Preliminary manuscript for external reviewers

Fixed-cluster Inference with Unbalanced Cluster Sizes

Jungbin Hwang*and Min Seong Kim†
Department of Economics,
University of Connecticut

May 30, 2021

Abstract

This paper provides asymptotic distribution and approximation theory for clustered data with a fixed number of large-sized clusters, allowing cluster sizes to be heterogeneous. In developing new inference procedures under fixed-cluster asymptotics, we make the following contributions. First, we show that the Wald statistic's asymptotic representation with possibly heterogeneous cluster sizes is a quadratic form of a standard normal vector with an independent random weighting matrix. The random weighting matrix is equal in distribution to a weighted sum of independent Wishart matrices. Second, we develop a convenient F-approximation to the nonstandard fixed-cluster limiting distribution. The key intuition is that a scaled Wishart matrix can well approximate the random weighting matrix in the fixed-cluster limit of the Wald statistic with a proper choice of degree of freedom. We also show that the equivalent degree of freedom in our F-approximation connects to the equivalent number of clusters in Carter et al.'s (2013). Lastly, we establish the fixed-G asymptotics and F approximation for the test based on the jackknife CCE and show the built-in bias adjustment functions in the jackknife CCE using the fixed-cluster asymptotics.

1 Introduction

Economic data is potentially dependent and heterogenous across individuals. In such cases, estimators of model parameters usually remain asymptotically valid, but it becomes challenging to correctly estimate their sampling variations. This is a crucial problem to make valid inference, because an estimator with positively correlated data tends to have a larger variation so failure to account for the correlation often causes a misleading inference result. To address this, it is common practice in empirical studies to employ the cluster variance estimator (CCE) that is robust to an arbitrary form of within-cluster correlations while assuming zero correlations across clusters. See White (1984, Theorem 6.3, p.136), Liang and Zeger (1986), Arellano (1987) for seminal methodological contributions, and Wooldridge (2003) and Cameron and Miller (2015) for overviews of the CCE and its applications. The main goal of this paper is to propose new and more accurate inference procedures based on the CCE.

For implementing the cluster-robust inference, two different types of asymptotics have been considered in the literature to characterize the limiting behavior of the CCE and associated test statistic: large-cluster (or large-G) asymptotics and fixed-cluster (or fixed-G) asymptotics.

*Email: jungbin.hwang@uconn.edu
†Email: min_seong.kim@uconn.edu

The most commonly used cluster-robust inference is based on the large-cluster asymptotics, under which the number of clusters G is assumed to increase with the sample size n. Under this asymptotic specification, the CCE becomes concentrated at the true asymptotic variance as $G \to \infty$ and the associated Wald statistic converges in distribution to a χ^2 distribution. Although the large-cluster asymptotic test works well in many cases, its finite sample performance can be very poor if the number of clusters is small and/or cluster sizes are heterogenous. This is because the estimation uncertainty in a CCE is large in such cases, but the approach treats the CCE as the true asymptotic variance. See Bester, Conley and Hansen (2011), and Carter, Schnepel and Steigerwald (2017) for further explanations.

The fixed-cluster asymptotics is proposed by Hansen (2007) and Bester, Conley and Hansen (2011, BCH hereafter) to design more accurate cluster-robust inference than the large-cluster asymptotics by capturing the estimation uncertainty of the CCE in the limiting distribution. The asymptotics sets G to be equal to a fixed number as $n \to \infty$ while letting the number of individuals in each cluster increase with n. As a result, the CCE converges to a random matrix that is proportional to the true asymptotic variance, and the Wald statistic becomes asymptotically nuisance parameter free but has a non-χ^2 limiting distribution. BCH show that when all cluster sizes are asymptotically the same the suitably modified Wald statistic converges weekly to a standard F distribution. They also show by simulation that the fixed-cluster asymptotic test has better size properties than the large-cluster χ^2 approximation.

In the paper, we generalize the fixed-cluster asymptotics to be applicable to the GMM setting allowing cluster sizes to be heterogenous. Heterogeneity in cluster sizes is indeed common in empirical studies, but the existing fixed-cluster theory focuses on the case that clusters are equally sized. In developing new inference procedures, we make several contributions. First, we derive the asymptotic representation of the Wald statistic with possibly heterogenous cluster sizes, which is a quadratic form of a standard normal vector with an independent and random weighting matrix. The representation implies the fixed-cluster distribution is pivotal but nonstandard except the case that cluster sizes are asymptotically homogenous. We then show that the random weighting matrix is equal in distribution to a weighted sum of independent Wishart matrices. Based on this result, we take the asymptotic expansion of the fixed-cluster limiting distribution around the large-cluster χ^2 distribution to examine the difference between these two distributions rigorously. The expansion indicates explicitly the former accounts for the effects of the number of clusters, heterogeneity in cluster sizes, the estimation error in parameter estimators, and the dimension of null restrictions, all of which the latter assumes away.

Next, we develop an F-approximation to the nonstandard fixed-cluster limiting distribution. While the inference based on the fixed-cluster asymptotics is more accurate, it is not as convenient to use as the large-cluster χ^2 test because extra computation is required to simulate critical values in general. To tackle this issue, we propose an F test by approximating the fixed-cluster distribution of a Wald statistic with a scaled F distribution. The key intuition is that the random weighting matrix in the fixed-cluster limit of the Wald statistic can be well approximated by a scaled Wishart matrix with certain choice of degree of freedom. Our approach includes the fixed-cluster asymptotic F test by BCH as a special case when all cluster sizes are asymptotically equivalent. Using the asymptotic expansion, we show that the deviation of our F approximation from the fixed-cluster distribution is small. The expansion also implies that our approximate F distribution reduces to the large-cluster χ^2 distribution as $G \to \infty$. Therefore, we can expect our F test to yield accurate inference results regardless of whether G is small or large. Our F-approximation is motivated from Satterthwaite (1946) who approximates the distribution of a

weighted sum of χ^2 random variables with a χ^2 distribution with an adjusted degree of freedom. In the HAR inference literature, Sun (2014) employs this idea to develop an F-approximation to the fixed-b limiting distribution. We apply the approach to the fixed-cluster asymptotics in the cluster-robust inference context.

Lastly, we extend our fixed-cluster asymptotics and its F-approximation to the cluster-robust inference with a bias reduced CCE in the linear regression setting. One weakness of the cluster-robust inference under the large-cluster asymptotics is that the CCE can have a nontrivial downward bias in small samples which leads to serious overrejection. In this regard, Bell and McCaffrey (2002) propose bias reduced CCEs that are analogous to the bias reduced versions of the Eicker-Huber-White variance estimator by White and MacKinnon (1985). We employ their Jackknife CCE and derive its asymptotic unbiasedness under the fixed-cluster asymptotics. We then establish the F-approximation and provide its theoretical justification by showing that its deviation from the fixed-cluster distribution is of small order as G increases. Our approach is related to Bell and McCaffrey (2002) and Imbens and Kolesar (2016) who suggest t-distribution approximations to the distribution of a t-statistic based on the bias reduced CCEs. Their methods are very attractive and encouraged to use in empirical studies because of their substantial improvement in sizes over the normal approximation. On the other hand, there is an issue from a theoretical point of view that no clear theory is developed for its justification. We address this by developing an accurate F approximation which is justified from the fixed-cluster asymptotic perspective.

The fixed-cluster asymptotics in the cluster-robust inference has gained growing attention in the literature. Ibragimov and Müller (2010, 2016) proposes a subsample based t test for a scalar parameter that is robust to heterogeneous clusters. Hansen (2007), Stock and Watson (2008), and BCH propose a cluster-robust F and t tests under cluster-size homogeneity. Hwang (2020) establishes the fixed-cluster asymptotics under two-step GMM framework. 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, Carter, Schnepel and Steigerwald (2017) propose a measure of the effective number of clusters and Cameron, Gelbach and Miller (2008), MacKinnon and Webb (2017), and Djogbenou, MacKinnon and Nielsen (2019) investigate cluster-robust bootstrap approaches. Recently, a large-cluster asymptotic distribution theory is developed by Hansen and Lee (2019).

Our F-approximation is related to the F-approximation to the fixed-b limiting distribution in the HAR inference literature. Sun (2014) establishes the validity of an F-approximation in the general GMM framework, and Kim and Sun (2013) and Hwang and Sun (2017) extend this approach to the panel data model and efficient GMM setting.

2 Basic Set-up

Let Y_i be a $d_y \times 1$ vector of observations for $i = 1, \ldots, n$. We want to estimate $d_\theta \times 1$ vector of parameters $\theta \in \Theta$. The true parameter vector θ_0 is assumed to be an interior point of parameter space $\Theta \subseteq \mathbb{R}^{d_\theta}$, and characterized by the moment condition

$$Ef(Y_i, \theta) = 0 \text{ holds if and only if } \theta = \theta_0, \tag{1}$$

where $f_i(\theta) = f(Y_i, \theta)$ is an $d_f \times 1$ vector of twice continuously differentiable function with $d_f - d_\theta \geq 0$. We assume that $\Gamma := \Gamma(\theta_0) = E\left[\partial f(Y_i, \theta_0)/\partial \theta'\right]$ has a full rank d_θ. We allow that the moment process $\{f_i(\theta_0)\}_{i=1}^n$ is cross-sectionally dependent with unknown form. Define $g_n(\theta) = n^{-1} \sum_{i=1}^n f_i(\theta)$. Let W_n be a $d_f \times d_f$ positive definite and a symmetric weighting matrix

that does not depend on the unknown parameter θ_0, and $\text{plim}_{n\to\infty} W_n = W > 0$. Given the moment condition in (1), the GMM estimator of θ_0 is given by

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

We consider the null hypothesis $H_0 : R\theta_0 = r$ against the alternative $H_1 : R\theta_0 \neq r$, where R is a $p \times d_\theta$ matrix.[1] Then, under some regularity conditions, one can show that

$$\begin{aligned}\sqrt{n}(R\hat{\theta} - r) &= -R(\Gamma' W^{-1}\Gamma)^{-1}\Gamma' W^{-1}\sqrt{n} g_n(\theta_0) + o_p(1) \\ &\xrightarrow{d} N(0, \tilde{R}\Omega\tilde{R}'),\end{aligned} \tag{2}$$

where $\tilde{R} = R(\Gamma' W^{-1}\Gamma)^{-1}\Gamma' W^{-1}$, and

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

Motivated by this, one can construct the F test version of Wald statistic

$$F(\hat{\theta}) = \frac{n}{p}(R\hat{\theta} - r)'\left\{\hat{R}\hat{\Omega}\hat{R}\right\}^{-1}(R\hat{\theta} - r)', \tag{3}$$

where $\hat{R}$ and $\hat{\Omega}$ are estimators for $\tilde{R}$ and Ω, respectively. It is obvious that the t-statistic can be used when $p = 1$. Let $\hat{\Gamma}(\theta) = n^{-1}\sum_{i=1}^{n} \partial f_i(\theta)/\partial\theta'$, and assume that $\text{plim}_{n\to\infty}\hat{\Gamma}(\tilde{\theta}) = \Gamma$ for any $\sqrt{n}$-consistent estimator $\tilde{\theta}$. Then, it is not difficult to formulate $\hat{R}$ as

$$\hat{R} = R(\hat{\Gamma}(\hat{\theta})' W_n^{-1}\hat{\Gamma}(\hat{\theta}))^{-1}\hat{\Gamma}(\hat{\theta})' W_n^{-1},$$

and show that $\hat{R} \xrightarrow{p} \tilde{R}$. To construct $\hat{\Omega}$, however, one has to properly account for the cross-sectional dependence which is embodied in Ω. The matrix 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 individuals. For easy reference, we call Ω the global variance, following Sun and Kim (2015).

In this paper, we assume that the structure of cross-sectional dependence can be characterized by clustered samples, where each of cluster has an unknown dependence structure. More specifically, we assume that the set of true moment process $\{f_i(\theta_0), i = 1, 2, ..., n\}$ can be partitioned by G nonoverlapping clusters $\{f_i^g(\theta_0) : i = 1, ..., n_g\}$ for $g = 1, \ldots, G$. Here, n_g denotes the size of cluster g, and the total number of observations is $n = \sum_{g=1}^{G} n_g$. Utilizing the clustered data structure, one can define the following clustered covariance estimator (CCE):

$$\hat{\Omega}(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} 1(i, j \in \text{the same cluster}) f_i(\hat{\theta}) f_j(\hat{\theta})', \tag{4}$$

which is one of the most popular variance estimators in applied econometrics. To ensure that $\hat{\Omega}(\hat{\theta})$ is positive definite, we assume that $G \geq d_f$, and maintain this condition throughout the rest of the paper. In the context of the (exactly identified) linear regression model, $d_f = d_\theta$, this

[1] The asymptotic theory developed in our paper can be generalized to non-linear restrictions using the Delta method.

is equal to restrict that the number of regressors (d_θ) is no greater than the number of clusters (G), which serves as the effective number of sample sizes in CCE.

The conventional large-cluster (G) asymptotics, $G \to \infty$ such that $\max_{1\le g\le G} n_g/n \to 0$, shows that the CCE becomes concentrated at the true asymptotic variance, i.e., $\hat{\Omega}(\hat{\theta}) \xrightarrow{p} \Omega$. See, for example, White (1984), Carter et al. (2017), and Hansen and Lee (2019). As a direct consequence, the Wald statistic converges in distribution to a standard χ^2_p limit. To be more specific, we re-express the CCE formula in (4) as

$$\hat{\Omega}(\hat{\theta}) = \frac{1}{G}\sum_{g=1}^{G} \hat{u}_g \hat{u}_g \text{ with } \hat{u}_g = \left(\sqrt{\frac{n_g}{\bar{n}}}\right) \times \left\{ \frac{1}{\sqrt{n_g}} \sum_{i=1}^{n_g} f_i^g(\hat{\theta}) \right\}, \tag{5}$$

where $\bar{n} = G^{-1}\sum_{g=1}^{n} n_g$ is the average size of clusters. As $G \to \infty$, the large-G asymptotics controls the sampling variations in $\hat{u}_g$ via the law of large number across different clusters. In effect, the large-G asymptotics ignores the estimation error in the CCE despite $\hat{u}_g$ can possess high variation in finite samples. The expression in (5) indicates that the high variation in $\hat{u}_g$ can be large, when there exists only a small number (G) of large-sized (n_g) clusters, and the clusters are unequal in sizes, i.e., $n_g/\bar{n}$ differs across different g. In fact, there is ample simulation evidence that points out the poor finite-sample property of the conventional normal and chi-square inferences via the large-G approximations. See, for example, Cameron, Gelbach, and Miller (2008), Carter, Schnepel, and Steigerwald (2017), Mackinnon and Webb (2017), and Hwang (2020). All of these studies commonly present problems of the large-G asymptotics when a relatively small number of clusters have severely unbalanced sizes.

Motivated by this, this paper follows an alternative asymptotic path by assuming that there is a *fixed* number (G) of *large-sized* (n_g) clusters, e.g., Bester, Conley and Hansen (2011, BCH hereafter), Ibragimov and Müller (2010, 2015), and Canay et al. (2019). Also, our theory explicitly considers that clusters can vary in size asymptotically, which fills a gap between the existing asymptotic theory and empirical practice.

3 Fixed-cluster Asymptotics with Unbalanced Cluster-sizes

To formally characterize the asymptotic property of $\hat{\Omega}(\hat{\theta})$ under the fixed-G asymptotics, we further maintain the following high level conditions.

