that there exists $m \in \mathbb{N}$ such that

$$\sum_{d=m}^{\infty} \alpha_{1,1}(d) d^{[\nu(2+\delta)/\delta]-1} < \epsilon \cdot \sum_{d=m}^{\infty} d^{\nu-1}\alpha_{1,1}(d)^{\delta/(2+\delta)} < \infty,$$

which leads to the desired result in (B.23). For the ϕ-mixing case, we check that our Assumptions B.4 and B.5 directly satisfy Assumptions 4 and 2, respectively, for Theorem 1 in JP (2009), which leads to the (joint) CLT for the ϕ-mixing random field $\cup_{g=1}^{G}\{\lambda_g t_g f_{i,n}^{g}(\theta_0) : i \in \mathcal{G}_{g,n}\}$. This completes the proof of (B.17). ■

B.2 Proof of Proposition 6

Proof. Let $U\Sigma V'$ be a singular value decomposition (SVD) of Γ_Λ. Also, define $B^U_{m,g} = U'B_{m,g}$, $\bar{B}^U_m = U'\bar{B}_m$, and

$$\tilde{\mathbb{D}}^U_\infty = U'\tilde{\mathbb{D}}_\infty U = \left(\begin{array}{cc} \underset{d\times d}{\tilde{\mathbb{D}}^U_{11}} & \underset{d\times q}{\tilde{\mathbb{D}}^U_{12}} \\ \underset{q\times d}{\tilde{\mathbb{D}}^U_{21}} & \underset{q\times q}{\tilde{\mathbb{D}}^U_{22}} \end{array} \right). \tag{B.25}$$

Using the similar argument in the proof of Proposition 8-(a), we can show that

$$\sqrt{n}(\hat{\theta}_2 - \theta_0) \overset{d}{\to} -VA^{-1}\sqrt{G}\left(\bar{B}^U_d - \tilde{\mathbb{D}}^U_{12}\left[\tilde{\mathbb{D}}^U_{22}\right]^{-1}\bar{B}^U_q\right) \text{ and } \hat{\Omega}(\hat{\theta}_1) \overset{d}{\to} \Lambda\tilde{\mathbb{D}}_\infty\Lambda',$$

where holds jointly. It then follows that

$$\begin{aligned} F_{\hat{\Omega}(\hat{\theta}_1)}(\hat{\theta}_2) &= \frac{1}{p}\cdot\left[R(\hat{\theta}_2-\theta_0)\right]'\left(R\widehat{var}_{\hat{\Omega}(\hat{\theta}_1)}(\hat{\theta}_2)R'\right)^{-1}\left[R(\hat{\theta}_2-\theta_0)\right] \\ &\overset{d}{\to} \frac{G}{p}\cdot(\bar{B}^U_d - \tilde{\mathbb{D}}^U_{12}\left[\tilde{\mathbb{D}}^U_{22}\right]^{-1}\bar{B}^U_q)'A^{-1\prime}V'R'\left[R\left(\Gamma'\left(\Lambda\tilde{\mathbb{D}}_\infty\Lambda'\right)^{-1}\Gamma\right)^{-1}R'\right]^{-1} \\ &\quad\times RVA^{-1}(\bar{B}^U_d - \tilde{\mathbb{D}}^U_{12}\left[\tilde{\mathbb{D}}^U_{22}\right]^{-1}\bar{B}^U_q) \\ &= \frac{G}{p}\cdot(\bar{B}^U_d - \tilde{\mathbb{D}}^U_{12}\left[\tilde{\mathbb{D}}^U_{22}\right]^{-1}\bar{B}^U_q)'A^{-1\prime}V'R'\cdot\left\{R\left[\Gamma'\left(\Lambda'\right)^{-1}U\left(U'\tilde{\mathbb{D}}_\infty U\right)^{-1}U'\Lambda^{-1}\Gamma\right]^{-1}R'\right\}^{-1} \\ &\quad\times RVA^{-1}(\bar{B}^U_d - \tilde{\mathbb{D}}^U_{12}\left[\tilde{\mathbb{D}}^U_{22}\right]^{-1}\bar{B}^U_d) \\ &= \frac{G}{p}\cdot(\bar{B}^U_d - \tilde{\mathbb{D}}^U_{12}\left[\tilde{\mathbb{D}}^U_{22}\right]^{-1}\bar{B}^U_q)'A^{-1\prime}V'R'\left\{RVA^{-1}\tilde{\mathbb{D}}^U_{11\cdot 2}A^{-1\prime}V'R'\right\}^{-1} \\ &\quad\times RVA^{-1}(\bar{B}^U_d - \tilde{\mathbb{D}}^U_{12}\left[\tilde{\mathbb{D}}^U_{22}\right]^{-1}\bar{B}^U_d). \end{aligned}$$

Let $\tilde{U}_{p\times p}\tilde{\Sigma}\tilde{V}'_{d\times d}$ be a SVD of RVA^{-1}, where $\tilde{\Sigma} = (\tilde{A}_{p\times p}, O_{p\times(d-p)})$. Also, define

$$\mathbb{V} = \left(\begin{array}{cc} \tilde{V}_{d\times d} & O \\ O & I_{q\times q} \end{array} \right),$$

and

$$\breve{\mathbb{D}} = \left(\begin{array}{cc} \breve{\mathbb{D}}_{11} & \breve{\mathbb{D}}_{12} \\ \breve{\mathbb{D}}_{21} & \breve{\mathbb{D}}_{22} \end{array} \right) = \left(\begin{array}{cc} \tilde{V}_{d\times d} & O \\ O & I_q \end{array} \right)'\left(\begin{array}{cc} \tilde{\mathbb{D}}^U_{11} & \tilde{\mathbb{D}}^U_{12} \\ \tilde{\mathbb{D}}^U_{21} & \tilde{\mathbb{D}}^U_{22} \end{array} \right)\left(\begin{array}{cc} \tilde{V}_{d\times d} & O \\ O & I_q \end{array} \right) = \mathbb{V}'\tilde{\mathbb{D}}^U_\infty\mathbb{V}.$$

Then,

$$\begin{aligned} \breve{\mathbb{D}} &= \frac{1}{G}\sum_{g=1}^{G}\mathbb{V}'U'(B_{m,g}-\bar{B}_m)(B_{m,g}-\bar{B}_m)'\mathbb{V}U + \left(\begin{array}{c} \tilde{V}'\beta_{\tilde{W}} \\ I_q \end{array} \right)\bar{B}^U_q(\bar{B}^U_q)'\left(\begin{array}{c} \tilde{V}'\beta_{\tilde{W}} \\ I_q \end{array} \right)' \\ &\overset{d}{=} \frac{1}{G}\sum_{g=1}^{G}(B_{m,g}-\bar{B}_m)(B_{m,g}-\bar{B}_m)' + \left(\begin{array}{c} \tilde{V}'\beta_{\tilde{W}} \\ I_q \end{array} \right)\bar{B}_q\bar{B}'_q\left(\begin{array}{c} \tilde{V}'\beta_{\tilde{W}} \\ I_q \end{array} \right)', \end{aligned}$$

which implies that

$$\breve{\mathbb{D}}_{11} := \left(\begin{array}{cc} \breve{\mathbb{D}}_{pp} & \breve{\mathbb{D}}_{p,d-p} \\ \breve{\mathbb{D}}_{d-p,p} & \breve{\mathbb{D}}_{d-p,d-p} \end{array} \right) \overset{d}{=} \frac{1}{G}\sum_{g=1}^{G}(B_{d,g}-\bar{B}_d)(B_{d,g}-\bar{B}_d)' + \left(\tilde{V}'\beta_{\tilde{W}}\right)\bar{B}_q\bar{B}'_q\left(\tilde{V}'\beta_{\tilde{W}}\right)', \tag{B.26}$$

and

$$\breve{\mathbb{D}}_{12} := \left(\begin{array}{c} \breve{\mathbb{D}}_{pq} \\ \breve{\mathbb{D}}_{d-p,q} \end{array} \right) \overset{d}{=} \frac{1}{G}\sum_{g=1}^{G}(B_{d,g} - \bar{B}_d)(B_{q,g} - \bar{B}_q)' + \left(\tilde{V}'\beta_{\tilde{W}}\right)\bar{B}_q\bar{B}_q'. \tag{B.27}$$

Now

$$\begin{aligned} & F_{\hat{\Omega}(\hat{\theta}_1)}(\hat{\theta}_2) \xrightarrow{d} \frac{G}{p} \cdot (\bar{B}_d^U - \breve{\mathbb{D}}_{12}\breve{\mathbb{D}}_{22}^{-1}\bar{B}_q^U)'\tilde{V}\tilde{\Sigma}'\tilde{U}' \left\{\tilde{U}\tilde{\Sigma}\tilde{V}'\breve{\mathbb{D}}_{11\cdot 2}\tilde{V}\tilde{\Sigma}'\tilde{U}'\right\}^{-1} \tilde{U}\tilde{\Sigma}\tilde{V}'(\bar{B}_d^U - \breve{\mathbb{D}}_{12}\breve{\mathbb{D}}_{22}^{-1}\bar{B}_q^U) \\ & = \frac{G}{p} \cdot (\bar{B}_d^U - \breve{\mathbb{D}}_{12}\breve{\mathbb{D}}_{22}^{-1}\bar{B}_q^U)'\tilde{V}\tilde{\Sigma}' \cdot \left\{\tilde{\Sigma}\tilde{V}'\breve{\mathbb{D}}_{11\cdot 2}\tilde{V}\tilde{\Sigma}'\right\}^{-1} \cdot \tilde{\Sigma}\tilde{V}'(\bar{B}_d^U - \breve{\mathbb{D}}_{12}\breve{\mathbb{D}}_{22}^{-1}\bar{B}_q^U) \\ & \overset{d}{=} \frac{G}{p} \cdot \left[\bar{B}_p - \breve{\mathbb{D}}_{pq}\breve{\mathbb{D}}_{qq}^{-1}\bar{B}_q\right]' \tilde{A}' \left\{\tilde{A}\left(\breve{\mathbb{D}}_{pp} - \breve{\mathbb{D}}_{pq}\breve{\mathbb{D}}_{qq}^{-1}\breve{\mathbb{D}}_{qp}\right)\tilde{A}'\right\}^{-1} \tilde{A}\left[\bar{B}_p - \breve{\mathbb{D}}_{pq}\breve{\mathbb{D}}_{qq}^{-1}\bar{B}_q\right] \\ & \overset{d}{=} \frac{G}{p} \cdot \left[\bar{B}_p - \breve{\mathbb{D}}_{pq}\breve{\mathbb{D}}_{qq}^{-1}\bar{B}_q\right]' \left(\breve{\mathbb{D}}_{pp} - \breve{\mathbb{D}}_{pq}\breve{\mathbb{D}}_{qq}^{-1}\breve{\mathbb{D}}_{qp}\right)^{-1} \left[\bar{B}_p - \breve{\mathbb{D}}_{pq}\breve{\mathbb{D}}_{qq}^{-1}\bar{B}_q\right], \end{aligned} \tag{B.28}$$

where $\breve{\mathbb{D}}_{pq}$, $\breve{\mathbb{D}}_{qq}$, and $\breve{\mathbb{D}}_{qp}$ in the last two equalities are understood to equal the corresponding components on the right hand sides of (B.26) and (B.27). Here, we have abused the notation a little bit. We have

$$\left(\begin{array}{cc} \breve{\mathbb{D}}_{pp} & \breve{\mathbb{D}}_{pq} \\ \breve{\mathbb{D}}_{pq}' & \breve{\mathbb{D}}_{qq} \end{array} \right) \overset{d}{=} \mathbb{V}_{p+q,p+q} = \left(\begin{array}{cc} \bar{\mathbb{S}}_{pp} & \bar{\mathbb{S}}_{pq} \\ \bar{\mathbb{S}}_{pq}' & \bar{\mathbb{S}}_{qq} \end{array} \right) + \tilde{w}\bar{B}_q\bar{B}_q'\tilde{w}' \tag{B.29}$$

for

$$\tilde{w} = \left(\begin{array}{c} \tilde{\beta}_{\tilde{W}}^p \\ I_q \end{array} \right) \in \mathbb{R}^{(p+q)\times q}.$$