Assumption 1 *Assume the followings:*

i) The number of clusters G is fixed, so that the size of each cluster n_g grows to infinity with the total sample size n such that

$$\begin{pmatrix} \frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1} f_i^1(\theta_0) \\ \vdots \\ \frac{1}{\sqrt{n_G}}\sum_{i=1}^{n_G} f_i^G(\theta_0) \end{pmatrix} \xrightarrow{d} N\left(0, \begin{pmatrix} \Omega_1 & & 0 \\ & \ddots & \\ 0 & & \Omega_G \end{pmatrix}\right), \tag{6}$$

and $\Omega_g > 0$ for $g = 1, \ldots, G$.

ii) For each $g = 1, \ldots G$, we have

$$\frac{n_g}{n} \to \lambda_g > 0 \text{ such that } \sum_{g=1}^{G} \lambda_g = 1. \tag{7}$$

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

Assumption 3 *Assume the followings:*
(i) For each $g = 1, ..., G$, *let* $\Gamma_g(\theta) := \lim_{n_g \to \infty} E\left[\frac{1}{n_g}\sum_{i=1}^{n_g} \partial f_i^g(\theta)/\partial\theta'\right]$. *Then,*

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

where $\mathcal{N}(\theta_0)$ *is an open neighborhood of* θ_0 *and* $\|\cdot\|$ *is the Euclidean norm.*
(ii) $\Gamma_g(\theta)$ *is continuous at* $\theta = \theta_0$, *and for* $\Gamma_g = \Gamma_g(\theta_0)$,

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

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

Assumption 1-i) implies that the fixed-cluster structure consists of asymptotically independent G clusters, but allows each cluster has an arbitrary weak dependence structure. Under Assumption 1, it is straightforward the total (scaled) sum of moment process can be decomposed as the following form of cluster-wise sums

$$\begin{aligned} \sqrt{n} g_n(\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} f_i(\theta_0) &= \sum_{g=1}^{G} \sqrt{\frac{n_g}{n}} \left(\frac{1}{\sqrt{n_g}} \sum_{i=1}^{n_g} f_i^g(\theta_0) \right) \\ &= \sum_{g=1}^{G} \sqrt{\lambda_g} \left(\frac{1}{\sqrt{n_g}} \sum_{i=1}^{n_g} f_i^g(\theta_0) \right) (1 + o_p(1)), \end{aligned} \tag{8}$$

and the following cross-sectional CLT holds:

$$\sqrt{n} g_n(\theta_0) \xrightarrow{d} N(0, \Omega), \tag{9}$$

where $\Omega := \sum_{g=1}^{G} \lambda_g \Omega_g$. Assumption 2 is made for convenience, and primitive sufficient conditions are available from the standard GMM asymptotic theory, e.g. Newey and Mcfadden (1994). Assumption 3 is a uniform law of large numbers (ULLN), from which we obtain $\hat{\Gamma}(\hat{\theta}) = G^{-1}\sum_{g=1}^{G} \Gamma_g + o_p(1) = \Gamma + o_p(1)$. The homogeneity conditions in Assumptions 4 and 5 guarantee the asymptotic pivotality of the cluster-robust GMM statistic we consider. Similar assumptions are made in BCH (2011) and Sun and Kim (2015), which develop asymptotically valid F tests that are robust to spatial autocorrelation in the same spirit as our fixed-G asymptotics.

Let $Z_{p,g} \overset{\text{i.i.d}}{\sim} N(0, I_p)$ for $g = 1, ..., G$. By construction, $\sqrt{\lambda_g} Z_{p,g} \sim N(0, \lambda_g I_p)$ and $\sum_{g=1}^{G} \sqrt{\lambda_g} Z_{p,g} \sim N(0, I_p)$, due to $\sum_{g=1}^{G} \lambda_g = 1$. Define a $p \times p$ random matrix

$$\bar{\mathbb{S}}_{pp} = \sum_{g=1}^{G} \left(\sqrt{\lambda_g} Z_{p,g} - \lambda_g \sum_{g=1}^{G} \sqrt{\lambda_g} Z_{p,g} \right) \left(\sqrt{\lambda_g} Z_{p,g} - \lambda_g \sum_{g=1}^{G} \sqrt{\lambda_g} Z_{p,g} \right)'. \tag{10}$$

For $G \times 1$ vector $a = (a_1, \ldots, a_G)'$, we denote $\text{Diag}(a)$ a $G \times G$ diagonal matrix of a. Similarly, $\text{Diag}(\sqrt{a})$ is a diagonal matrix of $\sqrt{a} = (\sqrt{a_1}, \ldots, \sqrt{a_G})'$.

Lemma 1 *Under Assumptions 1–5, if H_0 is true, we have that: (i)*

$$\sqrt{n}(R\hat{\theta} - r) \xrightarrow{d} \tilde{\Lambda} \sum_{g=1}^{G} \sqrt{\lambda_g} Z_{p,g}; \tag{11}$$

$$\hat{R}\hat{\Omega}\hat{R} \xrightarrow{d} \tilde{\Lambda}\bar{\mathbb{S}}_{pp}\tilde{\Lambda}' \tag{12}$$

hold jointly, where $\sum_{g=1}^{G} \sqrt{\lambda_g} Z_{p,g} \perp \bar{\mathbb{S}}_{pp}$ and $\tilde{\Lambda}$ is a $p \times p$ matrix such that $\tilde{\Lambda}\tilde{\Lambda}' = \tilde{R}\Omega\tilde{R}'$. (ii) The following distributional equivalence holds

$$\bar{\mathbb{S}}_{pp} \overset{d}{=} \sum_{g=1}^{G} \kappa_g \tilde{Z}_{p,g} \tilde{Z}'_{p,g}, \tag{13}$$

where each $\tilde{Z}_{p,g}$ is independent copy of $Z_{p,g}$, and $\kappa'_g s$ are G eigenvalues of DD', with $G \times G$ matrix $D = (I_G - Diag(\lambda)11')Diag(\sqrt{\lambda})$.

The proof of Lemma 1 is provided in Appendix. Lemma 1(i) shows that, under the fixed-G asymptotics, the CCE matrix weakly converges to a random matrix limit $\tilde{\Lambda}\bar{\mathbb{S}}_{pp}\tilde{\Lambda}'$, which is proportional to the true variance $\tilde{R}\Omega\tilde{R}'$ in a matrix sense and is independent of the limit of $\sqrt{n}(R\hat{\theta} - r)$. We can also see from this result that the fixed-G asymptotics accommodates the effect of cluster size heterogeneity represented by $\lambda = (\lambda_1, \ldots, \lambda_G)'$ and the effect of the estimation error in $\hat{\theta}$ captured by the subtraction term in $\bar{\mathbb{S}}_{pp}$. This is in sharp contrast to the large-G asymptotics that ignores these two effects by assuming $\max_g n_g/n \to 0$ to achieve consistency (e.g., Hansen and Lee, 2019). Hence, the fixed-G asymptotics is expected to yield more accurate inference results than the large-G asymptotics, and the gap between these two asymptotics will become substantial when the number of clusters is small and/or the degree of cluster size heterogeneity is high. We will examine this rigorously in Theorem 3 below.

Lemma 1(ii) presents another distributional representation of $\bar{\mathbb{S}}_{pp}$. (13) indicates that $\bar{\mathbb{S}}_{pp}$ is equal in distribution to the weighted sum of independent Wishart distributions. The weights κ'_gs are the eigenvalues of the $G \times G$ matrix DD',which depends only on the cluster-size ratios λ_g.

Since we allow λ_g's to have asymptotically unbalanced cluster-sizes, Lemma 1 generalizes the existing fixed-G limit with the equally sized clusters. We note that the fixed-G limiting distribution in (11) has also been obtained by Proposition 2-iii) in BCH (2011), but they consider only linear regressions and M-estimators for an exactly identified model. Also, their results do not cover Lemma 1(ii), which is a key device to establish the stochastic expansion of fixed-G asymptotics and to develop our convenient F approximation to the fixed-G limiting distribution of the test statistic $F(\hat{\theta})$.

It is straightforward from Lemma 1(i) that given G fixed

$$F(\hat{\theta}) \xrightarrow{d} \mathbb{F}_\infty = \frac{1}{p} \left(\sum_{g=1}^{G} \sqrt{\lambda_g} Z_{p,g} \right)' \bar{\mathbb{S}}_{pp}^{-1} \left(\sum_{g=1}^{G} \sqrt{\lambda_g} Z_{p,g} \right) \tag{14}$$

if H_0 is true. Note that $\sum_{g=1}^{G} \sqrt{\lambda_g} Z_{p,g} \sim N(0,1)$. Thus, the fixed-$G$ limit of $F(\hat{\theta})$ is a quadratic form in a standard normal vector $\sum_{g=1}^{G} \sqrt{\lambda_g} Z_{p,g}$ with a random and independent weighting matrix $\bar{\mathbb{S}}_{pp}$. We summarize the fixed-G asymptotics of $F(\hat{\theta})$ in the following proposition.

Proposition 2 *Under Assumptions 1–5, if H_0 is true, we have*

$$P\left(F(\hat{\theta}) \leq x\right) = P\left(\mathbb{F}_\infty \leq x\right) + o\left(1\right)$$

as $n \to \infty$.

The proof is omitted because it is directly from Lemma 1(i). Proposition 2 shows the validity of the fixed-G asymptotic test. A special case of this test is when all clusters are asymptotically equal in sizes, i.e., $\lambda_g = 1/G$ for $g = 1, \ldots, G$. In this case, we have

$$\bar{\mathbb{S}}_{pp} \overset{d}{=} \frac{1}{G}\sum_{g=1}^{G-1} \tilde{Z}_{p,g}\tilde{Z}'_{p,g} \sim \frac{1}{G}\mathbb{W}_p(G-1, I_p),$$

where $\mathbb{W}_p(G-1, I_p)$ denotes a Wishart distribution with the degree of freedom $G-1$ and $p \times p$ scale matrix I_p. Then, $\mathbb{F}_\infty$ follows Hotelling's T^2distribution. According to the well-known relationship between the T^2 and standard F distributions such that $\mathbb{F}_\infty \overset{d}{=} (G/G-p)\mathcal{F}_{p,G-p}$, we can conduct a valid F test. This result is consistent with BCH (2011) and Hwang (2020), which covers the case of balanced cluster sizes.

In practice, however, it is more likely that cluster sizes are heterogenous. Then, as is implied in (14), the fixed-G limit of $F(\hat{\theta})$ is still pivotal but becomes nonstandard. We consider two different approaches for implementing the fixed-G asymptotic test in this case. The first is to employ simulation. The relative cluster sizes λ_g is readily obtained from data n_g/n and practitioners can simulate one random draw from $\mathbb{F}_\infty$ by generating G i.i.d. standard normal vectors $\{Z_{p,g}\}$ and by plugging them into the representation in (14). The other is to approximate the fixed-G limiting distribution with a standard F distribution with some scale factor, which is discussed in the next section.

Define subcomponents of the random matrix $\bar{\mathbb{S}}_{pp}$

$$\bar{\mathbb{S}}_{pp} \overset{d}{=} \sum_{g=1}^{G} \kappa_g \tilde{Z}_{p,g}\tilde{Z}'_{p,g} = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix}, \tag{15}$$

where $s_{11} \in \mathbb{R}$ and $s_{22} \in \mathbb{R}^{(p-1)\times(p-1)}$. Denote $s_{11\cdot 2} = s_{11} - s_{12}s_{22}^{-1}s_{21}$. In Lemma XX in Appendix, we show that $\mathbb{F}_\infty$ has another representation below

$$\mathbb{F}_\infty \overset{d}{=} \frac{1}{p}\frac{\chi^2_p}{s_{11\cdot 2}}, \text{ where } \chi^2_p \perp s_{11\cdot 2}. \tag{16}$$

The expression in (16) shows that the fixed-G limit of $F(\hat{\theta})$ is equal in distribution to a scale-mixture of chi-square random variable. Since $s_{11\cdot 2}$ is independent of the chi-square random variable in the numerator, the cdf of $\mathbb{F}_\infty$ can be written as

$$P\left(p\mathbb{F}_\infty \leq x\right) = E1\left\{\chi^2_p \leq s_{11\cdot 2}x\right\} = \Xi_p\left(s_{11\cdot 2}x\right),$$

where Ξ_p denotes the cdf of a chi-square random variable with the degree of freedom p. Note that the random scaling factor $s_{11\cdot 2}$ depends on κ_g, which is a function of the relative cluster size ratios λ_g. In Lemma 8 (a) and (b) in Appendix, we derive $\sum_{g=1}^{G}\kappa_g = 1 - O(G^{-1})$ and $\sum_{g=1}^{G}\kappa_g^2 = O(G^{-1})$, which leads to $\bar{\mathbb{S}}_{pp} \xrightarrow{p} I_p$ as $G \to \infty$. Therefore, $s_{11\cdot 2}$ converge in probability to 1 as $G \to \infty$, and the fixed-G asymptotics reduce to the standard χ^2_p/p limit. This finding enables us to expand $\Xi_p\left(s_{11\cdot 2}x\right)$ around $\Xi_p\left(x\right)$ to establish the following theorem.

Theorem 3 *As $G \to \infty$, we have*

$$P\left(p\mathbb{F}_\infty \leq x\right) = \Xi_p\left(x\right) + \left(\Xi_p''\left(x\right)x^2 - \Xi_p'\left(x\right)px\right)\sum_{g=1}^{G}\lambda_g^2 + o\left(G^{-1}\right). \tag{17}$$

The proof is provided in Appendix. Theorem 3 characterizes the difference of the fixed-G limiting distribution from the chi-square distribution. Since $\Xi_p'\left(x\right) > 0$ and $\Xi_p''\left(x\right) < 0$ when x is large, the theorem explicitly shows the fixed-G asymptotic test addresses the overrejection of chi-square approximation by accounting for the effects of the number of restrictions in H_0, the number of clusters, and the degree of cluster size unboundedness on the sampling distribution of $F(\hat{\theta})$. For each G, $\sum_{g=1}^{G}\lambda_g^2$ is minimized with equally sized clusters, and becomes larger as variation in cluster sizes grows. Thus, the chi-square approximation based on the large-G asymptotics would not be reliable, if the variation of cluster sizes is large. This finding is related with Carter, et al. (2017) that show cluster heterogeneity governs the behavior of the CCE and develop a measure of the adequacy of chi-square approximation, which they name as the "effective number of clusters," by scaling down the actual number of clusters based on the degree of cluster heterogeneity. We examine the relation between Carter, et al. (2017)'s effective number of clusters and our approach further in Section 5.1.

4 F-approximation of Non-standard Fixed-G Limits

The fixed-G asymptotic test yields more accurate inference results than the chi-square test based on the large-G asymptotics, and the gain becomes substantial when the number of restrictions in H_0 is large, the number of clusters is small, and/or the degree of cluster size unboundedness is high. On the other hand, the former is not as convenient to use as the latter in practice since extra computation is required to simulate critical values unless all clusters are equally sized. To address this, in this section, we propose an F test which obtains critical values by approximating the fixed-G limit of $F(\hat{\theta}_1)$, $\mathbb{F}_\infty$, by a scaled F distribution, and examine the properties.

Recall that

$$\bar{\mathbb{S}}_{pp} \overset{d}{=} \sum_{g=1}^{G}\kappa_g\tilde{Z}_{p,g}\tilde{Z}_{p,g}',$$

where $\tilde{Z}_{p,g} \overset{i.i.d}{\sim} N(0, I_p)$. Since each $\tilde{Z}_{p,g}\tilde{Z}_{p,g}'$ is a Wishart random matrix $\mathbb{W}_p\left(1, I_p\right)$, $\bar{\mathbb{S}}_{pp}$ is the weighted sum of independent Wishart random matrices. Suppose that $\bar{\mathbb{S}}_{pp}$ is approximated by

$$\sum_{g=1}^{G}\kappa_g\tilde{Z}_{p,g}\tilde{Z}_{p,g}' \overset{d}{\approx} \frac{\mu}{G^*}\sum_{g=1}^{G^*}\tilde{Z}_{p,g}\tilde{Z}_{p,g}' \tag{18}$$

for some positive constant μ and the effective degree of freedom G^*. Our approximation starts by finding proper values of μ and G^* such that the approximating Wishart distribution in (18) has the first two moments in common with $\bar{\mathbb{S}}_{pp}$. Then, it follows that

$$\mu = \sum_{g=1}^{g}\kappa_g \text{ and } G^* = \left\lfloor \frac{\left(\sum_{g=1}^{G}\kappa_g\right)^2}{\sum_{g=1}^{G}\kappa_g^2} \right\rfloor. \tag{19}$$

The idea of matching the first two moments for distributional approximation is suggested by Satterthwaite (1946) who approximates the distribution of a weighted sum of chi-square random variables with a single chi-square random variable with an adjusted degree of freedom. Recently Sun (2014) employs this idea to develop F-approximation to the fixed-b limiting distribution in the HAR inference. We apply the approach to the fixed-G asymptotics in the cluster-robust inference setting.

Since $\{\kappa_g\}$ are eigenvalues of DD' with $D = (I_G - \text{Diag}(\lambda)11')\text{Diag}(\sqrt{\lambda})$, direct calculations give us

$$\mu = 1 - \sum_{g=1}^{G} \lambda_g^2 \text{ and } G^* = \left\lfloor \frac{\left(1 - \sum_{g=1}^{G} \lambda_g^2\right)^2}{\sum_{g=1}^{G} \lambda_g^2 + \left(\sum_{g=1}^{G} \lambda_g^2\right)^2 - 2\sum_{g=1}^{G} \lambda_g^3} \right\rfloor. \tag{20}$$

Combining the distributional approximation in (18)-(20) with our previous result in (14), we have

$$\mathbb{F}_\infty \overset{d}{\approx} \frac{1}{\mu p}\left[\left(\sum_{g=1}^{G}\sqrt{\lambda_g}Z_{p,g}\right)' \left[\frac{1}{G^*}\mathbb{W}_p\left(I_p, G^*\right)\right]^{-1}\left(\sum_{g=1}^{G}\sqrt{\lambda_g}Z_{p,g}\right)\right]$$

where $\mathbb{W}_p(I_p, G^*)$ is independent of $\sum_{g=1}^{G}\sqrt{\lambda_g}Z_{p,g}$. The right side is a scaled Hotelling's T^2-distribution with the scale factor $(\mu p)^{-1}$ and the degrees of freedom p and G^*. Using the relation between Hotelling's T^2 distribution and F distribution, we finally approximate $\mathbb{F}_\infty$ by a scaled F distribution as follows:

$$\mathbb{F}_\infty \overset{d}{\approx} \underbrace{\frac{G^*}{\mu\left(G^* - p + 1\right)}}_{:=\tau}\mathcal{F}_{p,G^*-p+1}. \tag{21}$$

If all clusters are asymptotically equally-sized, we have $\mu = 1 - 1/G$ and $G^* = G - 1$ from (20). This leads to $\tau\mathcal{F}_{p,G^*-p+1} = \frac{G}{G-p}\mathcal{F}_{p,G-p}$, and our F approximation reduces to the fixed-G asymptotic distribution and there is no approximation error.

With the selected μ and G^* in (20), the proof of Theorem 4 in the appendix states the following expansion of the F approximation

$$P\left(\frac{pG^*}{\mu\left(G^* - p + 1\right)}\mathcal{F}_{p,G^*-p+1} \leq x\right) = \Xi_p\left(x\right) + \left[\Xi_p''\left(x\right)x^2 - \Xi_p'\left(x\right)px\right]\sum_{g=1}^{G}\lambda_g^2 + o\left(\frac{1}{G}\right). \tag{22}$$

Comparison between (22) and Theorem 3 indicates that the difference of the approximate F-distribution in (21) from the distribution of $\mathbb{F}_\infty$ is $o\left(G^{-1}\right)$. This implies our F test captures the primary impacts of the number of restrictions, the number of clusters, and the degree of cluster size unboundedness on the sampling distribution of $F\left(\hat{\theta}\right)$ that the chi-square test ignores.

The theorem below formally justifies our F approximation.

Theorem 4 *Let $\mathbb{F}_\infty$ be the non-standard fixed-G limits in (16). Then, as $G \to \infty$, we have*

$$P\left(p\mathbb{F}_\infty \leq x\right) = P\left(\frac{pG^*}{\mu\left(G^* - p + 1\right)}\mathcal{F}_{p,G^*-p+1} \leq x\right) + o\left(\frac{1}{G}\right),$$

We can relate our F approximation to the literature that provide t-distribution approximations to the distribution of a cluster robust t-statistic in the context of linear models. See, for example,

Bell and McCaffrey (2002), Donald and Lang (2007), Imbens and Kolesar (2016) and MacKinnon and Webb (2017). These procedures are encouraged to use in empirical research (e.g., Imbens and Kolesar, 2016) since they yield substantial improvement in sizes than normal approximation without extra computation costs. While they are practically very useful, one crucial issue is that, to the best of our knowledge, no theoretical foundation has been established except the case that clusters are equally sized. No theory shows their asymptotic validity or compares their asymptotic behavior against the asymptotic normal test. We fill the gap in the literature by addressing this problem. Motivated by F approximation methods, we develop our F test built on the approximation to fixed-G asymptotics, and show that the approximation error is small.

For implementing the F test, we define

$$\lambda_{ng} = \frac{n_g}{n},\ \mu_n = 1 - \sum_{g=1}^{G} \lambda_{ng}^2,\ \text{and}\ G_n^* = \left\lfloor \frac{\left(1 - \sum_{g=1}^{G} \lambda_{ng}^2\right)^2}{\sum_{g=1}^{G} \lambda_{ng}^2 + \left(\sum_{g=1}^{G} \lambda_{ng}^2\right)^2 - 2\sum_{g=1}^{G} \lambda_{ng}^3} \right\rfloor.$$

It is straightforward that $\lambda_{ng} = \lambda_g\left(1 + o\left(1\right)\right)$, $\mu_n = \mu\left(1 + o\left(1\right)\right)$ and $G_n^* = G^*\left(1 + o\left(1\right)\right)$ as $n \to \infty$. We also modify the effective degree of freedom by setting $G_{0n}^* = \max\left(3, \lfloor G_n^* - p + 1 \rfloor\right)$ to ensure the mean of $\mathcal{F}_{p,G_{0n}^*}$ exists. Let $\mathcal{F}_{p,G_{0n}^*}^{\alpha}$ denote the $1-\alpha$ quantile of the F distribution with the degrees of freedom p and G_{0n}^*. Based on Theorem 4, we propose an F-test based on the test statistic $F\left(\hat{\theta}\right)$ that uses

$$\tau_n \mathcal{F}_{p,G_{0n}^*}^{\alpha},\ \text{with}\ \tau_n = \frac{G_n^*}{\mu_n \max\left(1, G_n^* - p + 1\right)}$$

as the critical value with nominal size α.

It is obvious from Theorem 4 that we can conduct t-test when $p = 1$. The t-statistic is given by

$$t\left(\hat{\theta}\right) = \frac{\sqrt{n}\left(R\hat{\theta} - r\right)}{\sqrt{\hat{R}\hat{\Omega}\hat{R}}},$$

and we obtain critical values from $\tau_n t\left(G_{0n}^*\right)$, where $t\left(G_{0n}^*\right)$ denotes the t distribution with the degree of freedom G_{0n}^*.

5 Application to linear regression model

As a special case of our general GMM setting, consider a linear regression with a cluster structure:

$$y_{i_g} = x_i'\theta_0 + \epsilon_{i_g} \text{ for } i_g = 1, \ldots, n_g \text{ and } g = 1, \ldots, G. \tag{23}$$

With the moment function $f_i^g(\theta_0) = x_{i_g}(y_{i_g} - x_{i_g}'\theta)$, the moment condition $E[f_i^g(\theta_0)] = 0$ is exactly identified, so the GMM weight matrix W_n can be trivially set as I_{d_θ}. Assuming the equal cluster sizes, $n_1 = \ldots = n_g$, the model (23) is studied in Hansen (2007), Stock and Watson (2008), and BCH (2011). To apply our fixed-G asymptotic theory, we denote g-cluster specific regressor matrix $X_g = \left[x_1, \ldots, x_{n_g}\right]'$. Then, the $n \times d_\theta$ covariance matrix X for all observations is represented as $X = \left[X_1', \ldots, X_G'\right]'$. Similarly, we define $y_g, \epsilon_g \in \mathbb{R}^{n_g}$, $y = (y_1, \ldots, y_g)'$, and $\epsilon = (\epsilon_1, \ldots \epsilon_g)'$. It is straightforward that the pooled OLS (GMM) estimator is $\hat{\theta} = \left(\sum_{g=1}^{G} X_g'X_g\right)^{-1}\left(\sum_{g=1}^{G} X_g'y_g\right)$.

Denote $\Gamma = E[x_i x_i']$. With $\hat{\Gamma} = n^{-1}X'X$, the cluster-robust variance estimator is formulated as

$$\hat{V}_R := R(X'X/n)^{-1}\left[\frac{1}{n}\sum_{g=1}^{G} X_g'\hat{\epsilon}_g\hat{\epsilon}_g'X_g\right](X'X/n)^{-1}R'.$$

The estimator $\hat{V}_R$ is a linear regression version of the CCE matrix $\hat{R}\hat{\Omega}\hat{R}$ whose asymptotic properties are developed in Section 3. Under Assumptions 1–5 in the model (23), our fixed-cluster asymptotic theory shows that

$$\hat{V}_R \xrightarrow{d} \tilde{\Lambda}\bar{\mathbb{S}}_{pp}\tilde{\Lambda}'. \tag{24}$$

Also, the F-approximation theory developed in Section 4 shows that

$$F(\hat{\theta}) = \frac{1}{p}n(R\hat{\theta}-r)'\left(\frac{G^*}{\mu\,(G^*-p+1)}\hat{V}_R\right)^{-1}(R\hat{\theta}-r)' \overset{d}{\approx} \mathcal{F}_{p,G^*-p+1};$$

$$\mu = 1-\sum_{g=1}^{G}\lambda_g^2 \text{ and } G^* = \left\lfloor \frac{\left(1-\sum_{g=1}^{G}\lambda_g^2\right)^2}{\sum_{g=1}^{G}\lambda_g^2+\left(\sum_{g=1}^{G}\lambda_g^2\right)^2-2\sum_{g=1}^{G}\lambda_g^3}\right\rfloor.$$

Here, μ is the bias adjustment in the CCE, and G^* is the effective degree of freedom (edf) adjustment, where both of them are functions of the relative cluster sizes λ_g. Together with the F-critical value, the two adjustment terms are key components to improve the conventional F-statistics with chi-square critical value. In the next subsections, we discuss how these our bias adjustment and edf are connected to the existing literature in cluster-robust inference.

5.1 Effective number of clusters in Carter et al. (2017)

When clusters vary in sizes in (23), Carter et al. (2017) points out that the inferences via the conventional large-G asymptotics can be very poor in finite samples. They show that the finite-sample behaviors of F and t statistics cruically depend on the variations in cluster sizes. Employing the large-G asymptotics, Carter et al. (2017) develop a sample-specific adjustment to the number of clusters G, which they termed as the effective number of clusters. It is shown that the standard normal/chi-square inferences can be unreliable when the effective number of clusters is substantially smaller than G. For this reason, Carter et al. (2017) recommend using the effective number of clusters as a diagnostic measure of the large-G asymptotics.

To illustrate the main idea of the effective number of clusters in Carter et a. (2017), suppose that we are interested in a single null hypothesis $H_0 : R\hat{\theta} = r$ where R is a d_θ-dimensional normalized selection vector, i.e. $||R||^2 = 1$. Also, define the infeasible cluster estimator which is a function of the true moment $\{x_{i_g}\epsilon_{i_g}\}$

$$\tilde{V}_R := R(X'X/n)^{-1}\left[\frac{1}{n}\sum_{g=1}^{G} X_g'\epsilon_g\epsilon_g'X_g\right](X'X/n)^{-1}R'.$$

Note that $\tilde{V}_R$ is unbiased in that $E\left[\tilde{V}_R|X\right] = V_R$. Define $A_g = R(X'X)^{-1}(X_g'X_g)$ and $\hat{\theta}_g = (X_g'X_g)^{-1}X_g'y_g$ which is the OLS estimator based only on the observations for the cluster g. Then, one can re-express $\tilde{V}_R$ by

$$\tilde{V}_R = n\sum_{g=1}^{G} A_g\left(\hat{\theta}_g-\theta\right)\left(\hat{\theta}_g-\theta\right)'A_g'.$$

Carter et al. (2017) shows that

$$E\left[\left[\frac{\tilde{V}_R - V_R}{V_R}\right]^2 \middle| X\right] \leq \frac{1+\Psi(V_R,X)}{G}\left[2+\frac{M_4-3}{n^*}\right], \tag{25}$$

where M_4 is a positive constant, and $\Psi(V_R, X)$ is a sample specific quantity which is defined as:

$$\Psi(V_R,X) = \frac{1}{G}\sum_{g=1}^{G}\frac{(\gamma_g-\bar{\gamma})^2}{\bar{\gamma}^2};$$

$$\gamma_g = RA_gVar(\hat{\theta}_g|X)A_g'R' \text{ and } \bar{\gamma} = \frac{1}{G}\sum\gamma_g$$

This result indicates that the relative mean squared error of $\tilde{V}_R$ is proportional to

$$G_{\text{eff}} = \left\lceil \frac{G}{1+\Psi(V_R,X)} \right\rceil,$$

which is termed as the effective number of clusters in Carter et al. (2017). By construction, if clusters are completely homogeneous, including $n_1 = \ldots = n_g$, $\Psi(V_R, X) = 0$ and $G_{\text{eff}} = G$. If there is heterogeneity across clusters, then $\Psi(V_R, X) > 0$ and $G_{\text{eff}} < G$. Thus, a larger gap between G and G_{eff} indicates a greater downward adjustment in the number of clusters and suggests standard normal/chi-square critical values are inappropriate. This is related with our observation from Theorem 3 that shows under homogeneity conditions in Assumptions 4 and 5 the gap between the fixed-G asymptotics and the chi-square approximation grows with $\sum_{g=1}^{G}\lambda_g^2$ that depends on the degree of cluster size heterogeneity.

Although Carter et al. (2017) focuses on using G_{eff} as a diagnostic measure to check the reliability of large-G asymptotics, practitioners can naturally question about the possibility of using t critical values from G_{eff}. Since the conventional normal approximation is well known to overreject with unbalanced clusters, it might be logically possible to reflect the data-driven measure G_{eff} in asymptotic critical values. The possibility of using a t distribution with the degree of freedom G_{eff} in inference is also conjectured in Carter et al. (2017), and numerically investigated by Mackinnon and Webb (2017). Still, the theoretical justification of this method has not been discussed in the literature. To fill this gap, we first discuss that G_{eff} can be viewed as a finite sample version of equivalent degree of freedom under the normality assumption, and then show that G_{eff} and our equivalent degree of freedom, $\tilde{G}$, are asymptotically equivalent.

Assume $\epsilon|X \sim N(0,\Sigma)$ where Σ has a cluster structure such that $E\left(\epsilon_g\epsilon_g'\right) = \Sigma_g$ and $E\left(\epsilon_g\epsilon_h'\right) = 0$ if $g \neq h$. Let $\epsilon_g = \Sigma_g^{1/2}Z_g$ with $Z_g \sim N\left(0, I_{n_g}\right)$. Due to the following distributional equivalence

$$A_g\left(\hat{\beta}_g - \beta\right)|X \overset{d}{=} \underbrace{R(X'X)^{-1}X_g'\Sigma_g^{1/2}}_{1\times n_g}\underbrace{Z_g}_{n_g\times 1}|X \sim N\left(0,\gamma_g\right),$$

we have

$$\frac{1}{n}\tilde{V}_R|X = \sum_{g=1}^{G}\left(A_g\left(\hat{\beta}_g-\beta\right)\right)^2|X \overset{d}{=} \sum_{g=1}^{G}\gamma_g Z_{1,g}^2$$

where $\{Z_{1,g}\}$ are independent standard normal random variables. Combining this result with $n^{-1}V_R = \sum_{g=1}^{G} \gamma_g$ leads to

$$\frac{\tilde{V}_R}{V_R}\bigg| X \overset{d}{=} \sum_{g=1}^{G} a_g Z_{1,g}^2 \text{ with } a_g = \frac{\gamma_g}{\sum_{g=1}^{G} \gamma_g}.$$

If $\sum_{g=1}^{G} a_g \chi_g^2$ is approximated with a scaled chi-square distribution by matching the first two moments as our approach does, we have

$$\sum_{g=1}^{G} a_g Z_{1,g}^2 \overset{d}{\approx} \frac{1}{G_0^*} \chi_{G_0^*}^2 \text{ with } G_0^* = \frac{\left(\sum_{g=1}^{G} \gamma_g\right)^2}{\sum_{g=1}^{G} \gamma_g^2}.$$

Going back to Carter et al.'s (2017) setup, under the normality assumption (25) is rewritten as

$$\begin{aligned} E\left[\left[\frac{\tilde{V}_R - V_R}{V_R}\right]^2 \middle| X\right] &= 2\frac{\sum_{g=1}^{G} \gamma_g^2}{\left(\sum_{g=1}^{G} \gamma_g\right)^2} \\ &= 2\frac{1 + \Psi(V_R, X)}{G} = 2\frac{1}{G_{\text{eff}}}, \end{aligned}$$

and it is straightforward that $G_{\text{eff}} = G_0^*$. Therefore, G_{eff} may be viewed as the equivalent degree of freedom to approximate $\tilde{V}_R/V_R$ by a single chi-square distribution.