Direct calculations show that the representation in (B.28) is numerically identical to

$$\frac{1}{p} \cdot \left[G \left(\begin{array}{c} \bar{B}_p \\ \bar{B}_q \end{array} \right)' \left(\begin{array}{cc} \mathbb{V}_{pp} & \mathbb{V}_{pq} \\ \mathbb{V}_{pq}' & \mathbb{V}_{qq} \end{array} \right)^{-1} \left(\begin{array}{c} \bar{B}_p \\ \bar{B}_q \end{array} \right) - G \cdot \bar{B}_q'\mathbb{S}_{qq}^{-1}\bar{B}_q \right],$$

which completes the proof of part (a). To prove part (b), we repeat the same argument in the proof of Proposition 8-(b) and obtain that

$$\begin{aligned} \sqrt{n}g_n(\hat{\theta}_2) &= \frac{1}{\sqrt{G}}\sum_{g=1}^{G}\left(\frac{1}{\sqrt{\bar{L}}}\sum_{i=1}^{L_g} f_i^g(\hat{\theta}_2)\right) + o_p(1) \\ &\xrightarrow{d} \Lambda\sqrt{G}\left(UU'\bar{B}_m - \Gamma_\Lambda\left[\Gamma_\Lambda'\tilde{\mathbb{D}}_\infty^{-1}\Gamma_\Lambda\right]^{-1}\Gamma_\Lambda'\tilde{\mathbb{D}}_\infty^{-1}\bar{B}_m\right) \\ &\overset{d}{=} \Lambda\sqrt{G}\left[U\bar{B}_m^U - \Gamma_\Lambda V A^{-1}\left(\bar{B}_d^U - \tilde{\mathbb{D}}_{12}^U\left[\tilde{\mathbb{D}}_{22}^U\right]^{-1}\bar{B}_q^U\right)\right] \end{aligned}$$

with $\tilde{\mathbb{D}}_{12}^U$ and $\tilde{\mathbb{D}}_{22}^U$ given in (B.25). Therefore, we have

$$
\begin{aligned}
& J(\hat{\theta}_2) = n g_n(\hat{\theta}_2)' \left(\hat{\Omega}(\hat{\theta}_1)\right)^{-1} g_n(\hat{\theta}_2) \\
& \xrightarrow{d} G\left\{U\bar{B}_m^U - \Gamma_\Lambda V A^{-1}\left(\bar{B}_d^U - \tilde{\mathbb{D}}_{12}^U\left[\tilde{\mathbb{D}}_{22}^U\right]^{-1}\bar{B}_q^U\right)\right\}' \times \Lambda'\left(\Lambda\tilde{\mathbb{D}}_\infty\Lambda'\right)^{-1}\Lambda \\
& \times \left\{U\bar{B}_m^U - \Gamma_\Lambda V A^{-1}\left(\bar{B}_d^U - \tilde{\mathbb{D}}_{12}^U\left[\tilde{\mathbb{D}}_{22}^U\right]^{-1}\bar{B}_q^U\right)\right\} \\
& = G\left\{\bar{B}_m^U - U'\Gamma_\Lambda V A^{-1}\left(\bar{B}_d^U - \tilde{\mathbb{D}}_{12}^U\left[\tilde{\mathbb{D}}_{22}^U\right]^{-1}\bar{B}_q^U\right)\right\}' U'\tilde{\mathbb{D}}_\infty^{-1}U \\
& \times \left\{\bar{B}_m^U - U'\Gamma_\Lambda V A^{-1}\left(\bar{B}_d^U - \tilde{\mathbb{D}}_{12}^U\left[\tilde{\mathbb{D}}_{22}^U\right]^{-1}\bar{B}_q^U\right)\right\},
\end{aligned}
$$

which is continued in

$$
\begin{aligned}
& = G\left\{\bar{B}_m^U - \left[\begin{array}{c} I_{d\times d} \\ O_{q\times d}\end{array}\right]\left(\bar{B}_d^U - \tilde{\mathbb{D}}_{12}^U\left[\tilde{\mathbb{D}}_{22}^U\right]^{-1}\bar{B}_q^U\right)\right\}'\left[\tilde{\mathbb{D}}_\infty^U\right]^{-1} \\
& \times \left\{\bar{B}_m^U - \left[\begin{array}{c} I_{d\times d} \\ O_{q\times d}\end{array}\right]\left(\bar{B}_d^U - \tilde{\mathbb{D}}_{12}^U\left[\tilde{\mathbb{D}}_{22}^U\right]^{-1}\bar{B}_q^U\right)\right\} \\
& = G\left(\begin{array}{c} \tilde{\mathbb{D}}_{12}^U\left[\tilde{\mathbb{D}}_{22}^U\right]^{-1}\bar{B}_q^U \\ \bar{B}_q^U \end{array}\right)'\left[\tilde{\mathbb{D}}_\infty^U\right]^{-1}\left(\begin{array}{c} \tilde{\mathbb{D}}_{12}^U\left[\tilde{\mathbb{D}}_{22}^U\right]^{-1}\bar{B}_q^U \\ \bar{B}_q^U \end{array}\right) \\
& = G(\bar{B}_q^U)'\left[\tilde{\mathbb{D}}_{22}^U\right]^{-1}\bar{B}_q^U \overset{d}{=} G\cdot\bar{B}_q'\mathbb{S}_{22}^{-1}\bar{B}_q
\end{aligned}
$$

where the second last equality follows from straightforward calculations. The joint convergence can be proved easily. ■

B.3 Proof of Proposition 8

Proof. Let $U\Sigma V'$ be a singular value decomposition (SVD) of Γ_Λ. By construction, $U'U = UU' = I_m$, $V'V = V'V = I_d$, and

$$\Sigma = \left[\begin{array}{c} A_{d\times d} \\ O_{q\times d} \end{array} \right],$$

where A is a diagonal matrix. Then,

$$\begin{aligned} R\left[\Gamma_\Lambda'\bar{\mathbb{S}}^{-1}\Gamma_\Lambda\right]^{-1}\Gamma_\Lambda'\bar{\mathbb{S}}^{-1}\bar{B}_m &= R\left[V\Sigma'U'\bar{\mathbb{S}}^{-1}U\Sigma V'\right]^{-1}V\Sigma'U'\bar{\mathbb{S}}^{-1}\bar{B}_m \\ &= RV\left[\Sigma'(U'\bar{\mathbb{S}}^{-1}U)\Sigma V'\right]^{-1}\Sigma'\left[U'\bar{\mathbb{S}}^{-1}U\right](U'\bar{B}_m) \\ &\overset{d}{=} RV\left[\Sigma'\bar{\mathbb{S}}^{-1}\Sigma\right]^{-1}\Sigma'\bar{\mathbb{S}}^{-1}\bar{B}_m, \end{aligned}$$

where the distributional equivalence holds by a rotation invariance property of $[\bar{B}_m, \bar{\mathbb{S}}^{-1}]$. Denoting

$$[\bar{\mathbb{S}}^{-1}]_{m\times m} = \left(\begin{array}{cc} \underset{d\times d}{\bar{\mathbb{S}}^{dd}} & \underset{d\times q}{\bar{\mathbb{S}}^{dq}} \\ \underset{q\times q}{\bar{\mathbb{S}}^{qd}} & \underset{q\times d}{\bar{\mathbb{S}}^{qq}} \end{array} \right),$$

we have

$$\begin{aligned} RV\left[\Sigma'\bar{\mathbb{S}}^{-1}\Sigma\right]^{-1}\Sigma'\bar{\mathbb{S}}^{-1}\bar{B}_m &= RVA^{-1}\left(\bar{\mathbb{S}}^{dd}\right)^{-1}\left(A'\right)^{-1}A'\left(\bar{\mathbb{S}}^{dd}, \bar{\mathbb{S}}^{dq}\right)\bar{B}_m \\ &= RVA^{-1}\left[\begin{array}{cc} I_d & \left(\bar{\mathbb{S}}^{dd}\right)^{-1}\bar{\mathbb{S}}^{dq} \end{array}\right]\bar{B}_m \\ &= RVA^{-1}\left[\bar{B}_d - \bar{\mathbb{S}}_{dq}\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q\right]. \end{aligned}$$

where the last equation follows by the partitioned inverse formula that $\bar{\mathbb{S}}^{dq} = -\bar{\mathbb{S}}^{dd}\bar{\mathbb{S}}_{dq}\bar{\mathbb{S}}_{qq}^{-1}$. Similarly, we obtain

$$\begin{aligned} R\left[\Gamma_\Lambda'\bar{\mathbb{S}}^{-1}\Gamma_\Lambda\right]^{-1}R' &\overset{d}{=} RV\left[\Sigma'\bar{\mathbb{S}}^{-1}\Sigma\right]^{-1}V'R' \\ &= RVA^{-1}\left(\bar{\mathbb{S}}^{dd}\right)^{-1}(A')^{-1}V'R' \\ &= RVA^{-1}\bar{\mathbb{S}}_{dd\cdot q}(A')^{-1}V'R', \end{aligned}$$

where $\bar{\mathbb{S}}_{dd\cdot q} = \bar{\mathbb{S}}_{dd} - \bar{\mathbb{S}}_{dq}\bar{\mathbb{S}}_{qq}^{-1}\bar{\mathbb{S}}_{qd}$. Therefore,

$$\begin{aligned} F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) \xrightarrow{d} \mathbb{F}_{2\infty} &\overset{d}{=} \frac{G}{p}\cdot\left[RVA^{-1}\left(\bar{B}_d - \bar{\mathbb{S}}_{dq}\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q\right)\right]' \\ &\times\left(RVA^{-1}\bar{\mathbb{S}}_{dd\cdot q}(A')^{-1}V'R'\right)^{-1}\left[RVA^{-1}\left(\bar{B}_d - \bar{\mathbb{S}}_{dq}\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q\right)\right]. \end{aligned} \tag{B.30}$$

Let $\tilde{U}_{p\times p}\tilde{\Sigma}\tilde{V}_{d\times d}'$ be a SVD of RVA^{-1}, where $\tilde{\Sigma} = (\tilde{A}_{p\times p}, O_{p\times(d-p)})$. By definition, $\tilde{V}$ is the matrix of eigenvectors of $(RVA^{-1})'(RVA^{-1})$. Then,

$$\begin{aligned} \mathbb{F}_{2\infty} &\overset{d}{=} \frac{G}{p}\left[\bar{B}_d - \bar{\mathbb{S}}_{dq}\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q\right]'\tilde{V}\tilde{\Sigma}'\tilde{U}'\left[\tilde{U}\tilde{\Sigma}\tilde{V}'\bar{\mathbb{S}}_{dd\cdot q}\tilde{V}\tilde{\Sigma}'\tilde{U}'\right]^{-1}\tilde{U}\tilde{\Sigma}\tilde{V}'\left[\bar{B}_d - \bar{\mathbb{S}}_{dq}\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q\right] \\ &= \frac{G}{p}\left[\bar{B}_d - \bar{\mathbb{S}}_{dq}\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q\right]'\tilde{V}\tilde{\Sigma}'\cdot\left[\tilde{\Sigma}\tilde{V}'\bar{\mathbb{S}}_{dd\cdot q}\tilde{V}\tilde{\Sigma}'\right]^{-1}\cdot\tilde{\Sigma}\tilde{V}'\left[\bar{B}_d - \bar{\mathbb{S}}_{dq}\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q\right] \\ &= \frac{G}{p}\left[\tilde{V}'\bar{B}_d - \tilde{V}'\bar{\mathbb{S}}_{dq}\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q\right]'\tilde{\Sigma}'\cdot\left[\tilde{\Sigma}\tilde{V}'\bar{\mathbb{S}}_{dd\cdot q}\tilde{V}\tilde{\Sigma}'\right]^{-1}\cdot\tilde{\Sigma}\left[\tilde{V}'\bar{B}_d - \tilde{V}'\bar{\mathbb{S}}_{dq}\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q\right]. \end{aligned}$$