The next proposition strengthens the theoretical connection between the equivalent number of clusters in Carter et al.'s (2013) and our equivalent degree of freedom by showing that G_{eff} is asymptotically equivalent to $\tilde{G}$, which is derived in our F-approximation theory.

Proposition 5 *Let Assumptions 1–5 hold in (23). Then, with $p = 1$, we have*

$$G_{\mathit{eff}} = \left\lceil \frac{G}{1 + \Psi(V_R, X)} \right\rceil = \tilde{G} + o_p(1),$$

as $n \to \infty$ such that G is fixed.

Proposition 5 together with distributional approximation in (**??**) may suggests using the scaled t statistic $G_{\text{eff}}^{-1/2} \left(R\hat{\theta} - r\right) / \sqrt{\hat{V}_R}$ and critical values from a t distribution with the degree of freedom G_{eff}. Note that this is different from our F approximation in that the former does not account for the bias of $\hat{V}_R$ by relying on the infeasible cluster variance estimator which is unbiased.

5.2 Fixed-G asymptotics and F-approximation with Jackknife bias corrected CCE

One weakness of $\hat{V}_R$ under large G asymptotics is that the bias from using the estimated residuals $\hat{\epsilon}_g = (\hat{\epsilon}_1, \ldots, \hat{\epsilon}_{n_g})'$ instead of the true ϵ_g is not taken into account, and this tends to cause overrejection in finite samples. This is mainly because $E[\hat{\epsilon}_g \hat{\epsilon}_g']$ is different from $E[\epsilon_g \epsilon_g']$ for each g. In the i.i.d setting, a number of different modifications of Eicker-Huber-White variance estimator

have been proposed to deal with this problem. See, for example, Mackinnon and White (1985), Angrist and Pischke (2009), and Hausman and Palmer (2011). In the cluster variance estimation setting, Bell and McCaffrey (2002) and Imbens and Kolesar (2016) consider modified cluster estimators that extend bias-adjustments in Mackinnon and White (1985). In this section, we establish the fixed-G asymptotics and its F approximation for the test based on the Jackknife CCE which is one of bias corrected CCEs considered by Bell and McCaffrey (2002).

Define $n_g \times n_g$ matrix $P_{gg} = X_g(X'X)^{-1}X_g'$ corresponding to cluster g. The Jackknife bias corrected CCE is

$$\hat{V}_{R,JK} = R(X'X/n)^{-1}\left[\frac{1}{n}\sum_{g=1}^{G} X_g'\hat{\epsilon}_g^*\hat{\epsilon}_g^{*\prime}X_g\right](X'X/n)^{-1}R'; \tag{26}$$

$$\hat{\epsilon}_g^* = \sqrt{\frac{n-n_g}{n}}(I_{n_g} - P_{gg})^{-1}\hat{\epsilon}_g.$$

For the transformed residuals in $\hat{V}_{R,JK}$, one can show that $\hat{\epsilon}_g^* = y_g - X_g'\hat{\theta}_{-g}$, where $\hat{\theta}_{-g}$ is the OLS estimator of θ based on all observations except the g-th cluster. Thus, the use of the transformed residual $\hat{\epsilon}_g^*$ leads to the corresponding $\hat{V}_{R,JK}$ that equal the delete-one-cluster jackknife estimate of the variance of the OLS estimator. The associated test statistic is given by

$$F_{JK}\left(\hat{\theta}\right) = \frac{1}{p}n\left(R\hat{\theta} - r\right)'\hat{V}_{R,JK}\left(R\hat{\theta} - r\right). \tag{27}$$

Define a $p \times p$ random matrix

$$\tilde{\mathbb{S}}_{pp} = \sum_{g=1}^{G}\frac{1}{(1-\lambda_g)}\left(\sqrt{\lambda_g}Z_{p,g} - \lambda_g\sum_{g=1}^{G}\sqrt{\lambda_g}Z_{p,g}\right)\left(\sqrt{\lambda_g}Z_{p,g} - \lambda_g\sum_{g=1}^{G}\sqrt{\lambda_g}Z_{p,g}\right)'.$$

In the next theorem, we establish the fixed-G asymptotics of $\hat{V}_{R,JK}$ and $F_{JK}\left(\hat{\theta}\right)$.

Theorem 6 *Let Assumptions 1–5 hold in (23). Then, as $n \to \infty$ such that G is fixed, we have*

$$\hat{V}_{R,JK} \xrightarrow{d} \tilde{\Lambda}\tilde{\mathbb{S}}_{pp}\tilde{\Lambda}', \tag{28}$$

where $\tilde{\Lambda}$ is a $p \times p$ matrix such that $\tilde{\Lambda}\tilde{\Lambda}' = R(\Gamma^{-1}\Omega\Gamma^{-1})R'$, and if H_0 is true

$$F_{JK}\left(\hat{\theta}\right) \xrightarrow{d} \mathbb{F}_{JK,\infty} = \left(\sum_{g=1}^{G}\sqrt{\lambda_g}Z_{p,g}\right)'\tilde{\mathbb{S}}_{pp}^{-1}\left(\sum_{g=1}^{G}\sqrt{\lambda_g}Z_{p,g}\right). \tag{29}$$

Together with (24), the first result in Theorem 6 shows that the Jackknife bias corrected CCE weakly converges to a non-standard limit $\tilde{\Lambda}\tilde{\mathbb{S}}_{pp}\tilde{\Lambda}'$, which is different from the limit of the original CCE formula $\tilde{\Lambda}\bar{\mathbb{S}}_{pp}\tilde{\Lambda}'$. For $\Phi_g = \sqrt{\lambda_g}Z_{p,g} - \lambda_g\sum_{g=1}^{G}\sqrt{\lambda_g}Z_{p,g}$, note that

$$E\left[\Phi_g\Phi_g'\right] = E\lambda_g Z_{p,g}Z_{p,g}' - 2E\left[\lambda_g\left(\sum_{g=1}^{G}\sqrt{\lambda_g}Z_{p,g}\right)\left(\sqrt{\lambda_g}Z_{p,g}\right)'\right]$$

$$+ \lambda_g^2 Var\left(\sum_{g=1}^{G}\sqrt{\lambda_g}Z_{p,g}\right) = \lambda_g(1-\lambda_g)\cdot I_p,$$

which implies that

$$E[\tilde{\Lambda}\bar{\mathbb{S}}_{pp}\tilde{\Lambda}'] = \tilde{\Lambda}\sum_{g=1}^{G} E\left[\Phi_g\Phi_g'\right]\tilde{\Lambda}' = \mu\cdot R(\Gamma^{-1}\Omega\Gamma^{-1})R', \text{ where } \mu = 1-\sum_{g=1}^{G}\lambda_g^2,$$

and

$$\begin{aligned} E[\tilde{\Lambda}\tilde{\mathbb{S}}_{pp}\tilde{\Lambda}'] &= \tilde{\Lambda}\sum_{g=1}^{G} E\left[\frac{\Phi_g\Phi_g'}{(1-\lambda_g)}\right]\tilde{\Lambda}' = \left(\sum_{g=1}^{G}\frac{\lambda_g(1-\lambda_g)}{(1-\lambda_g)}\right)\cdot\tilde{\Lambda}\tilde{\Lambda}' \\ &= \left(\sum_{g=1}^{G}\lambda_g\right)\times R(\Gamma^{-1}\Omega\Gamma^{-1})R' = R(\Gamma^{-1}\Omega\Gamma^{-1})R'. \end{aligned} \tag{30}$$

Thus, $\hat{V}_{R,JK}$ is asymptotically unbiased with $E[\tilde{\Lambda}\tilde{\mathbb{S}}_{pp}\tilde{\Lambda}'] = R(\Gamma^{-1}\Omega\Gamma^{-1})R'$. In contrast, the original CCE formula $\hat{V}_R$ is asymptotically biased with $E[\tilde{\Lambda}\bar{\mathbb{S}}_{pp}\tilde{\Lambda}'] = \mu\cdot R(\Gamma^{-1}\Omega\Gamma^{-1})R'$. This result implies that our fixed-G asymptotic theory is able to explicitly show the built-in bias adjustment functions in the modified formula $\hat{V}_{R,JK}$. In other words, the bias-adjusted functions in $\hat{V}_{R,JK}$ are to the scale adjusted CCE $\mu^{-1}\hat{V}_{R,}$ appeared in our F-approximation theory.

The second result in Theorem 6 presents the fixed-G asymptotics of $F_{JK}\left(\hat{\theta}\right)$, which is directly from 28. As a special case, if cluster sizes are homogenous, simple manipulations yield

$$\tilde{\mathbb{S}}_{pp} \sim \frac{1}{G-1}\mathbb{W}\left(G-1, I_p\right) \text{ and } \mathbb{F}_{JK,\infty} \sim \frac{G-1}{G-p}F\left(p, G-p\right),$$

so critical values are obtained from a standard F distribution. When cluster sizes are heterogenous, $\mathbb{F}_{JK,\infty}$ is nonstandard. In this case, as the fixed-G test with the standard CCE, critical values can be obtained either by simulation or by F-approximation. The simulation procedure is similar to the one for the fixed-G asymptotic test with the original CCE in Section 3. We first simulate G i.i.d. $p\times 1$ standard normal vector $\{Z_{p,g}\}$, and then plug in them into the representation in (29) to generate one draw from the distribution of $\mathbb{F}_{JK,\infty}$. The theorem below formalizes the F-approximation to $\mathbb{F}_{JK,\infty}$.

Theorem 7 *Let $\mathbb{F}_{JK,\infty}$ be the fixed-G limit of $F_{JK}\left(\hat{\theta}\right)$. Define the effective degree of freedom*

$$G_{JK}^{*} = \frac{1}{\sum_{g=1}^{G}\frac{\lambda_g^2}{(1-\lambda_g)^2} + \left(\sum_{g=1}^{G}\frac{\lambda_g^2}{1-\lambda_g}\right)^2 - 2\sum_{g=1}^{G}\frac{\lambda_g^3}{(1-\lambda_g)^2}}.$$

Then, as $G\to\infty$,

$$P\left(p\mathbb{F}_{JK,\infty}\leq x\right) = P\left(\frac{pG_{JK}^{*}}{G_{JK}^{*}-p+1}\mathcal{F}_{p,G_{JK}^{*}-p+1}\leq x\right) + o\left(\frac{1}{G}\right).$$

Theorem 7 suggests an F test based on $F_{JK}\left(\hat{\theta}\right)$ that obtains critical values from the scaled F distribution $\frac{pG_{JK}^{*}}{G_{JK}^{*}-p+1}\mathcal{F}_{p,G_{JK}^{*}-p+1}$. In the literature, Bell and McCaffrey (2002) and Imbens and Kolesar (2016) propose t-tests using bias-corrected CCE. Cameron, Gelbach and Miller (2008, Table 3) find that the jackknife-type correction in $\hat{V}_{R,JK}$ performs a considerable finite sample

improvement for the standard CCE formula $\hat{V}_R$.Their procedures consist of two steps. The first is to employ a t statistic that is based on the Jackknife bias reduced variance estimator

$$t_{JK} = \frac{\sqrt{n}\left(R\hat{\theta} - r\right)}{\sqrt{\hat{V}_{R,JK}}},$$

and the second is to obtain critical values from a t distribution with an adjusted degree of freedom they propose. To illustrate the second step, let's consider a test based on $\hat{V}_{R,(3)}$. Let $P = X(X'X)^{-1}X'$ and $(I_n - P)_g$ be an $n \times n_g$ matrix that corresponds to the n_g columns of the $n \times n$ matrix $(I_n - P)$ for cluster g. With these notations, we define an $n \times G$ matrix E, where its g-th column is equal to the n-vector:

$$E_g = \sqrt{\frac{n - n_g}{n}}(I_n - P)_g(I_{n_g} - P_{gg})^{-1}X_g(X'X)^{-1}R'.$$

Bell and McCaffrey (2002) show that the distribution of t_{JK} can be well approximated by the t distribution with the degree of freedom adjustment given by

$$v^*_{\text{JK}} = \left\lfloor \frac{\left(\sum_{g=1}^{G} \eta_g\right)^2}{\sum_{g=1}^{G} \eta_g^2} \right\rfloor,$$

where η_g are the eigenvalues of the $G \times G$ matrix $E'E$. The

While this procedure produces more accurate inference results than standard normal approximation, conditions for the bias correction are very restrictive. According to Bell and McCaffrey (2002), $\hat{V}_{R,JK}$ requires $E[\epsilon_g\epsilon_g'|X] = \sigma^2 I_{n_g}$ and $X_g'X_g\left(X'X\right)^{-1}$ be constant over g to be unbiased, both of which are unrealistic under the cluster setting.

6 Simulation Evidence

In this section, we provide the finite sample performance of our asymptotic F tests and compare them with the existing cluster-robust tests. We focus on a simple linear regression with a low-order moving average spatial correlations. The model to be estimated is

$$y_s = x_s'\theta_0 + \epsilon_s \text{ with } E\left(x_s\epsilon_s\right) = 0,$$

where x_s is a $d_\theta \times 1$ vector of regressor including a constant, and corresponding true parameter vector $\theta_0 = (\alpha_0, \beta_0')'$ consists of intercept α_0 and regressor coefficients $\beta_0 = (\beta_{0,(1)}, \ldots \beta_{0,(d_\theta - 1)})$. Following BCH (2011) and Sun and Kim (2015), we characterize the cross-sectional dependence in the (un)observable data set $(x_s, \epsilon_s)'$ using a spatial moving average (MA) data generating process (DGP). The spatial units are indexed by a two-dimensional $K \times K$ lattice $\mathbb{D}$. For an observation with a location index $s \in \mathbb{D}$ and $x_s = (1, x_{s,(1)}, \ldots, x_{s,(d_\theta - 1)})'$ the DGP is

$$x_{s,(i)} = \sum_{\text{dist}(s,h) \leq D} \gamma^{\text{dist}(s,h)} v_{h,(i)}, \text{ where } v_{h,(i)} \overset{\text{i.i.d}}{\sim} N(0,1)$$

for $i = 1, \ldots, d_\theta - 1$, and

$$\epsilon_s = \sum_{\text{dist}(s,h) \leq D} \gamma^{\text{dist}(s,h)} u_h, \text{ where } u_h \overset{\text{i.i.d}}{\sim} N\left(0, \sum_{i=1}^{D} x_{h,(i)}\right)$$

with $\text{dist}(s,h) = \max\{s_1 - h_1, s_2 - h_2\}$. Setting $D = 4$ and $\gamma = 0.40$, we set a total sample size of $n = 1296$, where the sampling units are located in $K \times K$ lattice. We set the number of clusters $G \in \{4, 9, 16, 25\}$. The baseline model assumes equally sized clusters with $n_g \in \{216, 144, 81, 52\}$ with the total number of observations n ranges between 1296–1300. The sequence of unbalanced cluster-size designs starts by dividing G number clusters into a large-sized $G_1 \in \{1, 3, 6, 6\}$ and small-sized clusters $G_2 \in \{3, 6, 10, 19\}$. In each succeeding cluster-size design, we subtract $\{40, 25, 15, 9\}$ individuals from each of the second group of G_2 clusters and equally add them to the first group of G_1 clusters. In this manner, we construct a series of four cluster-size designs, in which the proportion of the equivalent degrees of freedoms G^* decreases monotonically up to 60% less than the original number of clusters (G). The design is similar to Carter et al. (2017) and Imbens and Kolesár (2016) which investigate the behavior of cluster-robust t statistic under cluster heterogeneity. Table 1 describes the unbalanced cluster-size Designs I-IV we consider. In Tables **??**–**??**, we also graphically illustrate the disposition of the sampling units and the boundaries of the geographic clusters for each scenario of the unbalanced cluster-sizes we consider.