Note that the rotational invariance property of the standard normal vector implies

$$\left[\tilde{V}'\bar{B}_d, \tilde{V}'\bar{\mathbb{S}}_{dq}, \bar{\mathbb{S}}_{qq}^{-1}, \tilde{V}'\bar{\mathbb{S}}_{dd\cdot q}\tilde{V}\right] \overset{d}{=} \left[\bar{B}_d, \bar{\mathbb{S}}_{dq}, \bar{\mathbb{S}}_{qq}^{-1}, \bar{\mathbb{S}}_{dd\cdot q}\right],$$

which leads us to obtain

$$\begin{aligned}\mathbb{F}_{2\infty} &\overset{d}{=} \frac{G}{p}\left[\bar{B}_d - \bar{\mathbb{S}}_{dq}\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q\right]'\tilde{\Sigma}' \cdot \left[\tilde{\Sigma}\bar{\mathbb{S}}_{dd\cdot q}\tilde{\Sigma}'\right]^{-1} \cdot \tilde{\Sigma}\left[\bar{B}_d - \bar{\mathbb{S}}_{dq}\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q\right] \\ &= \frac{G}{p} \cdot \left[\bar{B}_p - \bar{\mathbb{S}}_{pq}\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q\right]'\tilde{A}'\left\{\tilde{A}\left(\bar{\mathbb{S}}_{pp\cdot q}\right)\tilde{A}'\right\}^{-1}\tilde{A}\left[\bar{B}_p - \bar{\mathbb{S}}_{pq}\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q\right] \\ &= \frac{G}{p} \cdot \left(\bar{B}_p - \bar{\mathbb{S}}_{pq}\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q\right)'\bar{\mathbb{S}}_{pp\cdot q}^{-1}\left(\bar{B}_p - \bar{\mathbb{S}}_{pq}\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q\right)',\end{aligned}$$

as desired. The proof of part (b) is similar, so we omit the details. To prove part (c), we use the same arguments and obtain

$$\begin{aligned}J(\hat{\theta}_2^c) \xrightarrow{d} \mathbb{J}_\infty &:= G \cdot \left\{U'\bar{B}_m - U'\Gamma_\Lambda\left(\Gamma_\Lambda'\bar{\mathbb{S}}^{-1}\Gamma_\Lambda\right)^{-1}\Gamma_\Lambda'\bar{\mathbb{S}}^{-1}\bar{B}_m\right\}' \times U'\bar{\mathbb{S}}^{-1}U \\ &\times \left\{U'\bar{B}_m - U'\Gamma_\Lambda\left(\Gamma_\Lambda'\bar{\mathbb{S}}^{-1}\Gamma_\Lambda\right)^{-1}\Gamma_\Lambda'\bar{\mathbb{S}}^{-1}\bar{B}_m\right\} \\ &\overset{d}{=} G\left\{\bar{B}_m - U'\Gamma_\Lambda VA^{-1}\left[\bar{B}_d - \bar{\mathbb{S}}_{dq}\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q\right]\right\}'\bar{\mathbb{S}}^{-1} \\ &\times \left\{\bar{B}_m - U'\Gamma_\Lambda VA^{-1}\left[\bar{B}_d - \bar{\mathbb{S}}_{dq}\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q\right]\right\},\end{aligned} \tag{B.31}$$

which is continued in

$$\begin{aligned}&= G\left\{\bar{B}_m - \begin{bmatrix} I_{d\times d} \\ O_{q\times d}\end{bmatrix}\left(\bar{B}_d - \bar{\mathbb{S}}_{dq}\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q\right)\right\}'\bar{\mathbb{S}}^{-1} \\ &\times \left\{\bar{B}_m - \begin{bmatrix} I_{d\times d} \\ O_{q\times d}\end{bmatrix}\left(\bar{B}_d - \bar{\mathbb{S}}_{dq}\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q\right)\right\} \\ &= G\begin{pmatrix}\bar{\mathbb{S}}_{dq}\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q \\ \bar{B}_q\end{pmatrix}'\bar{\mathbb{S}}^{-1}\begin{pmatrix}\bar{\mathbb{S}}_{dq}\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q \\ \bar{B}_q\end{pmatrix} \\ &= G \cdot \bar{B}_q'\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q,\end{aligned}$$

where the last equality follows from straightforward calculations. Lastly, it is not difficult to check that the convergence results in (B.28) and (B.31) hold jointly. This completes the proof. ■

B.4 Asymptotics for LM and QLR statistics

We construct the Quasi-Likelihood Ratio (QLR) and the Lagrangian Multiplier (LM) statistics in the GMM setting and investigate their fixed-G limits. Define the restricted and centered two-step estimator $\hat{\theta}_2^{c,r}$

$$\hat{\theta}_2^{c,r} := \arg\min_{\theta\in\Theta} g_n(\theta)' \left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} g_n(\theta) \text{ such that } R\theta = r.$$

The QLR statistic is then given by

$$LR_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c, \hat{\theta}_2^{c,r}) := \frac{n}{p}\left\{g_n(\hat{\theta}_2^{c,r})' \left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} g_n(\hat{\theta}_2^{c,r}) - g_n(\hat{\theta}_2^c)' \left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} g_n(\hat{\theta}_2^c)\right\}.$$

To define LM or score statistic in the GMM setting, let $S_{\hat{\Omega}^c(\cdot)}(\theta)$ be the gradient of the GMM criterion function $\hat{\Gamma}(\theta)'[\hat{\Omega}^c(\cdot)]^{-1}g_n(\theta)$, then the GMM score test statistic is given by

$$LM_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^{c,r}) := \frac{n}{p}\left[S_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^{c,r})\right]' \left\{\hat{\Gamma}(\hat{\theta}_2^{c,r})' \left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} \hat{\Gamma}(\hat{\theta}_2^{c,r})\right\}^{-1} \left[S_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^{c,r})\right].$$

In the definition of all three types of the GMM test statistics, we plug the first-step estimator $\hat{\theta}_1$ into $\hat{\Omega}^c(\cdot)$, but Lemma 7 indicates that replacing $\hat{\theta}_1$ with any $\sqrt{n}$-consistent estimator (e.g., $\hat{\theta}_2$ and $\hat{\theta}_2^c$) does not affect the fixed-G asymptotic results. This contrasts with the fixed-G asymptotics for the uncentered two-step estimator $\hat{\theta}_2$. Lastly, using the same multiplicative factors as in Section 4 in the main text, we can also construct the modified QLR and LM statistics.

Proposition B.4 *Let Assumptions 1∼5 hold. Then,,*

(*a*) $LR_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c, \hat{\theta}_2^{c,r}) = F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) + o_p(1)$;

(*b*) $LM_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^{c,r}) = F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) + o_p(1)$.

(*c*) *The modified QLR and LM statistics all converge in distribution to* $\mathcal{F}_{p,G-p-q}$.

Proposition B.4 shows that QLR and LM types of test statistics are asymptotically equivalent to the Wald statistics $F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)$. This also implies that all three types of test statistics share the same fixed-G limit as given in the main text. Similar results are obtained by Sun (2014) and Hwang and Sun (2017), which focus on the two-step GMM estimation and HAR inference in a time series setting.

B.4.1 Proof of Proposition B.4

Proof. To prove (a), it suffices to show the asymptotic equivalence between $\mathrm{LR}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c, \hat{\theta}_2^{c,r})$ and $F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)$ holds under the fixed-G asymptotics. Recall that the restricted two-step GMM estimator $\hat{\theta}_2^{c,r}$ minimizes

$$g_n(\theta)' \left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} g_n(\theta)/2 + \lambda_n'(R\theta - r).$$

The first order conditions are

$$\hat{\Gamma}'(\hat{\theta}_2^{c,r}) \left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} g_n(\hat{\theta}_2^{c,r}) + R'\lambda_n = 0 \text{ and } R\hat{\theta}_2^{c,r} = r.$$

Using a Taylor expansion and Assumption 3, we can combine two FOC's to get

$$\begin{aligned}\sqrt{n}(\hat{\theta}_2^{c,r} - \theta_0) = & -\Phi^{-1}\Gamma' \left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} \sqrt{n}g_n(\theta_0) \\ & - \Phi^{-1}R' \left(R\Phi^{-1}R'\right)^{-1} R\Phi^{-1}\Gamma' \left[\hat{\Omega}_n^c(\hat{\theta}_1)\right]^{-1} \sqrt{n}g_n(\theta_0) + o_p(1)\end{aligned} \tag{B.32}$$

and

$$\sqrt{n}\lambda_n = -(R\Phi^{-1}R')^{-1}R\Phi^{-1}\Gamma'\left[\hat{\Omega}_n^c(\hat{\theta}_1)\right]^{-1}\sqrt{n}g_n(\theta_0) + o_p(1), \tag{B.33}$$

where $\Phi := \Gamma'\left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1}\Gamma$. Subtracting (B.32) from (13), we have

$$\begin{aligned}\sqrt{n}(\hat{\theta}_2^{c,r} - \hat{\theta}_2^c) &= -\Phi^{-1}R'\left(R\Phi^{-1}R'\right)^{-1}R\Phi^{-1}\Gamma'\left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1}\sqrt{n}g_n(\theta_0) + o_p(1) &\text{(B.34)}\\ &= O_p(1), &\text{(B.35)}\end{aligned}$$

where the equation (B.35) comes from Lemma 7-(b) and Assumption 1-ii). Thus, we can approximate $g_n(\hat{\theta}_2^{c,r})$ around $\hat{\theta}_2^c$ as

$$g_n'(\hat{\theta}_2^{c,r}) = g_n'(\hat{\theta}_2^c) - (\hat{\theta}_2^{c,r} - \hat{\theta}_2^c)'\hat{\Gamma}'(\hat{\theta}_2^c) + o_p(n^{-1/2}).$$

Plugging this into the definition of $LR_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c, \hat{\theta}_2^{c,r})$,

$$\begin{aligned}LR_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c, \hat{\theta}_2^{c,r}) = \frac{n}{p}\Big\{&(\hat{\theta}_2^{c,r} - \hat{\theta}_2^c)'\hat{\Gamma}'(\hat{\theta}_2^c)\left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1}\hat{\Gamma}(\hat{\theta}_2^c)(\hat{\theta}_2^{c,r} - \hat{\theta}_2^c) &\text{(B.36)}\\ &-2g_n'(\hat{\theta}_2^c)\left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1}\hat{\Gamma}(\hat{\theta}_2^c)(\hat{\theta}_2^{c,r} - \hat{\theta}_2^c)\Big\} + o_p(1),\end{aligned}$$

where the last term in (B.36) is always zero from the FOC of $\hat{\theta}_2^c$. Combining (B.34) and (B.36), we have

$$\begin{aligned}LR_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c, \hat{\theta}_2^{c,r}) &= \frac{n}{p}(\hat{\theta}_2^{c,r} - \hat{\theta}_2^c)'\hat{\Gamma}'(\hat{\theta}_2^c)\left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1}\hat{\Gamma}(\hat{\theta}_2^c)(\hat{\theta}_2^{c,r} - \hat{\theta}_2^c) + o_p(1)\\ &= \frac{n}{p}\left[\Phi^{-1}R'\left(R\Phi^{-1}R'\right)^{-1}R\Phi^{-1}\Gamma'\left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1}\sqrt{n}g_n(\theta_0)\right]'\times\\ &\Phi\left[\Phi^{-1}R'\left(R\Phi^{-1}R'\right)^{-1}R\Phi^{-1}\Gamma'\left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1}\sqrt{n}g_n(\theta_0)\right] + o_p(1)\\ &= \frac{1}{p}\left[R\Phi^{-1}\Gamma'\left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1}\sqrt{n}g_n(\theta_0)\right]'\left(R\Phi^{-1}R'\right)^{-1}\\ &\times\left[R\Phi^{-1}\Gamma'\left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1}\sqrt{n}g_n(\theta_0)\right] + o_p(1)\\ &= F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) + o_p(1),\end{aligned}$$

as desired. To prove part (b), we show $LM_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^{c,r}) = LR_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c, \hat{\theta}_2^{c,r}) + o_p(1)$. From the first order condition of $\hat{\theta}_2^{c,r}$ and the equation (B.33), we expand the score vector by

$$\begin{aligned}\sqrt{n}S_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^{c,r}) &= \hat{\Gamma}(\hat{\theta}_2^{c,r})'\left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1}\sqrt{n}g_n(\hat{\theta}_2^{c,r}) = -R'\sqrt{n}\lambda_n\\ &= R'(R\Phi^{-1}R')^{-1}R\Phi^{-1}\Gamma'\left[\hat{\Omega}_n^c(\hat{\theta}_1)\right]^{-1}\sqrt{n}g_n(\theta_0) + o_p(1)\\ &= -\Phi\sqrt{n}(\hat{\theta}_2^{c,r} - \hat{\theta}_2^c) + o_p(1),\end{aligned}$$

and so

$$\begin{aligned}LM_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^{c,r}) &= n(\hat{\theta}_2^{c,r} - \hat{\theta}_2^c)'\Phi(\hat{\theta}_2^{c,r} - \hat{\theta}_2^c)/p + o_p(1)\\ &= LR_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c, \hat{\theta}_2^{c,r}) + o_p(1)\\ &= F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) + o_p(1),\end{aligned}$$

which leads the desired result. The proof of (c) directly follows from the results in (a), (b), and Proposition 8. ■

B.5 Asymptotically unbalanced sizes in clusters

We present the fixed-G limits of the centered CCE and corresponding test statistics, which is mentioned at Remark 11 in the main body of the paper, assuming the asymptotically unbalanced cluster sizes. We assume that the cluster size $L_g \to \infty$ as $n \to \infty$ such that

$$\frac{L_g}{n} \to \lambda_g > 0 \text{ for each } g = 1, \ldots, G. \tag{B.37}$$

By construction, $\sum_{g=1}^{G} \lambda_g = 1$ for all possible $\lambda_g \in (0,1)$. Also, Assumption 1-ii) guarantees that the total (scaled) sum of moment process satisfies

$$\begin{aligned} \sqrt{n} g_n(\theta_0) &= \sum_{g=1}^{G} \sqrt{\frac{L_g}{n}} \frac{1}{\sqrt{L_g}} \sum_{i=1}^{L_g} f_i^g(\theta_0) \\ &\xrightarrow{d} \sum_{g=1}^{G} \sqrt{\lambda_g} \Lambda_g B_{m,g} \sim N(0, \Omega), \end{aligned}$$

where $\Omega = \sum \lambda_g \Omega_g$. Keeping the same notations in the main text, we define an $m \times m$ random matrix

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

Also, we denote $\bar{\mathbb{M}}_{pp}, \bar{\mathbb{M}}_{pq}, \bar{\mathbb{M}}_{qp}$, and $\bar{\mathbb{M}}_{qq}$ as sub-matrices of $\bar{\mathbb{M}}$ in a similar manner in the main text.

Proposition B.5 *Let Assumptions 1–5 hold, where Assumption 1-iii) is replaced with (B.37). Define* $\bar{\mathbb{M}}_{pp\cdot q} = \bar{\mathbb{M}}_{pp} - \bar{\mathbb{M}}_{pq} \bar{\mathbb{M}}_{qq}^{-1} \bar{\mathbb{M}}_{qp}$*. Then,*

(a) $\hat{\Omega}^c(\tilde{\theta}) \xrightarrow{d} \Lambda \bar{\mathbb{M}} \Lambda'$ *for any* $\sqrt{n}$*-consistent estimator* $\tilde{\theta}$.;

(b) $F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) \xrightarrow{d} \tilde{\mathbb{F}}_{2\infty} := \tilde{\mathbb{F}}_{2\infty}(\lambda)$*, where*

$$\begin{aligned} \tilde{\mathbb{F}}_{2\infty}(\lambda) \overset{d}{=} \frac{1}{p} & \left[\left(\sum_{g=1}^{G} \sqrt{\lambda_g} B_{p,g} \right) - \bar{\mathbb{M}}_{pq} \bar{\mathbb{M}}_{qq}^{-1} \left(\sum_{g=1}^{G} \sqrt{\lambda_g} B_{q,g} \right) \right]' \bar{\mathbb{M}}_{pp\cdot q}^{-1} \\ & \times \left[\left(\sum_{g=1}^{G} \sqrt{\lambda_g} B_{p,g} \right) - \bar{\mathbb{M}}_{pq} \bar{\mathbb{M}}_{qq}^{-1} \left(\sum_{g=1}^{G} \sqrt{\lambda_g} B_{q,g} \right) \right]. \end{aligned}$$

(c) $t_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) \xrightarrow{d} \tilde{\mathbb{T}}_{2\infty} := \tilde{\mathbb{T}}_{2\infty}(\lambda)$*, where*

$$\tilde{\mathbb{T}}_{2\infty}(\lambda) = \frac{\sum_{g=1}^{G} \sqrt{\lambda_g} B_{p,g} - \bar{\mathbb{M}}_{pq} \bar{\mathbb{M}}_{qq}^{-1} \sum_{g=1}^{G} \sqrt{\lambda_g} B_{q,g}}{\sqrt{\bar{\mathbb{M}}_{pp\cdot q}}}.$$

B.5.1 Proof of Proposition B.5

Proof. We only prove parts (a) and (b), as the proof of part (c) can be done similarly. To prove (a), note that the centered CCE can be represented as:

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

For each $g = 1, ..., G$, we use the same arguments in the proof of Lemma 7 and obtain that

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

where the last equation follows from (B.37). Similarly,

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{L_g}\left(\frac{1}{n}\sum_{h=1}^{G}\sum_{i=1}^{L_h} f_i^h(\tilde{\theta})\right) &= \frac{\lambda_g}{\sqrt{n}}\sum_{h=1}^{G}\sum_{i=1}^{L_h} f_i^h(\tilde{\theta})(1+o_p(1) \\ &= \lambda_g\sum_{h=1}^{G}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{L_h}\left(f_i^g(\theta_0) + \frac{\partial f_i^g(\tilde{\theta})}{\partial\theta'}(\tilde{\theta}-\theta_0)\right)\right\}(1+o_p(1)) \\ &= \lambda_g\sum_{h=1}^{G}\left\{\sqrt{\lambda_h}\cdot\frac{1}{\sqrt{L_h}}\sum_{i=1}^{L_h} f_i^h(\theta_0) + \lambda_h\Gamma\sqrt{n}(\tilde{\theta}-\theta_0)\right\}(1+o_p(1)) \\ &= \left\{\lambda_g\sum_{h=1}^{G}\left(\sqrt{\lambda_h}\cdot\frac{1}{\sqrt{L_h}}\sum_{i=1}^{L_h} f_i^h(\theta_0)\right) + \lambda_g\Gamma\sqrt{n}(\tilde{\theta}-\theta_0)\right\}(1+o_p(1)),\end{aligned}$$

where the last equation holds from $\sum_{h=1}^{G}\lambda_h = 1$. It then follows that

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{L_g}\left(f_i^g(\tilde{\theta}) - \frac{1}{n}\sum_{h=1}^{G}\sum_{i=1}^{L_g} f_i^g(\tilde{\theta})\right) &= \Bigg\{\sqrt{\lambda_g}\cdot\frac{1}{\sqrt{L_g}}\sum_{i=1}^{L_g} f_i^g(\theta_0) \\ &\qquad -\lambda_g\sum_{h=1}^{G}\left(\sqrt{\lambda_h}\cdot\frac{1}{\sqrt{L_h}}\sum_{i=1}^{L_h} f_i^h(\theta_0)\right)\Bigg\}(1+o_p(1)) \\ &\xrightarrow{d} \sqrt{\lambda_g}B_{m,g} - \lambda_g\sum_{h=1}^{G}\sqrt{\lambda_h}B_{m,h},\end{aligned}$$

where the convergence holds jointly for $g = 1, \ldots, G$ by Assumption 1-ii). By continuous mapping theorem, this leads us to the desired result in (a). To prove part (b), note that the result in (a) and (B.37) imply

$$\begin{aligned}\sqrt{n}(\hat{\theta}_2^c - \theta_0) &= -\left(\Gamma' \left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} \Gamma\right)^{-1} \Gamma' \left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} \sqrt{n} g_n(\theta_0) + o_p(1) \\ &\xrightarrow{d} -\left[\Gamma'_\Lambda \bar{\mathbb{M}}^{-1} \Gamma_\Lambda\right]^{-1} \Gamma'_\Lambda \bar{\mathbb{M}}^{-1} \left(\sum_{g=1}^G \sqrt{\lambda_g} B_{m,g}\right),\end{aligned}$$

and leads us to obtain $F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) \xrightarrow{d} \tilde{\mathbb{F}}_{2\infty}$ where

$$\begin{aligned}\tilde{\mathbb{F}}_{2\infty} := \tilde{\mathbb{F}}_{2\infty}(\lambda) = \frac{1}{p} \cdot &\left[R\left(\Gamma'_\Lambda \bar{\mathbb{M}}^{-1} \Gamma_\Lambda\right)^{-1} \Gamma'_\Lambda \bar{\mathbb{M}}^{-1} \left(\sum_{g=1}^G \sqrt{\lambda_g} B_{m,g}\right)\right]' \left[R\left(\Gamma'_\Lambda \bar{\mathbb{M}}^{-1} \Gamma_\Lambda\right)^{-1} R'\right]^{-1} \\ \times &\left[R\left(\Gamma'_\Lambda \bar{\mathbb{M}}^{-1} \Gamma_\Lambda\right)^{-1} \Gamma'_\Lambda \bar{\mathbb{M}}^{-1} \left(\sum_{g=1}^G \sqrt{\lambda_g} B_{m,g}\right)\right].\end{aligned}$$

Let $U \Sigma V'$ be a singular value decomposition (SVD) of Γ_Λ. Also, denote $\tilde{U}\tilde{\Sigma}\tilde{V}'$ as a SVD of RVA^{-1}. Then, it is check that the following rotational invariance properties hold

$$\left[\sum_{g=1}^G \sqrt{\lambda_g} B_{m,g}, \bar{\mathbb{M}}^{-1}\right] \overset{d}{=} \left[U' \left(\sum_{g=1}^G \sqrt{\lambda_g} B_{m,g}\right), U' \bar{\mathbb{M}}^{-1} U\right],$$

and

$$\begin{aligned}&\left[\sum_{g=1}^G \sqrt{\lambda_g} B_{d,g}, \bar{\mathbb{M}}_{dq}, \bar{\mathbb{M}}_{qq}^{-1}, \bar{\mathbb{M}}_{dd\cdot q}\right] \\ \overset{d}{=} &\left[\tilde{V}' \sum_{g=1}^G \sqrt{\lambda_g} B_{d,g}, \tilde{V}' \bar{\mathbb{M}}_{dq}, \bar{\mathbb{M}}_{qq}^{-1}, \tilde{V}' \bar{\mathbb{M}}_{dd\cdot q} \tilde{V}\right]\end{aligned}$$

for $U'U = UU' = I_m$ and $\tilde{V}'\tilde{V} = \tilde{V}\tilde{V}' = I_d$. Then, the rest of proof is similar to that of Proposition 8-(a). The only differences is that the limit of CCE matrix $\bar{\mathbb{S}}$ and $\bar{B}_m$, and corresponding subcomponents are replaced with $\bar{\mathbb{M}}$, $\sum_{g=1}^G \sqrt{\lambda_g} B_{m,g}$, and corresponding submatrices, respectively. ■

B.6 Iterative Two-step and continuous updating Schemes

We consider two types of continuous updating schemes first suggested in Hansen et al. (1996). The first is motivated by the iterative scheme that updates the FOC of two-step GMM estimation until it converges. The FOC for $\hat{\theta}_{\mathrm{IE}}^{j}$ is

$$\hat{\Gamma}(\hat{\theta}_{\mathrm{IE}}^{j})'\hat{\Omega}^{-1}(\hat{\theta}_{\mathrm{IE}}^{j-1})g_n(\hat{\theta}_{\mathrm{IE}}^{j}) = 0 \text{ for } j \geq 1.$$