6.1 Choice of tests

In our simulations, we consider a variety of Wald-type tests with Wald-statistic and asymptotic critical values. Let

$$F(\hat{\theta}) = \frac{n}{p}(R\hat{\theta} - r)' \left(\hat{V}_R\right)^{-1} (R\hat{\theta} - r)$$

be the cluster-robust F-statistics without any modification. The first two tests we consider use the original number of clusters G with the plain F-statistics $F(\hat{\theta})$, and with the degrees-of-freedom modified F statistic $[(G-p)/G]\, F(\hat{\theta})$. The associated critical values $\chi^{1-\alpha}_p/p$ and $\mathcal{F}^{1-\alpha}_{p,G-p}$ are justified under the large-G asymptotics and under the fixed-G asymptotics with equal-sized clusters (BCH, 2011), respectively. The next two tests are based on our new F-approximation theory in Sections (3)–(4). The first test uses the plain F statistics $F(\hat{\theta})$ with the associated critical value which is obtained from simulating the non-standard distribution $\mathbb{F}_\infty$ in Section (3). The second test uses our F-approximation theory in (4) employs the equivalent degrees of freedom G^* and the bias adjustment factor μ. The corresponding test statistics is $(\mu\,(G^* - p + 1)/G^*)F(\hat{\theta})$ with associated critical value $\mathcal{F}^{1-\alpha}_{p,G^*-p+1}$. We also check the performance of other types of F-tests discussed in Section 5. The first F-test is based on $(G_{\text{eff}} - p + 1)/G_{\text{eff}} F(\hat{\theta})$ with $\mathcal{F}^{1-\alpha}_{p,G_{\text{eff}}-p}$, where G_{eff} is the effective number of clusters proposed by Carter et al. (2017). The next two F-tests are based on the modified F-statistics whose denominator are replaced by the two bias adjusted estimators, $\hat{V}_{R,(2)}$ and $\hat{V}_{R,(3)}$, considered in Subsection **??.** The corresponding critical value is obtained by the non-standard fixed-G limiting distributions $\breve{\mathbb{F}}_\infty \overset{d}{=} Z_p' \bar{\mathbb{S}}^{-1}_{pp} Z_p/p$. In Table 2, we summarize the aforementioned tests considered in our simulations and corresponding names.

We also consider a cluster-robust version of wild bootstrap procedure proposed by Cameron et al. (2008) that uses the restricted OLS estimator and Rademacher random variable. With $p = 1$, the consistency and the higher-order refinement of the wild bootstrap procedure require the number of cluster G tends to infinity (Djogbenou et al, 2019). A recent study by Canay et al. (2019) also shows that, under certain homogeneity restrictions on the covariates and $p = 1$, the wild-bootstrap test using Rademacher random variable can be very close to be size-corrected under the fixed-G asymptotics.

Lastly, we also consider the Fama-Macbeth type procedure in Ibragimov and Müller (2010, 2016, IM hereafter) to test hypotheses in (31)–(**??**). With $p = 1$, The IM's t-test is formed by

a cluster-specific OLS estimator $\hat{\theta}_g = (X_g'X_g)^{-1}(X_g'y_g)$, which only use observations within the cluster g, and construct a t-statistic based on $\{R\hat{\theta}_g\}_{g=1}^G$, i.e.,

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

where $\bar{\theta}_G = \sum_{g=1}^G \hat{\theta}_g$. IM (2010, 2016) shows that using t_{IM} with $1-\alpha$ quantile of t_{G-1} yields asymptotically valid inference in the presence of cluster heterogeneity for certain ranges of α. When p is greater than one, there is no known asymptotic validity result for testing the joint hypothesis, e.g., BCH (2011, pp. 143). However, one can still apply the IM's method to test the joint hypotheses by constructing the F-statistics based on the clustered specific estimates $\{R\hat{\theta}_g\}_{g=1}^G$. To be more specific, the F-test generalization of IM (2010, 2016)'s t-test uses

$$F_{\text{IM}} := \frac{G}{p}\left(R\hat{\theta}_g - r\right)' \left[R\left(\sum_{g=1}^G (\hat{\theta}_g - \bar{\theta}_G)(\hat{\theta}_g - \bar{\theta}_G)'\right)R'/(G-p)\right]^{-1}\left(R\hat{\theta}_g - r\right),$$

with associated critical value $\mathcal{F}_{p,G-p}^{1-\alpha}$.

6.1.1 Size experiment results

Setting $d_\theta = 4$, we examine the empirical size of a variety of testing procedures. The null hypotheses of interests are

$$H_{01} : \beta_{0,(1)} = 0, \tag{31}$$

$$H_{02} : \beta_{0,(1)} = \beta_{0,(2)} = 0, \tag{32}$$

where corresponding testing matrix is $R = (0,1,0,0)$ with $r = 0$, and $R = (0,1,1,0)$ with $r = (0,0)'$. All of our of simulation results are based on $20,000$ times of Monte Carlo repetitions, and the number of wild-bootstrap replication is $B = 2^G$ for $G = 4$, and $B = 499$ for $G = 9, 16$, and 25.

Tables 3–5 report the empirical size of the cluster-robust tests we consider in the presence of unbalanced cluster sizes. The results first indicate that the conventional large-G asymptotic Wald test (`Chi-square`) based on chi-square critical value suffer from severe size distortions, especially when the number of clusters G is small, p is large. For example, with $G = 9$ and $p = 3$, the empirical sizes of the chi-squared Wald test are 31.2% when there is moderate heterogeneity in cluster sizes at Design I. The size distortion becomes more severe with 61.2% when the heterogeneity in cluster size increases at Design IV. With the same values of G and p, the Wald tests' empirical sizes reduce to 6.9%–29.2% when the F critical values with the original number of clusters are employed (`BCH`). This result is consistent with the findings in BCH (2011) and Hansen (2007), which highlight the improved finite sample performance of the fixed-G approximation. When the size heterogeneity increases, however, Tables 3–5 indicate that there are still considerable amount of size distortions using the original number of clusters in F-tests. For example, with $G = 9$ and 16, the substantial size heterogeneity in Design IV increases the size distortions of the test BCH up to 20.8% and 14.7%, respectively, for $p = 2$. This motivates us to implement our non-standard fixed-G critical value $\mathbb{F}_\infty^{1-\alpha}$ (`HK-simul`), or use the alternative F-approximation critical value $\mathcal{F}_{p,G^*-p+1}^{1-\alpha}$ together with the equivalent degrees of freedom G^* and the bias-adjustment

factor μ (`HK-F-approx`). Tables 3–5 shows that the two tests, `HK-simul and HK-F-approx` value significantly alleviate the previous size distortions. The size distortions of HK-simul in the previous example are reported to be 10.4% and 8.4%, which have been significantly improved from BCH. Also, the size distortions are reduced by 9.5% and 8.3% in `HK-F-approx.` Lastly, we find evidence that the bias-adjusted F statistics, which are equipped with the non-standard critical values, result in similar finite-sample improvements for the empirical sizes (`BM-HC2-simul` and `BM-HC3-simul`). For example, the size distortions of `BM-HC3-simul` in the same example are improved by 8.8% and 7.2%, which are close to the targeted level 5%. This indicates our fixed-G asymptotic inferences with the bias-adjusted estimators successfully captures finite-sample distributions of corresponding Wald statistics and yields more accurate testing in finite-sample. We find similar conclusions for other values of G and p, and cluster-size designs.

Tables 3–5 also shows the empirical rejection probabilities of the cluster-robust wild bootstrap procedure by Cameron et al. (1996), which is denoted `Wild-boot` in the tables. We find that the empirical rejection probabilities of the wild bootstrap procedure become close to the nominal size when testing a single null hypothesis with $p = 1$, but the improved performances no longer exists when we test joint hypotheses with $p = 2$–3.

Tables 3–5 shows that the finite sample size distortions of the sub-sample based t and F tests, `Sub-t/F`, are also substantially reduced and close to the nominal size of 5%.

6.1.2 Power experiment results

We investigate the finite-sample power performances of three cluster-robust test statistics: $F(\hat{\theta})$, $F_{\mathrm{JK}}(\hat{\theta})$, and F_{IM}. Note that the first two test statistics is constructed by $\hat{\theta}$ that uses the entire observations, whereas the IM F-statistics F_{IM} uses estimators that use cluster specific observations. We use the finite-sample critical values under the null for each test statistic, 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 $R\theta_0 = R\theta_0 + c/\sqrt{n}$ for $||c|| \in [0, 15]$ and $p \in \{1, 2\}$.

Table 1: Design of unbalanced cluster size

$G = 4$				
	Equal-sized			
·	n_1	$n_2 = ... = n_4$	n	G^*
Baseline	324	324	1296	4
	Unbalanced cluster sizes			
Design	n_1	$n_2 = ... = n_4$	n	G^*
I	444	284	1296	3
II	564	244	1296	3
III	684	204	1296	3
IV	804	164	1296	3
$G = 9$				
	Equal-sized			
·	$n_1 = \ldots = n_3$	$n_4 = ... = n_9$	n	G^*
Baseline	144	144	1296	9
	Unbalanced cluster sizes			
Design	$n_1 = \ldots = n_3$	$n_4 = ... = n_9$	n	G^*
I	194	119	1296	8
II	244	94	1296	7
III	294	69	1296	6
IV	344	44	1296	4
$G = 16$				
	Equal-sized			
·	$n_1 = \ldots = n_6$	$n_7 = ... = n_{16}$	n	G^*
Baseline	81	81	1296	16
	Unbalanced cluster sizes			
·	$n_1 = \ldots = n_6$	$n_7 = ... = n_{16}$	n	G^*
Design I	106	66	1296	15
Design II	131	51	1296	13
Design III	156	36	1296	10
Design IV	181	21	1296	8
$G = 25$				
	Equal-sized			
·	$n_1 = \ldots = n_6$	$n_{10} = ... = n_{25}$	n	G^*
Baseline	52	52	1300	25
	Unbalanced cluster sizes			
·	$n_1 = ... = n_6$	$n_7 = ... = n_{25}$	n	G^*
Design I	80	43	1297	23
Design II	109	34	1300	18
Design III	137	25	1300	14
Design IV	166	16	1300	10

Table 2: Summary of the first-step, two-step tests, and J test

Name	Test Statistics	Critical Value
Chi-square	$F(\hat{\theta})$	$\chi_p^{1-\alpha}$
BCH	$[(G-p)/G]\,F(\hat{\theta})$	$\mathcal{F}_{p,G-p}^{1-\alpha}$
HK-F-simul.	$F(\hat{\theta})$	$\mathbb{F}_{\infty}^{1-\alpha}$
HK-F-approx.	$(\mu\,(G^*-p+1)/G^*)\cdot F(\hat{\theta})$	$\mathcal{F}_{p,G^*-p+1}^{1-\alpha}$
BM-HC3-simul.	$F_{\text{BM-HC3}}(\hat{\theta})$	$\breve{\mathbb{F}}_{\infty}^{1-\alpha}$
BM-HC3-F-approx.	$((G_{JK}^*-p+1)/G_{JK}^*)\cdot F_{\text{BM-HC3}}(\hat{\theta})$	$\mathcal{F}_{p,G_{JK}^*-p+1}^{1-\alpha}$
CSS	$F(\hat{\theta})$	$\mathcal{F}_{p,G_{\text{eff}}}^{1-\alpha}$
BM-IK	$F_{\text{BM-HC2}}(\hat{\theta})$	$\mathcal{F}_{p,v_{\text{BM-IK}}^*}^{1-\alpha}$
BM-HC3	$F_{\text{BM-HC3}}(\hat{\theta})$	$\mathcal{F}_{p,v_{\text{JK}}^*}^{1-\alpha}$
Wild-boot	$F(\hat{\theta})$	Cluster-robust wild bootstrap
Sub-t/F	t_{IM} and F_{IM}	$t_{G-1}^{1-\alpha}$ or $\mathcal{F}_{p,G-p}^{1-\alpha}$

Table 3: Empirical size of cluster-robust tests when there is a heterogeneity in cluster size: $G=4$

Unbalanced sizes in clusters with G=4								
	Design I		**Design II**		**Design III**		**Design IV**	
Tests	$p=1$	$p=2$	$p=1$	$p=2$	$p=1$	$p=2$	$p=1$	$p=2$
Chi-square	0.199	0.423	0.225	0.464	0.266	0.523	0.338	0.620
BCH	0.052	0.052	0.063	0.060	0.080	0.080	0.119	0.111
HK-F-simul.	0.044	0.046	0.051	0.049	0.057	0.058	0.074	0.068
HK-F-approx.	0.051	0.100	0.059	0.111	0.067	0.129	0.090	0.160
BM-HC3-simul.	0.043	0.040	0.049	0.043	0.054	0.049	0.068	0.059
BM-HC3-F-approx.	0.049	0.093	0.058	0.109	0.069	0.134	0.096	0.145
CSS	0.062	0.160	0.074	0.183	0.093	0.217	0.130	0.280
BM-IK	0.050	0.142	0.060	0.164	0.068	0.190	0.057	0.169
BM-HC3	0.049	0.138	0.058	0.158	0.063	0.174	0.062	0.172
Wild-boot	0.077	0.209	0.079	0.203	0.079	0.194	0.082	0.191
Sub-t/F	0.047	0.048	0.049	0.048	0.049	0.049	0.051	0.048

For test statistics and critical value in each test, see Table 2.

Table 4: Empirical size of cluster-robust tests when there is a heterogeneity in cluster size: $G = 9$

Unbalanced sizes in clusters with G=9								
	Design I		**Design II**		**Design III**		**Design IV**	
Tests	$p = 1$	$p = 2$	$p = 1$	$p = 2$	$p = 1$	$p = 2$	$p = 1$	$p = 2$
Chi-square	0.122	0.206	0.135	0.233	0.168	0.303	0.206	0.399
BCH	0.065	0.069	0.072	0.085	0.099	0.128	0.134	0.213
HK-F-simul.	0.057	0.059	0.058	0.063	0.070	0.079	0.081	0.110
HK-F-approx.	0.064	0.068	0.063	0.066	0.076	0.078	0.089	0.101
BM-HC3-simul.	0.053	0.054	0.053	0.057	0.062	0.072	0.068	0.094
BM-HC3-F-approx.	0.059	0.063	0.059	0.063	0.061	0.053	0.066	0.056
CSS	0.046	0.060	0.045	0.064	0.056	0.083	0.069	0.124
BM-IK	0.061	0.092	0.060	0.095	0.067	0.113	0.071	0.143
BM-HC3	0.058	0.088	0.057	0.090	0.064	0.106	0.067	0.134
Wild-boot	0.059	0.134	0.055	0.128	0.060	0.131	0.058	0.128
Sub-t/F	0.055	0.053	0.053	0.053	0.053	0.052	0.054	0.051

For test statistics and critical value in each test, see Table 2.