In view of the above FOC, $\hat{\theta}_{\mathrm{IE}}^{j}$ can be regarded as a generalized-estimating-equations (GEE) estimator, which is a class of estimators first proposed by Liang and Zeger (1986) and further studied by Jiang, Luan, and Wang (2007). When the number of iterations j goes to infinity until $\hat{\theta}_{\mathrm{IE}}^{j}$ converges, we obtain the continuous updating GEE (CU-GEE) estimator $\hat{\theta}_{\text{CU-GEE}}$. A recent paper by Hansen and Lee (2019) provides a rigorous theory for the existence of the iterated GMM estimator in Hansen et al (1996) in the context of both correctly specified and misspecified population moment conditions. The iterated GMM estimator is obtained by iterating the GMM object function instead of the FOC. Under correctly specified moment conditions, it is not difficult to check that the CU-GEE estimator considered here is asymptotically equivalent to the iterated GMM estimator. The FOC for $\hat{\theta}_{\text{CU-GEE}}$ is given by

$$\hat{\Gamma}(\hat{\theta}_{\text{CU-GEE}})'\hat{\Omega}^{-1}(\hat{\theta}_{\text{CU-GEE}})g_n(\hat{\theta}_{\text{CU-GEE}}) = 0. \tag{B.38}$$

Define

$$\hat{\Omega}^*(\hat{\theta}_{\text{CU-GEE}}) = \hat{\Omega}(\hat{\theta}_{\text{CU-GEE}}) - \left(\frac{1}{n}\sum_{g=1}^{G} L_g^2\right) \cdot g_n(\hat{\theta}_{\text{CU-GEE}})g_n(\hat{\theta}_{\text{CU-GEE}})'$$

which is an asymptotically equivalent eversion of the centered CCE $\hat{\Omega}^c(\hat{\theta}_{\text{CU-GEE}})$. While we employ the uncentered CCE, $\hat{\Omega}(\cdot)$ in the definition of $\hat{\theta}_{\text{CU-GEE}}$, it is not difficult to show that

$$\begin{aligned}
&\hat{\Gamma}(\hat{\theta}_{\text{CU-GEE}})'\hat{\Omega}^{-1}(\hat{\theta}_{\text{CU-GEE}})g_n(\hat{\theta}_{\text{CU-GEE}}) \\
&= \hat{\Gamma}(\hat{\theta}_{\text{CU-GEE}})'\left(\hat{\Omega}^*(\hat{\theta}_{\text{CU-GEE}})\right)^{-1} g_n(\hat{\theta}_{\text{CU-GEE}}) \cdot \frac{1}{1+\nu_n^*(\hat{\theta}_{\text{CU-GEE}})},
\end{aligned}$$

where

$$\nu_n^*(\hat{\theta}_{\text{CU-GEE}}) = \left(\frac{1}{n}\sum_{g=1}^{G} L_g^2\right) \cdot g_n(\hat{\theta}_{\text{CU-GEE}})'\left(\hat{\Omega}^*(\hat{\theta}_{\text{CU-GEE}})\right)^{-1} g_n(\hat{\theta}_{\text{CU-GEE}}).$$

Since $1/(1+\nu_n^*(\hat{\theta}_{\text{CU-GEE}}))$ is always positive, the first order condition in (B.38) holds if and only if

$$\hat{\Gamma}(\hat{\theta}_{\text{CU-GEE}})'\left[\hat{\Omega}^*(\hat{\theta}_{\text{CU-GEE}})\right]^{-1} g_n(\hat{\theta}_{\text{CU-GEE}}) = 0, \tag{B.39}$$

which indicates that replacing the CCE in (B.38) by $\hat{\Omega}^*(\hat{\theta}_{\text{CU-GEE}})$ has no effect on the iteration GMM estimator.

The second CU scheme continuously updates the GMM criterion function, which leads to the familiar continuous updating GMM (CU-GMM) estimator:

$$\hat{\theta}_{\text{CU-GMM}} = \underset{\theta\in\Theta}{\arg\min} g_n(\theta)'\hat{\Omega}^{-1}(\theta)g_n(\theta).$$

Although we use the uncentered CEE $\hat{\Omega}(\theta)$ in the above definition, the original idea of $\hat{\theta}_{\text{CU-GMM}}$ in Hansen, Heaton and Yaron (1996) is based on the centered CCE weighting matrix $\hat{\Omega}^c(\theta)$. With $M_n = n^{-1}\sum_{g=1}^{G} L_g^2$, it is easy to show that

$$\begin{aligned} M_n \cdot g_n(\theta)\hat{\Omega}^{-1}(\theta)g_n(\theta) &= M_n \cdot g_n(\theta)'\hat{\Omega}^{-1}(\theta)\left[\hat{\Omega}(\theta) - M_n \cdot g_n(\theta)g_n(\theta)'\right]\left[\hat{\Omega}^*(\theta)\right]^{-1} g_n(\theta) \\ &= M_n \cdot g_n(\theta)'\left[\hat{\Omega}^*(\theta)\right]^{-1} g_n(\theta)\left\{1 - M_n \cdot g_n(\theta)'\hat{\Omega}^{-1}(\theta)g_n(\theta)\right\}. \end{aligned}$$

Thus, we have

$$M_n \cdot g_n(\theta)'\left[\hat{\Omega}^*(\theta)\right]^{-1} g_n(\theta) = \frac{M_n \cdot g_n(\theta)'\hat{\Omega}^{-1}(\theta)g_n(\theta)}{1 - M_n \cdot g_n(\theta)'\hat{\Omega}^{-1}(\theta)g_n(\theta)}.$$

The above equation reveals the fact that the CU-GMM estimator will not change if the uncentered weighting matrix $\hat{\Omega}(\theta)$ is replaced by the centered one $\hat{\Omega}^*(\theta)$, that is,

$$\hat{\theta}_{\text{CU-GMM}} = \underset{\theta\in\Theta}{\arg\min} g_n(\theta)'\left[\hat{\Omega}^*(\theta)\right]^{-1} g_n(\theta). \tag{B.40}$$

Similar to the centered two-step GMM estimator, the two CU estimators can be regarded as having a built-in recentering mechanism via the version of CCE weight matrix $\hat{\Omega}^*(\theta)$. For this reason, the limiting distributions of the two CU estimators are the same as that of the centered two-step GMM estimator, as is shown below.

Proposition B.6 *Let Assumptions 1, 3–5 hold. Assume that $\hat{\theta}_{CU\text{-}GEE}$ and $\hat{\theta}_{CU\text{-}GMM}$ are $\sqrt{n}$-consistent. Then*

$$\sqrt{n}(\hat{\theta}_{CU\text{-}GEE} - \theta_0) \xrightarrow{d} -\left[\Gamma'\left(\Omega_\infty^c\right)^{-1}\Gamma\right]^{-1}\Gamma'\left(\Omega_\infty^c\right)^{-1}\Lambda\sqrt{G}\bar{B}_m$$

and

$$\sqrt{n}(\hat{\theta}_{CU\text{-}GMM} - \theta_0) \xrightarrow{d} -\left[\Gamma'\left(\Omega_\infty^c\right)^{-1}\Gamma\right]^{-1}\Gamma'\left(\Omega_\infty^c\right)^{-1}\Lambda\sqrt{G}\bar{B}_m.$$

In the proof of Proposition B.6, we show the asymptotic equivalence between two versions of centered CCE weighting matrix, i.e. $\hat{\Omega}^*(\tilde{\theta}) = \hat{\Omega}^c(\tilde{\theta}) + o_p(1)$ for any $\sqrt{n}$-consistent $\tilde{\theta}$. As a result, the CU estimators and the centered two-step GMM estimator are asymptotically equivalent under the fixed-G asymptotics. Based on the two CU estimators, we construct the Wald statistics as

$$F_{\hat{\Omega}^c(\hat{\theta}_{\text{CU-GEE}})}(\hat{\theta}_{\hat{\theta}_{\text{CU-GEE}}}) = \frac{1}{p}(R\hat{\theta}_{\text{CU-GEE}} - r)'\{R\widehat{var}_{\hat{\Omega}^c(\hat{\theta}_{\text{CU-GEE}})}(\hat{\theta}_{\text{CU-GEE}})R'\}^{-1}(R\hat{\theta}_{\text{CU-GEE}} - r) \tag{B.41}$$

and

$$F_{\hat{\Omega}^c(\hat{\theta}_{\text{CU-GMM}})}(\hat{\theta}_{\hat{\theta}_{\text{CU-GMM}}}) = \frac{1}{p}(R\hat{\theta}_{\text{CU-GMM}} - r)'\{R\widehat{var}_{\hat{\Omega}^c(\hat{\theta}_{\text{CU-GMM}})}(\hat{\theta}_{\text{CU-GMM}})R'\}^{-1}(R\hat{\theta}_{\text{CU-GMM}} - r). \tag{B.42}$$

We construct $t_{\hat{\Omega}^c(\hat{\theta}_{\text{CU-GEE}})}(\hat{\theta}_{\text{CU-GEE}})$ and $t_{\hat{\Omega}^c(\hat{\theta}_{\text{CU-GMM}})}(\hat{\theta}_{\text{CU-GMM}})$ in a similar way when $p = 1$. It follows from Proposition B.6 that the Wald statistics based on $\hat{\theta}_{\text{CU-GEE}}$ and $\hat{\theta}_{\text{CU-GMM}}$ are asymptotically equivalent to $F_{\hat{\Omega}^c(\hat{\theta}_\cdot)}(\hat{\theta}_2^c)$. As a result,

$$F_{\hat{\Omega}^c(\hat{\theta}_{\text{CU-GEE}})}(\hat{\theta}_{\text{CU-GEE}}) \xrightarrow{d} \mathbb{F}_{2\infty} \text{ and } F_{\hat{\Omega}^c(\hat{\theta}_{\text{CU-GMM}})}(\hat{\theta}_{\text{CU-GMM}}) \xrightarrow{d} \mathbb{F}_{2\infty}.$$

Similarly,

$$t_{\hat{\Omega}^c(\hat{\theta}_{\text{CU-GEE}})}(\hat{\theta}_{\text{CU-GEE}}) \xrightarrow{d} \mathbb{T}_{2\infty} \text{ and } t_{\hat{\Omega}^c(\hat{\theta}_{\text{CU-GMM}}}(\hat{\theta}_{\text{CU-GMM}}) \xrightarrow{d} \mathbb{T}_{2\infty}.$$

Note that changing the centered CCE $\hat{\Omega}^c(\cdot)$ by $\hat{\Omega}^*(\cdot)$ is innocuous to our fixed-G limiting distributions. In summary, we have shown that all three estimators $\hat{\theta}_2^{\text{c}}$, $\hat{\theta}_{\text{CU-GEE}}$ and $\hat{\theta}_{\text{CU-GMM}}$, and the corresponding Wald test statistics converge in distribution to the same nonstandard distributions. Proposition 8-(c) and (d) continues to hold for the CU-GEE and CU-GMM estimators, leading to the asymptotic equivalence of the three test statistics based on the CU-type estimators. That is, the CU-GMM estimator shares the first order fixed-smoothing limit with the two-step GMM estimator in our paper. Similar results have been found in a recent paper by Zhang (2016) in a time series setting who develops the fixed-smoothing asymptotic theory for the CU-GMM estimator.

Together with Theorem 10, Proposition B.6 imply that the modified of Wald, LR,LM, and t statistics based on the CU estimators are all asymptotically F and t distributed under the fixed-G asymptotics. For the finite-sample corrected variance formula, we have the following expansion

$$\begin{aligned}
&\sqrt{n}(\hat{\theta}_{\text{CU-GEE}} - \theta_0) \\
&= -\left(\Gamma'\left(\hat{\Omega}^c(\theta_0)\right)^{-1}\Gamma\right)^{-1}\Gamma'\left(\hat{\Omega}^c(\theta_0)\right)^{-1}\sqrt{n}g_n(\theta_0) + \mathcal{E}_{2n}\sqrt{n}(\hat{\theta}_{\text{CU-GEE}} - \theta_0) + o_p(1).
\end{aligned} \tag{B.43}$$

This can be regarded as a special case of (29) wherein the first-step estimator $\hat{\theta}_1$ is replaced by the CU-GEE estimator. Thus,

$$\sqrt{n}(\hat{\theta}_{\text{CU-GEE}} - \theta_0) \overset{a}{\sim} -(I_d - \mathcal{E}_{2n})^{-1}\left(\Gamma'\left(\hat{\Omega}^c(\theta_0)\right)^{-1}\Gamma\right)^{-1}\Gamma'\left(\hat{\Omega}^c(\theta_0)\right)^{-1}\sqrt{n}g_n(\theta_0), \tag{B.44}$$

We can obtain the same expression for the CU-GMM estimator $\sqrt{n}(\hat{\theta}_{\text{CU-GMM}} - \theta_0)$.

In view of the representation in (B.44), the corrected variance estimator for the CU type estimators can be constructed as follows:

$$\begin{aligned}
\widehat{var}^{\text{w}}_{\hat{\Omega}^c(\hat{\theta}_{\text{CU-GEE}})}(\hat{\theta}_{\text{CU-GEE}}) &= \left(I_d - \widehat{\mathcal{E}}_{\text{CU-GEE}}\right)^{-1}\widehat{var}\left(\hat{\theta}_{\text{CU-GEE}}\right)\left(I_d - \widehat{\mathcal{E}}'_{\text{CU-GEE}}\right)^{-1} \\
\widehat{var}^{\text{w}}_{\hat{\Omega}^c(\hat{\theta}_{\text{CU-GEE}})}(\hat{\theta}_{\text{CU-GMM}}) &= \left(I_d - \widehat{\mathcal{E}}_{\text{CU-GMM}}\right)^{-1}\widehat{var}\left(\hat{\theta}_{\text{CU-GMM}}\right)\left(I_d - \widehat{\mathcal{E}}'_{\text{CU-GMM}}\right)^{-1},
\end{aligned}$$

where

$$\begin{aligned}
\widehat{\mathcal{E}}_{\text{CU-GEE}}[.,j] = &\left\{\hat{\Gamma}'\left[\hat{\Omega}^c(\hat{\theta}_{\text{CU-GEE}})\right]^{-1}\hat{\Gamma}'\right\}^{-1} \\
&\times \hat{\Gamma}'\left\{\left[\hat{\Omega}^c(\hat{\theta}_{\text{CU-GEE}})\right]^{-1}\frac{\partial\hat{\Omega}^c(\hat{\theta}_{\text{CU-GEE}})}{\partial\theta_j}\left[\hat{\Omega}^c(\hat{\theta}_{\text{CU-GEE}})\right]^{-1}\right\}g_n(\hat{\theta}_{\text{CU-GEE}})
\end{aligned}$$

and $\widehat{\mathcal{E}}_{\text{CU-GMM}}$ is defined in the same way but with $\hat{\theta}_{\text{CU-GEE}}$ replaced by $\hat{\theta}_{\text{CU-GMM}}$. With the finite sample corrected and adjusted variance estimators in place, the Wald and t statistics based on the CU estimators also converge in distribution to the same nonstandard distributions in Proposition 8 (a) and (b), respectively.

B.6.1 Proof of Proposition B.6

Proof. For the result with CU-GEE estimator $\hat{\theta}_{\text{CU-GEE}}$, we have

$$\sqrt{n}(\hat{\theta}_{\text{CU-GEE}} - \theta_0) = -\left(\Gamma'\left[\hat{\Omega}^*(\hat{\theta}_{\text{CU-GEE}})\right]^{-1}\Gamma\right)^{-1}\Gamma'\left[\hat{\Omega}^*(\hat{\theta}_{\text{CU-GEE}})\right]^{-1}\sqrt{n}g_n(\theta_0) + o_p(1).$$

We first show that

$$\hat{\Omega}^*(\tilde{\theta}) = \hat{\Omega}^c(\tilde{\theta}) + o_p(1) \tag{B.45}$$

for any $\sqrt{n}$-consistent estimator $\tilde{\theta}$. Recall that

$$\begin{aligned}
\hat{\Omega}^*(\tilde{\theta}) &= \hat{\Omega}(\tilde{\theta}) - \left(\frac{1}{n}\sum_{g=1}^{G} L_g^2\right) \cdot g_n(\tilde{\theta}) g_n(\tilde{\theta}); \\
\hat{\Omega}^c(\tilde{\theta}) &= \frac{1}{n}\sum_{g=1}^{G} \left(\sum_{i=1}^{L_g} (f_i^g(\tilde{\theta}) - g_n(\tilde{\theta})\right) \left(\sum_{i=1}^{L_g} f_i^g(\tilde{\theta}) - g_n(\tilde{\theta})\right)' \\
&= \hat{\Omega}(\tilde{\theta}) - g_n(\tilde{\theta}) \cdot \left(\frac{1}{n}\sum_{g=1}^{G} L_g \sum_{i=1}^{L_g} f_i^g(\tilde{\theta})\right)' \\
&\quad - \left(\frac{1}{n}\sum_{g=1}^{G} L_g \sum_{i=1}^{L_g} f_i^g(\tilde{\theta})\right)' \cdot g_n(\tilde{\theta})' + \left(\frac{1}{n}\sum_{g=1}^{G} L_g^2\right) \cdot g_n(\tilde{\theta}) g_n(\tilde{\theta})'.
\end{aligned}$$

Thus, (B.45) can be proved by showing

$$g_n(\tilde{\theta}) \left(\frac{1}{n}\sum_{g=1}^{G} L_g \sum_{i=1}^{L_g} f_i^g(\tilde{\theta})\right)' = \left(\frac{1}{n}\sum_{g=1}^{G} L_g^2\right) \cdot g_n(\tilde{\theta}) g_n(\tilde{\theta})' + o_p(1). \tag{B.46}$$

By Assumption 1-iii) and $\sqrt{n}$-consistency of $\tilde{\theta}$,

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

Similarly, we obtain

$$\begin{aligned}
\left(\frac{1}{n}\sum_{g=1}^{G} L_g^2\right) \cdot g_n(\tilde{\theta}) g_n(\tilde{\theta})' &= \sum_{g=1}^{G} \left(\frac{L_g}{n}\right)^2 \cdot n g_n(\tilde{\theta}) g_n(\tilde{\theta})' \\
&= \frac{1}{G}\sum_{g=1}^{G} \left(\frac{L_g}{\bar{L}}\right)^2 \cdot \bar{L} g_n(\tilde{\theta}) g_n(\tilde{\theta})' \\
&= \bar{L} g_n(\tilde{\theta}) g_n(\tilde{\theta})' (1 + o_p(1)),
\end{aligned}$$

which gives the desired result in (B.46).

Now, using$\sqrt{n}$-consistency of $\hat{\theta}_{\text{CU-GEE}}$, we can apply Lemma 7 to obtain $\hat{\Omega}^*(\hat{\theta}_{\text{CU-GEE}}) = \hat{\Omega}^c(\theta_0) + o_p(1)$. Invoking the continuous mapping theorem yields

$$\sqrt{n}(\hat{\theta}_{\text{CU-GEE}} - \theta_0) \xrightarrow{d} -\left\{\Gamma' \left(\Omega_\infty^c\right)^{-1} \Gamma\right\}^{-1} \left\{\Gamma' \left(\Omega_\infty^c\right)^{-1} \Lambda \sqrt{G} \bar{B}_m\right\},$$

as desired.

For the CU-GMM estimator, we let $\hat{\Gamma}^j(\hat{\theta}_{\text{CU-GMM}})$ be the j-th column of $\hat{\Gamma}^j(\hat{\theta}_{\text{CU-GMM}})$. Then, the FOC with respect to the j-th element of $\hat{\theta}_{\text{CU-GMM}}$ is

$$\begin{aligned} 0 = {} & \hat{\Gamma}^j(\hat{\theta}_{\text{CU-GMM}})' \left[\hat{\Omega}^*(\hat{\theta}_{\text{CU-GMM}})\right]^{-1} g_n(\hat{\theta}_{\text{CU-GMM}}) \\ & - g_n(\hat{\theta}_{\text{CU-GMM}})' \left[\hat{\Omega}^*(\hat{\theta}_{\text{CU-GMM}})\right]^{-1} \Upsilon_j(\hat{\theta}_{\text{CU-GMM}}) \left[\hat{\Omega}^*(\hat{\theta}_{\text{CU-GMM}})\right]^{-1} g_n(\hat{\theta}_{\text{CU-GMM}}), \end{aligned} \tag{B.47}$$

where

$$\begin{aligned} \Upsilon_j^*(\theta) &= \frac{1}{n}\sum_{g=1}^{G} \left(\sum_{i=1}^{L_g} f_i^g(\theta)\right) \left(\sum_{i=1}^{L_g} \frac{\partial f_i^g(\theta)}{\partial \theta_j}\right)' - \left(\frac{1}{n}\sum_{g=1}^{G} L_g^2\right) \cdot g_n(\theta) \left(\frac{\partial g_n(\theta)}{\partial \theta_j}\right)' \\ &= \frac{1}{n}\sum_{g=1}^{G} \left(\sum_{i=1}^{L_g} f_i^g(\theta)\right) \left(\sum_{i=1}^{L_g} \frac{\partial f_i^g(\theta)}{\partial \theta_j}\right)' - \bar{L}(1+o(1)) \cdot g_n(\theta) \left(\frac{\partial g_n(\theta)}{\partial \theta_j}\right)'. \end{aligned}$$

Thus, the second term in (B.47) can be rewritten as

$$\begin{aligned} & g_n(\hat{\theta}_{\text{CU-GMM}})' \left[\hat{\Omega}^*(\hat{\theta}_{\text{CU-GMM}})\right]^{-1} \Upsilon_j^*(\hat{\theta}_{\text{CU-GMM}}) \left[\hat{\Omega}^*(\hat{\theta}_{\text{CU-GMM}})\right]^{-1} g_n(\hat{\theta}_{\text{CU-GMM}}) \\ & = \sqrt{\bar{L}} g_n(\hat{\theta}_{\text{CU-GMM}})' \left[\hat{\Omega}^*(\hat{\theta}_{\text{CU-GMM}})\right]^{-1} \left[\frac{1}{G}\sum_{g=1}^{G} \left(\frac{1}{\bar{L}}\sum_{i=1}^{L_g} f_i^g(\hat{\theta}_{\text{CU-GMM}})\right)\right. \\ & \times \left.\left\{\left(\frac{1}{\bar{L}}\sum_{i=1}^{L_g} \frac{\partial f_i^g(\hat{\theta}_{\text{CU-GMM}})}{\partial \theta_j}\right) - \frac{1}{G}\sum_{g=1}^{G}\left(\frac{1}{\bar{L}}\sum_{i=1}^{L_g} \frac{\partial f_i^g(\hat{\theta}_{\text{CU-GMM}})}{\partial \theta_j}\right)\right\}'\right] \\ & \times \left[\hat{\Omega}^*(\hat{\theta}_{\text{CU-GMM}})\right]^{-1} \sqrt{\bar{L}} g_n(\hat{\theta}_{\text{CU-GMM}})(1+o_p(1)). \end{aligned}$$