Table 5: Empirical size of cluster-robust tests when there is a heterogeneity in cluster size: $G = 16$

Unbalanced sizes in clusters with G=16								
	Design I		**Design II**		**Design III**		**Design IV**	
Tests	$p = 1$	$p = 2$	$p = 1$	$p = 2$	$p = 1$	$p = 2$	$p = 1$	$p = 2$
Chi-square	0.098	0.139	0.104	0.154	0.112	0.183	0.129	0.221
BCH	0.067	0.071	0.069	0.080	0.079	0.105	0.095	0.139
HK-F-simul.	0.060	0.062	0.057	0.064	0.060	0.072	0.064	0.081
HK-F-approx.	0.067	0.071	0.064	0.071	0.065	0.073	0.068	0.080
BM-HC3-simul.	0.056	0.056	0.052	0.057	0.053	0.063	0.057	0.071
BM-HC3-F-approx.	0.063	0.065	0.059	0.065	0.058	0.066	0.062	0.073
CSS	0.047	0.051	0.041	0.049	0.038	0.049	0.043	0.060
BM-IK	0.063	0.077	0.059	0.080	0.060	0.090	0.063	0.102
BM-HC3	0.061	0.074	0.057	0.076	0.056	0.084	0.060	0.095
Wild-boot	0.063	0.103	0.058	0.107	0.056	0.118	0.059	0.128
Sub-t/F	0.059	0.058	0.056	0.056	0.054	0.054	0.051	0.050

For test statistics and critical value in each test, see Table 2.

Table 6: Empirical size of cluster-robust tests when there is a heterogeneity in cluster size: $G = 25$

Unbalanced sizes in clusters with G=25								
	Design I		**Design II**		**Design III**		**Design IV**	
Tests	$p = 1$	$p = 2$	$p = 1$	$p = 2$	$p = 1$	$p = 2$	$p = 1$	$p = 2$
Chi-square	0.086	0.125	0.101	0.145	0.115	0.177	0.133	0.232
BCH	0.068	0.080	0.080	0.097	0.093	0.124	0.111	0.175
HK-F-simul.	0.060	0.069	0.066	0.078	0.068	0.083	0.074	0.098
HK-F-approx.	0.067	0.077	0.072	0.083	0.075	0.084	0.077	0.093
BM-HC3-simul.	0.056	0.063	0.059	0.070	0.058	0.071	0.061	0.080
BM-HC3-F-approx.	0.061	0.069	0.064	0.073	0.064	0.074	0.066	0.080
CSS	0.049	0.055	0.044	0.049	0.040	0.046	0.044	0.060
BM-IK	0.063	0.080	0.066	0.086	0.066	0.092	0.069	0.111
BM-HC3	0.060	0.076	0.063	0.081	0.061	0.086	0.063	0.101
Wild-boot	0.064	0.100	0.065	0.107	0.062	0.118	0.060	0.136
Sub-t/F	0.061	0.063	0.059	0.062	0.055	0.059	0.053	0.050

For test statistics and critical value in each test, see Table 2.

Appendix: Proofs

Proof of Lemma 1. (i) Let $Z_{m,g} \overset{\text{i.i.d}}{\sim} N(0, I_m)$. Under Assumptions 1–3, standard asymptotic arguments of GMM estimator give us

$$\sqrt{n}(\hat{\theta} - \theta_0) = R(\Gamma' W^{-1}\Gamma)^{-1}\Gamma' W^{-1} g_n(\theta_0) + o_p(1),$$

and so

$$\begin{aligned}\sqrt{n}(R\hat{\theta} - r) &= R(\Gamma' W^{-1}\Gamma)^{-1}\Gamma' W^{-1} g_n(\theta_0) + o_p(1) \\ &= \tilde{R}\left(\sum_{g=1}^{G}\frac{1}{\sqrt{n}}\sum_{i=1}^{n_g} f_i^g(\theta_0)\right)(1 + o_p(1)) \\ &\xrightarrow{d} \tilde{R}\Lambda\left(\sum_{g=1}^{G}\sqrt{\lambda_g} Z_{m,g}\right) \overset{d}{=} \tilde{\Lambda}\left(\sum_{g=1}^{G}\sqrt{\lambda_g} Z_{p,g}\right), \end{aligned} \tag{33}$$

where $\tilde{\Lambda}\tilde{\Lambda}' = \tilde{R}\Omega\tilde{R}'$. The weak convergence result in (33) follows from (9), Assumption 1, and continuous mapping theorem. This gives us the first result in (11). To show $\hat{R}\hat{\Omega}\hat{R} \xrightarrow{d} \tilde{\Lambda}\bar{\mathbb{S}}_{pp}\tilde{\Lambda}'$, we use mean-value theorem, for each $g = 1, ..., G$, and obtain that

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

where $\tilde{\theta}^*$ is between $\hat{\theta}$ and θ_0. Here, $\tilde{\theta}^*$ is different for different rows of $\partial f_i^g(\tilde{\theta}^*)/\partial\theta'$. For notational simplicity, we do not make this explicit. Using this result, and Assumptions 2 and 4, we have that

$$\begin{aligned}\frac{1}{\sqrt{n_g}}\sum_{i=1}^{n_g} f_i^g(\hat{\theta}) &= \left\{\frac{1}{\sqrt{n_g}}\sum_{i=1}^{n_g} f_i^g(\theta_0) - \sqrt{\frac{n_g}{n}}\left(\frac{1}{n_g}\sum_{i=1}^{n_g}\frac{\partial f_i^g(\tilde{\theta}^*)}{\partial\theta'}\right)\right. \\ &\quad \left.\times(\Gamma' W^{-1}\Gamma)^{-1}\Gamma' W^{-1}\left(\sum_{\tilde{g}=1}^{G}\frac{1}{\sqrt{n}}\sum_{i=1}^{n_{\tilde{g}}} f_i^{\tilde{g}}(\theta_0)\right)\right\}(1 + o_p(1)) \\ &= \left\{\frac{1}{\sqrt{n_g}}\sum_{i=1}^{n_g} f_i^g(\theta_0) - \sqrt{\lambda_g}\Gamma(\Gamma' W^{-1}\Gamma)^{-1}\Gamma' W^{-1}\right. \\ &\quad \left.\times\sum_{\tilde{g}=1}^{G}\sqrt{\lambda_{\tilde{g}}}\left(\frac{1}{\sqrt{n_{\tilde{g}}}}\sum_{i=1}^{n_{\tilde{g}}} f_i^{\tilde{g}}(\theta_0)\right)\right\}(1 + o_p(1)).\end{aligned}$$

Using the CLT assumption in (9) and Assumption 5, we then have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n_g} f_i^g(\hat{\theta}) \xrightarrow{d} \Lambda\sqrt{\lambda_g} Z_{m,g} - \lambda_g\Gamma(\Gamma' W^{-1}\Gamma)^{-1}\Gamma' W^{-1}\Lambda\sum_{\tilde{g}=1}^{G}\sqrt{\lambda_{\tilde{g}}} Z_{m,\tilde{g}}.$$

It follows that

$$\hat{\Gamma}(\hat{\theta})' W_n^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n_g} f_i^g(\hat{\theta})\right) \xrightarrow{d} \Gamma' W^{-1}\Lambda\left[\sqrt{\lambda_g} Z_{m,g} - \lambda_g\sum_{\tilde{g}=1}^{G}\sqrt{\lambda_{\tilde{g}}} Z_{m,\tilde{g}}\right].$$

Thus, the scaled CCE matrix $\hat{R}\hat{\Omega}\hat{R}$ converges in distribution to a random matrix:

$$\begin{aligned}
\hat{R}\hat{\Omega}\hat{R} &= \hat{R}\sum_{g=1}^{G}\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n_g}f_i^g(\hat{\theta})\right)\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n_g}f_i^g(\hat{\theta})\right)'\right]\hat{R}' \\
&\xrightarrow{d} \tilde{R}\Lambda\left\{\sum_{g=1}^{G}\left(\sqrt{\lambda_g}Z_{m,g}-\lambda_g\sum_{\tilde{g}=1}^{G}\sqrt{\lambda_{\tilde{g}}}Z_{m,\tilde{g}}\right)\right. \\
&\left.\times\left(\sqrt{\lambda_g}Z_{m,g}-\lambda_g\sum_{\tilde{g}=1}^{G}\sqrt{\lambda_{\tilde{g}}}Z_{m,\tilde{g}}\right)'\right\}\Lambda'\tilde{R}' \\
&\overset{d}{=} \tilde{\Lambda}\bar{\mathbb{S}}_{pp}\tilde{\Lambda}',
\end{aligned} \tag{34}$$

which proves the weak convergence of CCE result in (11). Also, it is not difficult to check that the convergence results in (33) and (34) hold jointly.

For any $g = 1, \ldots, G$, it is easy to check

$$\begin{aligned}
&cov\left(\sum_{\tilde{g}=1}^{G}\sqrt{\lambda_{\tilde{g}}}Z_{p,\tilde{g}}, \sqrt{\lambda_g}Z_{p,g}-\lambda_g\sum_{\tilde{g}=1}^{G}\sqrt{\lambda_{\tilde{g}}}Z_{p,\tilde{g}}\right) \\
&= cov\left(\sum_{\tilde{g}=1}^{G}\sqrt{\lambda_{\tilde{g}}}Z_{p,\tilde{g}}, \sqrt{\lambda_g}Z_{p,g}\right)-\lambda_g Var\left(\sum_{\tilde{g}=1}^{G}\sqrt{\lambda_{\tilde{g}}}Z_{p,\tilde{g}}\right) \\
&= \lambda_g Var(Z_{p,g})-\lambda_g Var\left(\sum_{\tilde{g}=1}^{G}\sqrt{\lambda_{\tilde{g}}}Z_{p,\tilde{g}}\right) \\
&= \lambda_g I_p-\lambda_g I_p \\
&= 0,
\end{aligned}$$

and this shows the independence result in (**??**) for all values of $\lambda_g \geq 0$ such that $\sum_{g=1}^{G}\lambda_g = 1$.

(ii) To show the distributional equivalence in (13), denote $p\times 1$ random vector $A_g = \sqrt{\lambda_g}Z_{p,g} - \lambda_g\sum_{g=1}^{G}\sqrt{\lambda_g}Z_{p,g}$. Also, define $p\times G$ random matrices $\mathbf{A}' = (A_1, A_2, \ldots, A_G)$ and $\mathbf{Z}' = (Z_{p,1}, \ldots, Z_{p,G})$. Then it is easy to check that $\bar{\mathbb{S}}_{pp} = \mathbf{A}'\mathbf{A}$ and

$$\begin{aligned}
\mathbf{A}' &= (Z_{p,1}, \ldots, Z_{p,G})\times \mathrm{Diag}(\sqrt{\lambda})\times(I_G - 11'\mathrm{diag}(\lambda)) \\
&= \mathbf{Z}'D.
\end{aligned}$$

Thus, we can write

$$\bar{\mathbb{S}}_{pp} = \mathbf{Z}'DD'\mathbf{Z} = \mathbf{Z}'\Phi\mathrm{Diag}(\kappa)\Phi'\mathbf{Z},$$

where $\Phi\mathrm{Diag}(\kappa)\Phi'$ is a spectral decomposition of $G\times G$ matrix DD'. By construction, $\mathrm{Diag}(\kappa)$ is a diagonal matrix of eigenvalues of DD', and $\Phi = (\phi_1, ..., \phi_G)$ is a matrix of associated eigenvectors which is orthonormal, i.e. $\Phi'\Phi = I_G$. Let a be arbitrary $p\times 1$ non-zero vector such that $a'a = 1$. Then, it is not difficult to check

$$\mathbf{Z}a = \begin{pmatrix} Z_{p,1}'a \\ Z_{p,2}'a \\ \vdots \\ Z_{p,G}'a \end{pmatrix} \overset{d}{=} N(0, I_G).$$

From the orthonormal property of the eigenvector matrix Φ, we have

$$\Phi'\mathbf{Z}a \overset{d}{=} N(0, I_G),$$

which implies that $\Phi'\mathbf{Z}a \overset{d}{=} \breve{\mathbf{Z}}a$, where $\breve{\mathbf{Z}}' = (\tilde{Z}_{p,1}, \ldots, \tilde{Z}_{p,G})$ is independent copy of $\mathbf{Z}$. Therefore, the following distributional equivalence holds

$$a'\bar{\mathbb{S}}_{pp}a \overset{d}{=} a'\breve{\mathbf{Z}}'\text{Diag}(\kappa)\breve{\mathbf{Z}}a = a'\left(\sum_{g=1}^{G}\kappa_g\tilde{Z}_{p,g}\tilde{Z}'_{p,g}\right)a.$$

Since the non-random vector a is arbitrarily chosen, we obtain the desired result in (13). ■

Before we prove Theorem 4, we introduce some technical results in the following Lemmas.

Lemma 8 *Let $\{\kappa_g\}_{g=1}^G$ be G eigenvalues of DD', where $D = (I_G - Diag(\lambda)11')Diag(\sqrt{\lambda})$. Then, the following results hold.*

(a) $\sum_{i=1}^G \kappa_g = 1 - \sum_{i=1}^G \lambda_g^2$

(b) $\sum_{i=1}^G \kappa_g^2 = \sum_{g=1}^G \lambda_g^2 + \left(\sum_{g=1}^G \lambda_g^2\right)^2 - 2\sum_{g=1}^G \lambda_g^3$

(c) $\sum_{i=1}^G \kappa_g^j = O\left(G^{-(j-1)}\right)$

(d) $\sum_{g=1}^G \kappa_g^2 / \sum_{g=1}^G \kappa_g = \sum_{g=1}^G \lambda_g^2 + o(G^{-1})$

(e) $\left(\sum_{i=1}^G \kappa_g\right)^2 / \sum_{i=1}^G \kappa_g^2 = \left(\sum_{g=1}^G \lambda_g^2\right)^{-1}(1 + o(1))$

(f) $\frac{\mu(G^*-p+1)}{G^*} = 1 - p\sum_{g=1}^G \lambda_g^2 + o\left(\frac{1}{G}\right)$

(g) $\frac{G^*}{\mu(G^*-p+1)} = 1 + p\sum_{g=1}^G \lambda_g^2 + o(G^{-1})$.

Proof of Lemma 8. (a) It is easy to calculate

$$\begin{aligned} DD' &= \text{diag}(\sqrt{\lambda})(I_G - 11'\text{diag}(\lambda))(I_G - \text{diag}(\lambda)11')\text{diag}(\sqrt{\lambda}) \\ &= \begin{pmatrix} \lambda_1\sum_{g=1}^G\lambda_g^2 & \sqrt{\lambda_1\lambda_2}\sum_{g=1}^G\lambda_g^2 & \cdots & \sqrt{\lambda_1\lambda_G}\sum_{g=1}^G\lambda_g^2 \\ +\lambda_1 - 2\lambda_1^2 & -\sqrt{\lambda_1\lambda_2}(\lambda_1+\lambda_2) & & -\sqrt{\lambda_1\lambda_G}(\lambda_1+\lambda_G) \\ \sqrt{\lambda_1\lambda_2}\sum_{g=1}^G\lambda_g^2 & \lambda_2\sum_{g=1}^G\lambda_g^2 & \cdots & \sqrt{\lambda_2\lambda_G}\sum_{g=1}^G\lambda_g^2 \\ -\sqrt{\lambda_1\lambda_2}(\lambda_1+\lambda_2) & +\lambda_2 - 2\lambda_2^2 & & -\sqrt{\lambda_2\lambda_G}(\lambda_2+\lambda_G) \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{\lambda_1\lambda_G}\sum_{g=1}^G\lambda_g^2 & \sqrt{\lambda_2\lambda_G}\sum_{g=1}^G\lambda_g^2 & \cdots & \lambda_G\sum_{g=1}^G\lambda_g^2 \\ -\sqrt{\lambda_1\lambda_G}(\lambda_1+\lambda_G) & -\sqrt{\lambda_2\lambda_G}(\lambda_2+\lambda_G) & & +\lambda_G - 2\lambda_G^2 \end{pmatrix}. \end{aligned}$$

Thus,

$$\begin{aligned} \sum_{i=1}^G \kappa_g = tr\left(DD'\right) &= \sum_{i=1}^G \lambda_g + \sum_{i=1}^G \lambda_g \sum_{g=1}^G \lambda_g^2 - 2\sum_{g=1}^G \lambda_g^2 \\ &= 1 - \sum_{i=1}^G \lambda_g^2. \end{aligned}$$

(b) We have

$$\begin{aligned}
\sum_{i=1}^{G}\kappa_g^2 &= tr\left(\left(DD'\right)^2\right)\\
&= tr\left(\left(\begin{array}{cccc}
\lambda_1\sum_{g=1}^{G}\lambda_g^2 & \sqrt{\lambda_1\lambda_2}\sum_{g=1}^{G}\lambda_g^2 & \cdots & \sqrt{\lambda_1\lambda_G}\sum_{g=1}^{G}\lambda_g^2\\
+\lambda_1-2\lambda_1^2 & -\sqrt{\lambda_1\lambda_2}\left(\lambda_1+\lambda_2\right) & & -\sqrt{\lambda_1\lambda_G}\left(\lambda_1+\lambda_G\right)\\
\sqrt{\lambda_1\lambda_2}\sum_{g=1}^{G}\lambda_g^2 & \lambda_2-2\lambda_2^2 & \cdots & \sqrt{\lambda_2\lambda_G}\sum_{g=1}^{G}\lambda_g^2\\
-\sqrt{\lambda_1\lambda_2}\left(\lambda_1+\lambda_2\right) & +\lambda_2\sum_{g=1}^{G}\lambda_g^2 & & -\sqrt{\lambda_2\lambda_G}\left(\lambda_2+\lambda_G\right)\\
\vdots & \vdots & \ddots & \vdots\\
\sqrt{\lambda_1\lambda_G}\sum_{g=1}^{G}\lambda_g^2 & \sqrt{\lambda_2\lambda_G}\sum_{g=1}^{G}\lambda_g^2 & \cdots & \lambda_G-2\lambda_G^2\\
-\sqrt{\lambda_1\lambda_G}\left(\lambda_1+\lambda_G\right) & -\sqrt{\lambda_2\lambda_G}\left(\lambda_2+\lambda_G\right) & & +\lambda_G\sum_{g=1}^{G}\lambda_g^2
\end{array}\right)^2\right)\\
&= \sum_{g=1}^{G}\lambda_g^2+\left(\sum_{g=1}^{G}\lambda_g^2\right)^2-2\sum_{g=1}^{G}\lambda_g^3\\
&= \sum_{g=1}^{G}\lambda_g^2+o\left(\frac{1}{G}\right)
\end{aligned}$$

(c) Using the fact that $\lambda_g=\lim_{n\to\infty}n_g/n=O\left(1/G\right)$ as G grows, we have

$$\begin{aligned}
DD' &= \text{diag}(\sqrt{\lambda})(I_G-11'\text{diag}(\lambda))(I_G-\text{diag}(\lambda)11')\text{diag}(\sqrt{\lambda})\\
&= \left(\begin{array}{ccc}\sqrt{\lambda_1} & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & \sqrt{\lambda_G}\end{array}\right)\left[I-\left(\begin{array}{ccc}\lambda_1 & \cdots & \lambda_G\\ \vdots & \vdots & \vdots\\ \lambda_1 & \cdots & \lambda_G\end{array}\right)\right]\\
&\times\left[I-\left(\begin{array}{ccc}\lambda_1 & \cdots & \lambda_1\\ \vdots & \vdots & \vdots\\ \lambda_G & \cdots & \lambda_G\end{array}\right)\right]\left(\begin{array}{ccc}\sqrt{\lambda_1} & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & \sqrt{\lambda_G}\end{array}\right)\\
&= O\left(\frac{1}{G}\right)\left(\begin{array}{ccc}1 & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & 1\end{array}\right)+O\left(\frac{1}{G^2}\right)\left(\begin{array}{ccc}1 & \cdots & 1\\ \vdots & \vdots & \vdots\\ 1 & \cdots & 1\end{array}\right). \qquad (35)
\end{aligned}$$

Based on (35), it is easy to show that

$$\left(DD'\right)^j=O\left(\frac{1}{G^j}\right)\left(\begin{array}{ccc}1 & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & 1\end{array}\right)+O\left(\frac{1}{G^{j+1}}\right)\left(\begin{array}{ccc}1 & \cdots & 1\\ \vdots & \vdots & \vdots\\ 1 & \cdots & 1\end{array}\right)$$

Thus, we have

$$\sum_{i=1}^{G}\kappa_g^j=tr\left\{\left(DD'\right)^j\right\}=O\left(\frac{1}{G^{j-1}}\right).$$

(d) Note that

$$\begin{aligned}\frac{\sum_{g=1}^{G}\kappa_g^2}{\sum_{g=1}^{G}\kappa_g} &= \frac{\sum_{g=1}^{G}\kappa_g^2}{1-\sum_{i=1}^{G}\lambda_g^2} \\ &= \sum_{g=1}^{G}\kappa_g^2\left(1+\frac{\sum_{i=1}^{G}\lambda_g^2}{1-\sum_{i=1}^{G}\lambda_g^2}\right) \\ &= \left(\sum_{g=1}^{G}\lambda_g^2+o\left(\frac{1}{G}\right)\right)\left(1+O\left(\frac{1}{G}\right)\right) \\ &= \sum_{g=1}^{G}\lambda_g^2+o\left(\frac{1}{G}\right).\end{aligned}$$

(e) Note that

$$\begin{aligned}\frac{\left(\sum_{g=1}^{G}\kappa_g\right)^2}{\sum_{g=1}^{G}\kappa_g^2} &= \frac{1-2\sum_{i=1}^{G}\lambda_g^2+\left(\sum_{i=1}^{G}\lambda_g^2\right)^2}{\sum_{g=1}^{G}\lambda_g^2+o\left(G^{-1}\right)} \\ &= \frac{1/\sum_{g=1}^{G}\lambda_g^2-2+o\left(1\right)}{1+o\left(1\right)} \\ &= \frac{1}{\sum_{g=1}^{G}\lambda_g^2}\left(1+o\left(1\right)\right).\end{aligned}$$

(f) Note that

$$\begin{aligned}\frac{\mu(G^*-p+1)}{G^*} &= \frac{\left(\sum_{g=1}^{G}\kappa_g\right)\left[\left(\sum_{g=1}^{G}\kappa_g\right)^2/\sum_{g=1}^{G}\kappa_g^2-p+1\right]}{\left(\sum_{g=1}^{G}\kappa_g\right)^2/\sum_{g=1}^{G}\kappa_g^2} \\ &= \frac{\left(\sum_{g=1}^{G}\kappa_g\right)^2-(p-1)\sum_{g=1}^{G}\kappa_g^2}{\sum_{g=1}^{G}\kappa_g} \\ &= \sum_{g=1}^{G}\kappa_g-(p-1)\frac{\sum_{g=1}^{G}\kappa_g^2}{\sum_{g=1}^{G}\kappa_g} \\ &= \left(1-\sum_{g=1}^{G}\lambda_g^2\right)-(p-1)\left(\sum_{g=1}^{G}\lambda_g^2+o\left(\frac{1}{G}\right)\right) \\ &= 1-p\sum_{g=1}^{G}\lambda_g^2+o\left(\frac{1}{G}\right).\end{aligned}$$

(g)

$$\begin{aligned}\frac{G^*}{\mu(G^*-p+1)} &= \frac{1}{1-p\sum_{g=1}^{G}\lambda_g^2+o\left(\frac{1}{G}\right)} \\ &= 1+\frac{p\sum_{g=1}^{G}\lambda_g^2}{1-p\sum_{g=1}^{G}\lambda_g^2+o\left(\frac{1}{G}\right)}+o\left(\frac{1}{G}\right) \\ &= 1+p\sum_{g=1}^{G}\lambda_g^2\left(1+\frac{p\sum_{g=1}^{G}\lambda_g^2}{1-p\sum_{g=1}^{G}\lambda_g^2+o\left(\frac{1}{G}\right)}\right)+o\left(\frac{1}{G}\right) \\ &= 1+p\sum_{g=1}^{G}\lambda_g^2+o\left(\frac{1}{G}\right)\end{aligned}$$

■

Lemma 9 *As $G\to\infty$, we have*

(a) $E[s_{11}-s_{12}s_{22}^{-1}s_{21}]=(1-\sum_{g=1}^{G}\lambda_g^2)-\left(\sum_{g=1}^{G}\lambda_g^2\right)(p-1)+o(G^{-1})$;

(b) $E[\left(s_{11}-s_{12}s_{22}^{-1}s_{21}\right)^2]=1-2\left(\sum_{g=1}^{G}\lambda_g^2\right)(p-1)+o(G^{-1})$;

(c) $E[\left(s_{11}-s_{12}s_{22}^{-1}s_{21}\right)-1]^2=2\sum_{g=1}^{G}\lambda_g^2+o(G^{-1})$.

Proof of Lemma 9. (a) For $\tilde{Z}_{p,g}=(\tilde{Z}_{1,g},\tilde{Z}'_{(p-1),g})'\overset{\text{i.i.d}}{\sim}N(0,I_p)$ in which $\tilde{Z}_{1,g}$ is a scalar and $\tilde{Z}_{(p-1),g}$ is a $(p-1)\times 1$ vector, recall that

$$\begin{aligned}\sum_{g=1}^{G}\kappa_g\tilde{Z}_{p,g}\tilde{Z}'_{p,g} &= \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22}\end{pmatrix} \\ &= \begin{pmatrix} \sum_{g=1}^{G}\kappa_g\tilde{Z}_{1,g}^2 & \sum_{g=1}^{G}\kappa_g\tilde{Z}_{1,g}\tilde{Z}'_{(p-1),g} \\ \sum_{g=1}^{G}\kappa_g\tilde{Z}_{(p-1),g}\tilde{Z}_{1,g} & \sum_{g=1}^{G}\kappa_g\tilde{Z}_{(p-1),g}\tilde{Z}'_{(p-1),g}\end{pmatrix}.\end{aligned}$$

By Lemma 8-(a) and $E[\tilde{Z}_{1,g}^2]=1$, it is easy to check that

$$E\left[\nu_{11}\right]=\sum_{g=1}^{G}\kappa_g E[\tilde{Z}_{1,g}^2]=\sum_{g=1}^{G}\kappa_g=1-\sum_{g=1}^{G}\lambda_g^2. \tag{36}$$

Also, we can write

$$\begin{aligned}E[\nu_{12}\nu_{22}^{-1}\nu_{21}] &= E\text{tr}[(\nu_{22}^{-1}\nu_{21}\nu_{12})] \\ &= E\text{tr}\left[\left(\sum_{g=1}^{G-1}\kappa_g\tilde{Z}_{(p-1),g}\tilde{Z}'_{(p-1),g}\right)^{-1}\sum_{g=1}^{G-1}\kappa_g^2\tilde{Z}_{(p-1),g}\tilde{Z}_{1,g}^2\tilde{Z}'_{(p-1),g}\right] \\ &= E\text{tr}\left[\left(\sum_{g=1}^{G}\kappa_g\tilde{Z}_{(p-1),g}\tilde{Z}'_{(p-1),g}\right)^{-1}\left(\sum_{g=1}^{G}\kappa_g^2\tilde{Z}_{(p-1),g}\tilde{Z}'_{(p-1),g}\right)\right],\end{aligned}$$

where the last equation follows from $E[\tilde{Z}_{1,g}^2] = 1$, and the independence between $\tilde{Z}_{1,g}$ and $\{\tilde{Z}_{(p-1),g}\}_{g=1}^G$. Denote M_g the $(p-1)\times(p-1)$ matrix $\tilde{Z}_{(p-1),g}\tilde{Z}'_{(p-1),g}$. Then, we can continue to express the last equation as

$$
\begin{aligned}
&= E\mathrm{tr}\left[\left\{\left(\sum_{g=1}^G \kappa_g\right)\cdot I_{p-1} - \sum_{g=1}^G \kappa_g\left(I_{p-1} - M_g\right)\right\}^{-1}\left(\sum_{g=1}^G \kappa_g^2 M_g\right)\right] \\
&= \frac{1}{\sum_{g=1}^G \kappa_g} E\mathrm{tr}\left[\left\{I_{p-1} - \left(\frac{1}{\sum_{g=1}^G \kappa_g}\right)\sum_{g=1}^G \kappa_g\left(I_{p-1} - M_g\right)\right\}^{-1}\left(\sum_{g=1}^G \kappa_g^2 M_g\right)\right], \\
&= \frac{1}{\sum_{g=1}^G \kappa_g} E\mathrm{tr}\left[\left\{I_{p-1} + \sum_{j=1}^{\infty}\left(\left(\frac{1}{\sum_{g=1}^G \kappa_g}\right)\sum_{g=1}^G \kappa_g\left(I_{p-1} - M_g\right)\right)^j\right\}\right. \\
&\left.\times\left(\sum_{g=1}^G \kappa_g^2 M_g\right)\right] \\
&= \frac{1}{\sum_{g=1}^G \kappa_g}\sum_{g=1}^G \kappa_g^2 E\mathrm{tr}\left[M_g\right] \qquad (37) \\
&+ \frac{1}{\sum_{g=1}^G \kappa_g} E\mathrm{tr}\left[\sum_{j=1}^{\infty}\left(\left(\frac{1}{\sum_{g=1}^G \kappa_g}\right)\cdot\sum_{g=1}^G \kappa_g\left(I_{p-1} - M_g\right)\right)^j\left(\sum_{g=1}^G \kappa_g^2 M_g\right)\right], \qquad (38)
\end{aligned}
$$

where the second last equation follows by the formula $(I-M)^{-1} = \sum_{j=0}^{\infty} M^k$ for any square matrix M such that $(I-M)$ is non-singular, and dominated convergence theorem. For the term in (37), we use $E[M_g] = E[\tilde{Z}_{(p-1),g}\tilde{Z}'_{(p-1),g}] = I_{p-1}$ and $\sum_{g=1}^G \kappa_g = 1 - \sum_{g=1}^G \lambda_g^2 = 1 + O(G^{-1})$, and obtain that

$$
\begin{aligned}
\frac{1}{\sum_{g=1}^G \kappa_g}\sum_{g=1}^G \kappa_g^2 E\mathrm{tr}\left[M_g\right] &= \frac{(p-1)}{1+O\left(G^{-1}\right)}\sum_{g=1}^G \kappa_g^2 \\
&= (p-1)\sum_{g=1}^G \kappa_g^2 + o(G^{-1}) \\
&= (p-1)\sum_{g=1}^G \lambda_g^2 + o(G^{-1}),
\end{aligned}
$$

where the last equation follows by Lemma 8. For the second term in (38), we have that

$$\begin{aligned}
&\left| E\mathrm{tr}\sum_{j=1}^{\infty}\left(\frac{1}{\sum_{g=1}^{G}\kappa_g}\sum_{g=1}^{G}\kappa_g\left(I_{p-1}-M_g\right)\right)^j\sum_{g=1}^{G}\kappa_g^2 M_g\right| \\
&\leq \frac{1}{1+O\left(G^{-1}\right)}\sum_{j=1}^{\infty}\sum_{g=1}^{G}\left(\kappa_g^{j+2}E\mathrm{tr}\left[\left|\left(\sum_{g=1}^{G}(I_{p-1}-M_g)\right)^j M_g\right|\right]\right) \\
&\leq O\left[\sum_{j=1}^{\infty}C^j\sum_{g=1}^{G}\kappa_g^{j+2}\right], \qquad (39)
\end{aligned}$$

where the positive number C satisfies

$$\left|E\mathrm{tr}(I_{p-1}-M_{g_1})\times\cdots(I_{p-1}-M_{g_j})\times M_g\right|\leq C^j.$$

Using the result in Lemma 8-(d), we can further express the upper bound (39) as

$$O\left[\frac{1}{G}\sum_{j=1}^{\infty}\left(\frac{C}{G}\right)^j\right]=o\left(G^{-1}\right).$$

Combining these results together, we obtain

$$\begin{aligned}
E[s_{12}s_{22}^{-1}s_{21}] &= E\mathrm{tr}\left[\left(\sum_{g=1}^{G}\kappa_g\tilde{Z}_{(p-1),g}\tilde{Z}'_{(p-1),g}\right)^{-1}\left(\sum_{g=1}^{G}\kappa_g^2\tilde{Z}_{(p-1),g}\tilde{Z}'_{(p-1),g}\right)\right] && (40)\\
&= (p-1)\sum_{g=1}^{G}\kappa_g^2+o(G^{-1}) && (41)\\
&= (p-1)\sum_{g=1}^{G}\lambda_g^2+o(G^{-1}), && (42)
\end{aligned}$$

which, together with (36), leads to the desired result.