Given that $\hat{\theta}_{\text{CU-GMM}} = \theta_0 + O_p(\bar{L}^{-1/2})$ and $\hat{\Omega}^*(\hat{\theta}_{\text{CU-GMM}}) = \hat{\Omega}^c(\theta_0) + o_p(1)$, we have

$$\begin{aligned} & \hat{\Omega}^*(\hat{\theta}_{\text{CU-GMM}}) = O_p(1), \\ & \sqrt{\bar{L}} g_n(\hat{\theta}_{\text{CU-GMM}}) = \frac{1}{G}\sum_{g=1}^{G}\left(\frac{1}{\sqrt{\bar{L}}}\sum_{i=1}^{L_g} f_i^g(\theta_0)\right) + \Gamma\sqrt{\bar{L}}(\hat{\theta}_{\text{CU-GMM}} - \theta_0) + o_p(1) = O_p(1), \\ & \frac{1}{\bar{L}}\sum_{i=1}^{L_g} f_i^g(\hat{\theta}_{\text{CU-GMM}}) = \frac{1}{\bar{L}}\sum_{i=1}^{L_g} f_i^g(\theta_0) + \frac{1}{\bar{L}}\sum_{i=1}^{L_g} \frac{\partial f_i^g(\tilde{\theta})}{\partial \theta}(\hat{\theta}_{\text{CU-GMM}} - \theta_0) = O_p\left(\frac{1}{\sqrt{\bar{L}}}\right), \end{aligned}$$

and for each $g = 1, ..., G$,

$$\begin{aligned} & \left(\frac{1}{\bar{L}}\sum_{i=1}^{L_g} f_i^g(\hat{\theta}_{\text{CU-GMM}})\right)\left\{\left(\frac{1}{\bar{L}}\sum_{i=1}^{L_g} \frac{\partial f_k^g(\hat{\theta}_{\text{CU-GMM}})}{\partial \theta}\right) - \frac{1}{G}\sum_{g=1}^{G}\left(\frac{1}{\bar{L}}\sum_{i=1}^{L_g} \frac{\partial f_i^g(\hat{\theta}_{\text{CU-GMM}})}{\partial \theta}\right)\right\}' \\ & = O_p\left(\frac{1}{\sqrt{\bar{L}}}\right) \cdot o_p(1) = o_p\left(\frac{1}{\sqrt{\bar{L}}}\right). \end{aligned}$$

Combining these together, the second term in FOC in (B.47) is $o_p(\bar{L}^{-1/2})$. As a result,

$$\hat{\Gamma}(\hat{\theta}_{\text{CU-GMM}})' \left[\hat{\Omega}^*(\hat{\theta}_{\text{CU-GMM}})\right]^{-1} g_n(\hat{\theta}_{\text{CU-GMM}}) = o_p\left(\frac{1}{\sqrt{\bar{L}}}\right).$$

Using this result and $\hat{\Omega}^*(\hat{\theta}_{\text{CU-GMM}}) = \hat{\Omega}^c(\theta_0) + o_p(1)$, we obtain the desired result as

$$\begin{aligned}\sqrt{n}(\hat{\theta}_{\text{CU-GMM}} - \theta_0) &= -\left\{\Gamma' \left[\hat{\Omega}^*(\hat{\theta}_{\text{CU-GMM}})\right]^{-1} \Gamma\right\}^{-1} \Gamma' \left[\hat{\Omega}^*(\hat{\theta}_{\text{CU-GMM}})\right]^{-1} \sqrt{n} g_n(\theta_0) + o_p(1) \\ &\overset{d}{\to} -\left\{\Gamma' \left(\Omega^c_\infty\right)^{-1} \Gamma\right\}^{-1} \Gamma' \left(\Omega^c_\infty\right)^{-1} \Lambda\sqrt{G}\bar{B}_m.\end{aligned}$$

■

B.7 Application to linear dynamic panel model

This subsection illustrates how to implement the GMM testing procedures in the context of linear dynamic panel model under clustered dependence. Consider

$$y_{it} = \gamma y_{it-1} + x_{it}'\beta + \eta_i + u_{it}, \tag{B.48}$$

for $i = 1, ..., n,\ t = 1, ..., T$, where $x_{it} = (x_{it}^1, ..., x_{it}^{d-1})' \in \mathbb{R}^{d-1}$. The unknown parameter vector is $\theta = (\gamma, \beta')' \in \mathbb{R}^d$. We assume that the vector of regressors $w_{it} = (y_{it-1}, x_{it}'\ )'$ is correlated with η_i and is predetermined with respect to u_{it}, i.e., $E(w_{it}u_{it+s}) = 0$ for $s = 0, ..., T-t$.

When T is small, popular panel estimators such as the fixed-effects estimator or first-differenced estimator suffer from the Nickel bias (Nickel, 1981). Anderson and Hsiao (1981) consider the first-differenced equation

$$\Delta y_{it} = \Delta w_{it}'\theta + \Delta u_{it},\ t = 2, ..., T,$$

and propose a consistent IV estimator that employs the lagged w_{it}, e.g. w_{it-1} or Δw_{it-1}, as the instrument. Building upon Anderson and Hsiao (1981), Arellano and Bond (1991, AB hereinafter) examine the problem in a GMM framework and find $dT(T-1)/2$ sequential instruments:

$$\underset{(T-1)\times d(T-1)T/2}{Z_i} = diag(z_{i2}', ..., z_{iT}'),$$

where $z_{it} = (y_{i0}, ..., y_{it-2}, x_{i1}', ..., x_{it-1}')', 2 \le t \le T$. The moment conditions are then given by

$$E\left(Z_i'\Delta u_i\right) = 0,$$

where Δu_i is the $(T-1)$ vector $(\Delta u_{i2}, ..., \Delta u_{iT})'$. The original AB method assumes away cross-sectional dependence, but clustered dependence can be easily accommodated. Here we assume that the moment vector $\{Z_i'\Delta u_i\}_{i=1}^n$ can be partitioned into independent clusters. That is, $\{Z_i'\Delta u_i\}_{i=1}^n = \cup_{g=1}^G \cup_{k=1}^{L_g} \{Z_k^{g\prime}\Delta u_k^g\}$ with $Z_k^{g\prime}\Delta u_k^g$ and $Z_l^{h\prime}\Delta u_l^h$ being independent for all $g \neq h$.

The first-step GMM estimator with initial weighting matrix W_n^{-1} is given by

$$\hat{\theta}_1 = \left(\Delta w' Z W_n^{-1} Z'\Delta w\right)^{-1} \Delta w' Z W_n^{-1} Z'\Delta y,$$

where Z' is the $dT(T-1)/2 \times n(T-1)$ matrix $(Z_1', Z_2', ..., Z_n')$, Δw_i is the $(T-1)\times d$ matrix $(\Delta w_{i2}, ..., \Delta w_{iT})'$, Δy_i is the $(\Delta y_{i2}, ..., \Delta y_{iT})'$, Δw and Δy are $(\Delta w_1', ..., \Delta w_n')'$ and $(\Delta y_1', ..., \Delta y_n')'$, respectively. The examples of W_n's can be $Z'Z/n$ for 2SLS and $n^{-1}\sum_{i=1}^n Z_i'HZ_i$ where H is a matrix that consists with 2's on the main diagonal, with -1's on the main diagonal, and zeros elsewhere.

The Wald statistic[3] for testing $H_0 : R\theta_0 = r$ vs $H_1 : R\theta_0 \neq r$ is given by

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

where

$$\widehat{var}(\hat{\theta}_1) = n\left(\Delta w' Z W_n^{-1} Z'\Delta w\right)^{-1}\left(\Delta w' Z W_n^{-1}\hat{\Omega}(\hat{\theta}_1) W_n^{-1} Z'\Delta w\right)\left(\Delta w' Z W_n^{-1} Z'\Delta w\right)^{-1}.$$

Let $Z_{(g)}$ be the $L(T-1)\times dT(T-1)/2$ matrix obtained by stacking all Z_i's belonging to cluster g. Similarly, let $\Delta\hat{u}_{(g)}$ be the $L_g(T-1)$ stacked vector of the estimated first-differenced errors

[3] The formula for the t statistic, which is omitted here, can be similarly constructed.

$\Delta\hat{u}_i = \Delta y_i - \Delta w_i'\hat{\theta}_1$. Then, in the presence of clustered dependence, the CCE and centered CCE are constructed as

$$\hat{\Omega}(\hat{\theta}_1) = \frac{1}{G}\sum_{g=1}^{G}\left(\frac{Z_{(g)}'\Delta\hat{u}_{(g)}}{\sqrt{\bar{L}}}\right)\left(\frac{Z_{(g)}'\Delta\hat{u}_{(g)}}{\sqrt{\bar{L}}}\right)'$$

and

$$\hat{\Omega}^c(\hat{\theta}_1) = \frac{1}{G}\sum_{g=1}^{G}\left(\frac{Z_{(g)}'\Delta\hat{u}_{(g)}}{\sqrt{\bar{L}}} - \frac{1}{G}\sum_{\tilde{g}=1}^{G}\frac{Z_{(\tilde{g})}'\Delta\hat{u}_{(\tilde{g})}}{\sqrt{\bar{L}}}\right)\left(\frac{Z_{(g)}'\Delta\hat{u}_{(g)}}{\sqrt{\bar{L}}} - \frac{1}{G}\sum_{\tilde{g}=1}^{G}\frac{Z_{(\tilde{g})}'\Delta\hat{u}_{(\tilde{g})}}{\sqrt{\bar{L}}}\right)'.$$

Using the centered CCE $\hat{\Omega}^c(\hat{\theta}_1)$ as the weighting matrix, the two-step GMM estimator $\hat{\theta}_2^c$ is

$$\hat{\theta}_2^c = \left(\Delta w' Z\left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} Z'\Delta w\right)^{-1}\Delta w' Z\left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} Z'\Delta y,$$

and the t and Wald statistics for $\hat{\theta}_2^c$ is

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

Under the conventional large-G asymptotics, both $F(\hat{\theta}_1)$ and $F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)$ are asymptotically χ_p^2/p. Under our fixed-G asymptotics, we have

$$F(\hat{\theta}_1) \overset{d}{\to} \frac{G}{G-p}F_{p,G-p} \text{ and}$$
$$F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) \overset{d}{\to} \frac{G}{G-p-q}F_{p,G-p-q}\left(\|\Delta\|^2\right). \tag{B.49}$$

In addition to utilizing these new approximations, we suggest a variance correction in order to capture the higher order effect of $\hat{\theta}_1$ on $\hat{\Omega}^c(\hat{\theta}_1)$. The finite sample corrected variance is

$$\widehat{var}^{\mathrm{w}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) = \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) + \widehat{\mathcal{E}}_n\widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) + \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)\widehat{\mathcal{E}}_n' + \widehat{\mathcal{E}}_n\widehat{var}(\hat{\theta}_1)\widehat{\mathcal{E}}_n', \tag{B.50}$$

where the j-th column is given by

$$\widehat{\mathcal{E}}_n[.,j] = -\left(\Delta w' Z\left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} Z'\Delta w\right)^{-1}\Delta w' Z\left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1}\times$$
$$\left.\frac{\partial\hat{\Omega}^c(\theta)}{\partial\theta_j}\right|_{\theta=\hat{\theta}_1}\left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} Z'\Delta\hat{u}_2,$$
$$\Delta\hat{u}_2 = \Delta y - \Delta w\hat{\theta}_2^c,$$

and

$$\left.\frac{\partial\hat{\Omega}^c(\theta)}{\partial\theta_j}\right|_{\theta=\hat{\theta}_1} = \Upsilon_j(\hat{\theta}_1) + \Upsilon_j'(\hat{\theta}_1),$$

$$\Upsilon_j(\hat{\theta}_1) = -\frac{1}{G}\sum_{g=1}^{G}\left(\frac{Z_{(g)}'\Delta w_{j,(g)}}{\sqrt{\bar{L}}} - \frac{1}{G}\sum_{\tilde{g}=1}^{G}\frac{Z_{(\tilde{g})}'\Delta w_{j,(\tilde{g})}}{\sqrt{\bar{L}}}\right)\left(\frac{Z_{(g)}'\Delta\hat{u}_{(g)}}{\sqrt{\bar{L}}} - \frac{1}{G}\sum_{\tilde{g}=1}^{G}\frac{Z_{(\tilde{g})}'\Delta\hat{u}_{(\tilde{g})}}{\sqrt{\bar{L}}}\right)',$$

$$\underset{(T-1)\times d}{\Delta w_i} = (\Delta w_{1,i}, ..., \Delta w_{d,i}) \text{ and } \underset{l(T-1)\times d}{\Delta w_{(g)}} = (\Delta w_{1,(g)}, ..., \Delta w_{d,(g)})$$

for each $j = 1, ..., d$. Here, $\Delta w_{(g)}$ is a $L_g(T-1) \times d$ matrix that stacks $\{\Delta w_i\}_{i=1}^n$ belonging to the group g.

Based on the finite sample corrected variance estimator in (B.50) and the usual J statistic, we construct the modified Wald and t statistics

$$\tilde{F}^{\mathrm{w}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) = \frac{G-p-q}{G} \frac{F^{\mathrm{w}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)}{1 + \frac{1}{G} J(\hat{\theta}_2^c)}, \tag{B.51}$$

where

$$\begin{aligned} J(\hat{\theta}_2^c) &= n \cdot g_n(\hat{\theta}_2^c)' \left(\hat{\Omega}^c(\hat{\theta}_2^c)\right)^{-1} (\theta) g_n(\hat{\theta}_2^c) \\ &= \left(\frac{Z'\Delta u(\hat{\theta}_2^c)}{\sqrt{n}}\right)' \left[\hat{\Omega}^c(\hat{\theta}_2^c)\right]^{-1} \left(\frac{Z'\Delta u(\hat{\theta}_2^c)}{\sqrt{n}}\right). \end{aligned}$$

From the F and t limit theory in Section 4 and the fixed-G approximation of the finite sample corrected variance in Section 5, we have

$$\tilde{F}^{\mathrm{w}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) \overset{d}{\to} F_{p,G-p-q} \tag{B.52}$$

and

$$\frac{G-q}{Gq} J(\hat{\theta}_2^c) \overset{d}{\to} F_{q,G-q}.$$

The critical values from the F and t distributions are readily available from statistical tables.

B.8 Procedure for cluster-robust Hall and Horowitz (1996) two-step GMM-bootstrap

Conditioning on the original sample $\{Z_{(g)}, \Delta u_{(g)}(\theta), \Delta y_{(g)}, \Delta w_{(g)}\}_{g=1}^{G}$, let $\{Z_i^*, \Delta y_i^*, \Delta u_i^*(\theta), \Delta w_i^*\}_{i=1}^{n} = \cup_{g=1}^{G}\{\ Z^*_{(g)},\ \Delta y^*_{(g)}, \Delta u^*_{(g)}(\theta), \Delta w^*_{(g)}\}$ be a cluster-wise bootstrap sample. Given the b-th resampled data $\cup_{g=1}^{G}\{\ Z^*_{(g)},\ \Delta y^*_{(g)}, \Delta u^*_{(g)}(\theta), \Delta w^*_{(g)}\}$, we can implement the GMM bootstrap procedure in Hall and Horowitz (1996) as follows.

Step 1: With the recentered moment function $(Z_i^*)'\Delta u_i^*(\theta) - E^*[(Z_i^*)'\Delta u_i^*(\hat{\theta}_1)]$, obtain a bootstrap version of the initial estimator:

$$\hat{\theta}^*_{1,(b)} = \left[(\Delta w^*)'\, Z^*\, [W_n^*]^{-1}\, (Z^*)'\Delta w^*\right]^{-1} (\Delta w^*)'\, Z^*\, [W_n^*]^{-1} \left((Z^*)'\Delta y^* - Z'\Delta u(\hat{\theta}_1)\right).$$

, where $W_n^* = n^{-1}\sum_{i=1}^{n}(Z_i^*)'Z_i^*$.

Step 2: Obtain a bootstrap version of the two-step GMM weighting matrix $\hat{\Omega}^{c*}(\hat{\theta}^*_{1,(b)})$ with the recentered moment process $(Z_i^*)'\Delta u_i^*(\theta) - E^*[(Z_i^*)'\Delta u_i^*(\hat{\theta}_2)]$:

$$\hat{\Omega}^{c*}(\hat{\theta}^*_{1,(b)};\hat{\theta}_2) = \frac{1}{G}\sum_{g=1}^{G}\left[\left(\frac{(Z^*_{(g)})'\Delta u^*_{(g)}(\hat{\theta}^*_{1,(b)})}{\sqrt{L}} - \frac{1}{G}\sum_{h=1}^{G}\frac{Z'_{(h)}\Delta u_{(h)}(\hat{\theta}_2)}{\sqrt{L}}\right)\times \right.$$
$$\left.\left(\frac{(Z^*_{(g)})'\Delta u^*_{(g)}(\hat{\theta}^*_{1,(b)})}{\sqrt{L}} - \frac{1}{G}\sum_{h=1}^{G}\frac{Z'_{(h)}\Delta u_{(h)}(\hat{\theta}_2)}{\sqrt{L}}\right)'\right].$$

, and corresponding two-step GMM estimator:

$$\hat{\theta}^*_{2,(b)} = \left[(\Delta w^*)'\, Z^* \left[\hat{\Omega}^{c*}(\hat{\theta}^*_{1,(b)})\right]^{-1} (Z^*)'\Delta w^*\right]^{-1} (\Delta w^*)'\, Z^* \left[\hat{\Omega}^{c*}(\hat{\theta}^*_{1,(b)})\right]^{-1} \times$$
$$\left[(Z^*)'\Delta y^* - Z'\Delta u(\hat{\theta}_1)\right].$$

Step 3: Construct the b-th bootstrap version of t statistic:

$$t^*(\hat{\theta}^*_{2,(b)}) = \frac{R(\hat{\theta}^*_{2,(b)} - \hat{\theta}_2)}{\sqrt{R\widehat{var}_{\hat{\Omega}^{c*}(\hat{\theta}^*_{1,(b)};\hat{\theta}_2)}(\hat{\theta}^*_{2,(b)})R'}},$$

where $\widehat{var}_{\hat{\Omega}^{c*}(\hat{\theta}^*_{1,(b)})}(\hat{\theta}^*_{2,(b)}) = n\left((\Delta w^*)'\, Z^* \left[\hat{\Omega}^{c*}(\hat{\theta}^*_{1,(b)};\hat{\theta}_2)\right]^{-1} (Z^*)'\Delta w^*\right)^{-1}$, and J statistic:

$$J^*(\hat{\theta}^*_{2,(b)};\hat{\theta}_2) = \left[\frac{Z^{*\prime}\Delta u^*(\hat{\theta}^*_{2,(b)}) - Z'\Delta u(\hat{\theta}_2)}{\sqrt{n}}\right]' \left[\hat{\Omega}^{c*}(\hat{\theta}^*_{2,(b)};\hat{\theta}_2)\right]^{-1} \times$$
$$\left[\frac{Z^{*\prime}\Delta u^*(\hat{\theta}^*_{2,(b)}) - Z'\Delta u(\hat{\theta}_2)}{\sqrt{n}}\right].$$

Step 4: After repeating 1∼3 steps B-times, compute the two-side bootstrap p-values for t and J statistics with

$$\hat{p}_{t,\mathrm{HH}} = \frac{1}{B}\sum_{b=1}^{B} 1\left(\left|t^*(\hat{\theta}^*_{2,(b)})\right| > \left|t(\hat{\theta}_2)\right|\right) \text{ and } \hat{p}_{J,\mathrm{HH}} = \frac{1}{B}\sum_{b=1}^{B} 1\left(J^*(\hat{\theta}^*_{2,(b)}) > J(\hat{\theta}_2)\right).$$

Then, we reject Reject the null hypothesis $H_0 : R\theta = r$ iff

$$\hat{p}_{t,HH} \leqq \alpha$$

, and reject the null hypothesis of J test $H_0 : E\left(Z_i' \Delta u_i\right) = 0$ iff

$$\hat{p}_{J,HH} \leqq \alpha.$$

B.9 Testing the level of clustering

The Ibragimov and Muller (2016, IM hereinafter)'s test considers a scenario that empirical researchers face a choice between a small number of coarse clusters and a large number of the finer level of clusters. The null hypothesis is that a finer level of clustering is appropriate with a consistent CCE estimator, against the alternative the only fewer clusters provide valid information. As a general motivation, consider a linear regression

$$y_i^g = (X_i^g)'\theta + \epsilon_i^g,$$

where y_i^g and X_i^g are the i-th of L_g observations from cluster $g = 1, \ldots, G$. Suppose a researcher is interested in the j-th element of $\theta \in \mathbb{R}^d$, $\beta_j = e_j'\theta$, where e_j is a j-th standard basis vector in $\mathbb{R}^d$. We first partition sample into G clusters, and estimate the model only using observations in cluster g and obtain $\hat{\beta}_j^g$ for each $g = 1, \ldots G$, and compute the following statistics S^2:

$$S^2 = \frac{1}{G-1}\sum_{g=1}^{G}(\hat{\beta}_j^g - \bar{\hat{\beta}})^2 \text{ where } \bar{\hat{\beta}} = \frac{1}{G}\sum_{g=1}^{G}\hat{\beta}_j^g. \tag{B.53}$$

In the estimation of $\hat{\beta}_j^g$, one also calculates its cluster-robust standard errors $\hat{\sigma}_g$ assuming the finer level of clustering within each group g is appropriate. To obtain the critical value of S^2, we draw $Y_g \overset{\text{i.i.d}}{\sim} N(0, \hat{\sigma}_g^2)$ for $g = 1, \ldots, G$, and compute

$$S_Y^2 = \frac{1}{G-1}\sum_{g=1}^{G}(Y_g - \bar{Y})^2 \text{ where } \bar{Y} = \frac{1}{G}\sum_{g=1}^{G}Y_g. \tag{B.54}$$

When the test statistics S^2 is larger than 95% percentile of the $10,000$ draws of S_Y^2, the test rejects the null hypothesis of validity of finer level of clustering. IM (2016) shows that the test is asymptotically size corrected, and has a non-trivial asymptotic power when the fine level of clustering ignores correlation among the observations in the coarser cluster.

In our empirical example in Section 7, we consider a finer clustering at the level of individuals. To apply the IM's test, we first need to estimate the key parameter of interests, β_d, β_s, and β_{ds} including all nuisance parameters at each county level cluster. However, we note that the number of observations at each county in Emran and Hou (2013)'s data set are insufficient to estimate the full regression equation in Section 7. This is mostly coming from having control variables that do not have sufficient within-cluster (within-county) variations. Thus, as the next best thing, we could estimate the cluster-specific OLS estimators, $\hat{\beta}_d^g$, $\hat{\beta}_s^g$, and $\hat{\beta}_{ds}^g$, for each county cluster g by only considering the key variables of interests in the regression. We also compute the corresponding robust standard errors, $\hat{\sigma}_d^g$, $\hat{\sigma}_s^g$, and $\hat{\sigma}_{ds}^g$ at the level of individual clustering. Based on these estimates, we construct the test statistics and corresponding critical values as in (B.53) and (B.54) for each parameter of interests. The corresponding p-values of tests are reported in Table 7 in the main body of the paper.

References

[1] Anderson, T. W. and Hsiao, C. (1981): "Estimation of dynamic models with error components." *Journal of the American statistical Association*, 76(375), 598–606.

[2] 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.

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

[4] Bolthausen, E. (1982): "On the central limit theorem for stationary mixing random fields." *The Annals of Probability*, 1047–1050.

[5] Conley, T. G. (1999): "GMM estimation with cross sectional dependence." *Journal of econometrics*, 92(1), 1–45.

[6] Dedecker, J. (1998): "A central limit theorem for stationary random fields." *Probability Theory and Related Fields*, 110(3), 397–426.

[7] 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.

[8] 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.

[9] Hansen, B. E. and Lee, S. (2018): "Inference for iterated GMM under misspecification and clustering." UNSW Business School Research Paper, (2018–07).

[10] 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.

[11] Hall, P. and Heyde, C. C. (2014): "Martingale limit theory and its application." Academic press.

[12] Hwang, J. and Sun, Y. (2017): "Asymptotic F and t Tests in an Efficient GMM Setting." *Journal of Econometrics*, 198(2), 277–295.

[13] 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.

[14] Jiang, J., Luan, Y., and Wang, Y. G. (2007): "Iterative estimating equations: Linear convergence and asymptotic properties." *The Annals of Statistics*, 35(5), 2233–2260.

[15] Liang, K.-Y. and Zeger, S. (1986): "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73(1):13–22.

[16] Nickell, S. (1981): "Biases in dynamic models with fixed effects." *Econometrica*

[17] , 1417–1426.

[18] Sun, Y. (2014): "Fixed-smoothing Asymptotics in a Two-step GMM Framework." *Econometrica* 82(6), 2327–2370.

[19] Zhang, X. (2016): "Fixed-smoothing asymptotics in the generalized empirical likelihood estimation framework." *Journal of Econometrics*, 193(1), 123–146.