(b) Note that

$$E(s_{11}-s_{12}s_{22}^{-1}s_{21})^2=Es_{11}^2+E[s_{12}s_{22}^{-1}s_{21}s_{12}s_{22}^{-1}s_{21}]-2E[s_{11}s_{12}s_{22}^{-1}s_{21}].$$

We want to show that

$$\begin{aligned}
Es_{11}^2 &= 1+o(G^{-1});\\
E\left(s_{11}s_{12}s_{22}^{-1}s_{21}\right) &= (p-1)\sum_{g=1}^{G}\lambda_g^2+o\left(G^{-1}\right);\\
E[s_{12}s_{22}^{-1}s_{21}s_{12}s_{22}^{-1}s_{21}] &= o(G^{-1}),
\end{aligned}$$

which directly imply the desired result. We prove each result in turn. First, we have $Es_{11}^2 = Var(s_{11})+\left(E[s_{11}]\right)^2$, and

$$Var(s_{11})=Var\left(\sum_{g=1}^{G}\kappa_g\tilde{Z}_{1,g}^2\right)=\sum_{g=1}^{G}\kappa_g^2Var(\tilde{Z}_{1,g}^2)=2\sum_{g=1}^{G}\kappa_g^2.$$

Thus,

$$
\begin{aligned}
Es_{11}^2 &= 2\sum_{g=1}^{G}\kappa_g^2 + \left[\sum_{g=1}^{G}\kappa_g\right]^2 \\
&= 2\left[\sum_{g=1}^{G}\lambda_g^2 + \left(\sum_{g=1}^{G}\lambda_g^2\right)^2 - 2\sum_{g=1}^{G}\lambda_g^3\right] + \left[1-\sum_{g=1}^{G}\lambda_g^2\right]^2 \\
&= 1 + 3\left(\sum_{g=1}^{G}\lambda_g^2\right)^2 - 4\sum_{g=1}^{G}\lambda_g^3 = 1 + o(G^{-1}). \qquad (43)
\end{aligned}
$$

Second, we have

$$
\begin{aligned}
E\left(s_{11}s_{12}s_{22}^{-1}s_{21}\right) &= E\left[\left(\sum_{g=1}^{G}\kappa_g\tilde{Z}_{1g}^2\right)\left(\sum_{g=1}^{G}\kappa_g\tilde{Z}_{1g}\tilde{Z}'_{(p-1),g}\right)\left(\sum_{g=1}^{G}\kappa_g\tilde{Z}_{(p-1),g}\tilde{Z}'_{(p-1),g}\right)^{-1}\right. \\
&\left.\times\left(\sum_{g=1}^{G}\kappa_g\tilde{Z}_{1g}\tilde{Z}_{(p-1),g}\right)\right] \\
&= \mathrm{tr}E\left[\left(\sum_{g_1=1}^{G}\kappa_{g_1}\tilde{Z}_{(p-1),g_1}\tilde{Z}'_{(p-1),g_1}\right)^{-1}\left(\sum_{g_2=1}^{G}\kappa_{g_2}\tilde{Z}_{1g_2}\tilde{Z}_{(p-1),g_2}\right)\right. \\
&\left.\times\left(\sum_{g_3=1}^{G}\kappa_{g_3}\tilde{Z}_{1g_3}\tilde{Z}'_{(p-1),g_3}\right)\left(\sum_{g_4=1}^{G}\kappa_{g_4}\tilde{Z}_{1g_4}^2\right)\right] \\
&= \mathrm{tr}E\left[\left(\sum_{g_1=1}^{G}\kappa_{g_1}\tilde{Z}_{(p-1),g_1}\tilde{Z}'_{(p-1),g_1}\right)^{-1}\sum_{g_2,g_3,g_4=1}^{G}\kappa_{g_2}\kappa_{g_3}\kappa_{g_4}\left(\tilde{Z}_{(p-1),g_2}\tilde{Z}'_{(p-1),g_3}\right)\right. \\
&\left.\times E\left(\tilde{Z}_{1g_2}\tilde{Z}_{1g3}\tilde{Z}_{1g_4}^2\right)\right].
\end{aligned}
$$

For $g_2, g_3, g_4 = 1, \ldots, G$, it is easy to check that

$$
\begin{aligned}
E\left(\tilde{Z}_{1g_2}\tilde{Z}_{1g3}\tilde{Z}_{1g_4}^2\right) &= \underbrace{E\left(\tilde{Z}_{1g_2}^2\tilde{Z}_{1g_4}^2\right)}_{=1} \text{ if } g_2 = g_3; \\
&= \underbrace{E\left(\tilde{Z}_{1g_2}^4\right)}_{=3} \text{ if } g_2 = g_3 = g_4; \\
&= 0 \text{ otherwise.}
\end{aligned}
$$

Using this, we can decompose the last equation into

$$\mathrm{tr}E\left[\left(\sum_{g_1=1}^{G}\kappa_{g_1}\tilde{Z}_{(p-1),g_1}\tilde{Z}'_{(p-1),g_1}\right)^{-1}\sum_{g_2=1}^{G}\kappa_{g_2}^2\tilde{Z}_{(p-1),g_2}\tilde{Z}'_{(p-1),g_2}\right]\times\left(\sum_{g_4=1}^{G}\kappa_{g_4}\right) \tag{44}$$

$$+3\mathrm{tr}E\left[\left(\sum_{g_1=1}^{G}\kappa_{g_1}\tilde{Z}_{(p-1),g_1}\tilde{Z}'_{(p-1),g_1}\right)^{-1}\sum_{g_2=1}^{G}\kappa_{g_2}^3\tilde{Z}_{(p-1),g_2}\tilde{Z}'_{(p-1),g_2}\right]. \tag{45}$$

Using the previous results in (40)–(41), we can express the first term (44) as

$$\begin{aligned}
&\mathrm{tr}E\left[\left(\sum_{g=1}^{G}\kappa_g\tilde{Z}_{(p-1),g}\tilde{Z}'_{(p-1),g}\right)^{-1}\sum_{g=1}^{G}\kappa_g^2\tilde{Z}_{(p-1),g}\tilde{Z}'_{(p-1),g}\right]\times\left(\sum_{g=1}^{G}\kappa_g\right)\\
&=\left[(p-1)\sum_{g=1}^{G}\lambda_g^2+o(G^{-1})\right]\left(1-\sum_{g=1}^{G}\lambda_g^2\right)\\
&=(p-1)\sum_{g=1}^{G}\lambda_g^2+o(G^{-1}),
\end{aligned} \tag{46}$$

and the second term (45) is equal to

$$\begin{aligned}
&3\mathrm{tr}E\left[\left(\sum_{g=1}^{G}\kappa_g\tilde{Z}_{(p-1),g}\tilde{Z}'_{(p-1),g}\right)^{-1}\sum_{g=1}^{G}\kappa_g^3\tilde{Z}_{(p-1),g}\tilde{Z}'_{(p-1),g}\right]\\
&=3\,(p-1)\sum_{g=1}^{G}\kappa_g^3+o\left(G^{-1}\right)=o(G^{-1}).
\end{aligned}$$

Therefore, we have

$$E\left(s_{11}s_{12}s_{22}^{-1}s_{21}\right)=(p-1)\sum_{g=1}^{G}\lambda_g^2+o\left(G^{-1}\right). \tag{47}$$

Finally, we can use similar arguments to show that

$$E[s_{12}s_{22}^{-1}s_{21}s_{12}s_{22}^{-1}s_{21}]=o(G^{-1}).$$

Details are omitted. This completes the proof of part (b).

(c) From the results in parts (a) and (b), it is straightforward that

$$\begin{aligned}
E[\big(s_{11} - s_{12}s_{22}^{-1}s_{21}\big) - 1]^2 &= 1 - 2\left(\sum_{g=1}^{G}\lambda_g^2\right)(p-1) \\
&\quad - 2\left[(1-\sum_{g=1}^{G}\lambda_g^2) - \left(\sum_{g=1}^{G}\lambda_g^2\right)(p-1)\right] + 1 + o(G^{-1}) \\
&= 2 - 2(1-\sum_{g=1}^{G}\lambda_g^2) + o(G^{-1}) \\
&= 2\sum_{g=1}^{G}\lambda_g^2 + o(G^{-1}),
\end{aligned}$$

which gives us the desired result. ■

Proof of Theorem 3. Let $Z_p = \sum_{g=1}^{G}\sqrt{\lambda_g}Z_{p,g}$. Then, we write

$$\begin{aligned}
p\mathbb{F}_\infty &= ||Z_p||^2\left(\frac{Z_p}{||Z_p||}\right)'\bar{\mathbb{S}}_{pp}^{-1}\left(\frac{Z_p}{||Z_p||}\right) \\
&\stackrel{d}{=} \frac{||Z_p||^2}{\left[\left(\frac{Z_p}{||Z_p||}\right)'\left(\sum_{g=1}^{G}\kappa_g\tilde{Z}_{p,g}\tilde{Z}_{p,g}'\right)^{-1}\left(\frac{Z_p}{||Z_p||}\right)\right]^{-1}}.
\end{aligned}$$

Since Z_p is normal $N(0, I_p)$, $||Z_p||^2$ is independent of $Z_p/||Z_p||$ that is the direction of Z_p. Thus, the numerator is independent of the denominator. Note that $||Z_p||^2 \stackrel{d}{=} \chi_p^2$. Let H be an orthonormal matrix such that $H = (Z_p/||Z_p||, H_{p-1})'$, where H_{p-1} is a $p \times (p-1)$ matrix. Then, by definition,

$$H\left(\frac{Z_p}{||Z_p||}\right) = e_1 = (1, 0, ..., 0)' \in \mathbb{R}^p.$$

Using this, the inverse of the denominator is equal to

$$\begin{aligned}
&\left(\frac{Z_p}{||Z_p||}\right)'H'H\left(\sum_{g=1}^{G}\kappa_g\tilde{Z}_{p,g}\tilde{Z}_{p,g}'\right)^{-1}H'H\left(\frac{Z_p}{||Z_p||}\right) \\
&\qquad = e_1'\left(\sum_{g=1}^{G}\kappa_g(H\tilde{Z}_{p,g})(H\tilde{Z}_{p,g})'\right)^{-1}e_1,
\end{aligned}$$

where $\{\tilde{Z}_{p,g}\}_{g=1}^{G}$ is independent of $||Z_p||^2$. Using the rotation invariance property of standard normal matrix, i.e. $H\tilde{Z}_{p,g} \stackrel{d}{=} \tilde{Z}_{p,g}$ for $g = 1, \dots G$, we then have

$$\mathbb{F}_\infty \stackrel{d}{=} \frac{1}{p}\frac{\chi_p^2}{\left[e_1'\left(\sum_{g=1}^{G}\kappa_g\tilde{Z}_{p,g}\tilde{Z}_{p,g}'\right)^{-1}e_1\right]^{-1}}.$$

By the matrix inverse formula, we have $e_1'(\sum_{g=1}^{G}\kappa_g\tilde{Z}_{p,g}\tilde{Z}_{p,g}')^{-1}e_1 = s_{11\cdot 2}^{-1}$, and this gives us the desire result.

Then, we have

$$\begin{aligned} P\left(p\mathbb{F}_\infty \leq x\right) &= E\left[\Xi_p\left(xe_1'\left[\sum\nolimits_{g=1}^{G}\kappa_g\zeta_g\zeta_g'\right]^{-1}e_1\right)\right] \\ &= E\left[\Xi_p(x(\nu_{11}-\nu_{12}\nu_{22}^{-1}\nu_{21})\right] \\ &= E[\Xi_p(x\nu_{11\cdot 2})] \end{aligned}$$

where

$$\sum_{g=1}^{G}\kappa_g\zeta_g\zeta_g' = \begin{pmatrix} \nu_{11} & \nu_{12} \\ \nu_{21} & \nu_{22} \end{pmatrix}.$$

and ν_{11} is a scalar.

Taking a Taylor expansion, we have

$$\begin{aligned} P\left(p\mathbb{F}_\infty \leq x\right) &= E\left[\Xi_p(x(\nu_{11}-\nu_{12}\nu_{22}^{-1}\nu_{21})\right] \\ &= \Xi_p(x) + \Xi_p'(x)xE\left[(\nu_{11}-\nu_{12}\nu_{22}^{-1}\nu_{21})-1\right] \\ &+ \frac{1}{2}\Xi_p''(x)x^2E\left[(\nu_{11}-\nu_{12}\nu_{22}^{-1}\nu_{21})-1\right]^2 \\ &+ \frac{1}{2}x^2E\left[\left(\Xi_p''(\tilde{x})-\Xi_p''(x)\right)\left[(\nu_{11}-\nu_{12}\nu_{22}^{-1}\nu_{21})-1\right]^2\right], \end{aligned}$$

where $\tilde{x}$ is between x and $x(\nu_{11}-\nu_{12}\nu_{22}^{-1}\nu_{21})$. Using Lemma 9, we have

$$\begin{aligned} P\left(p\mathbb{F}_\infty \leq x\right) &= \Xi_p(x) - \Xi_p'(x)xp\sum_{g=1}^{G}\lambda_g^2 + \Xi_p''(x)x^2\sum_{g=1}^{G}\lambda_g^2 + o(G^{-1}) \\ &= \Xi_p(x) + \left(\Xi_p''(x)x^2 - \Xi_p'(x)xp\right)\sum_{g=1}^{G}\lambda_g^2 + o(G^{-1}). \end{aligned}$$

■

Proof of Theorem ??. Lemma 8 (f) states

$$\frac{1}{\tau} = \frac{\mu(G^*-p+1)}{G^*} = 1 - p\sum_{g=1}^{G}\lambda_g^2 + o\left(\frac{1}{G}\right).$$

Thus,

$$
\begin{aligned}
&P\left(\frac{pG^*}{\mu(G^*-p+1)}\mathcal{F}_{p,G^*-p+1}\leq x\right)\\
&=E\Xi_p\left(\frac{1}{\tau}x\frac{\chi^2_{G^*-p+1}}{(G^*-p+1)}\right)\\
&=\Xi_p\left(\frac{1}{\tau}x\right)+\Xi_p'\left(\frac{1}{\tau}x\right)\frac{1}{\tau}xE\left(\frac{\chi^2_{G^*-p+1}}{(G^*-p+1)}-1\right)\\
&+\frac{1}{2}\Xi_p''\left(\frac{1}{\tau}x\right)\left(\frac{1}{\tau}x\right)^2\frac{1}{(G^*-p+1)^2}E\left(\chi^2_{G^*-p+1}-(G^*-p+1)\right)^2+o\left(\frac{1}{G}\right)\\
&=\Xi_p\left(\frac{1}{\tau}x\right)+\Xi_p''\left(\frac{1}{\tau}x\right)\left(\frac{1}{\tau}x\right)^2\frac{1}{G^*-p+1}+o\left(\frac{1}{G}\right)\\
&=\Xi_p\left(x\right)+\Xi_p'\left(x\right)x\left(\frac{1}{\tau}-1\right)+\Xi_p''\left(x\right)x^2\left(\frac{\mu(G^*-p+1)}{G^*}\right)^2\frac{1}{G^*-p+1}+o\left(\frac{1}{G}\right)\\
&=\Xi_p(x)+\left(\Xi_p''(x)x^2-\Xi_p'(x)xp\right)\sum_{g=1}^{G}\lambda_g^2+o(G^{-1}).
\end{aligned}
\tag{48}
$$

because

$$
\begin{aligned}
\frac{1}{G^*-p+1}&=\frac{1}{\left(\sum_{g=1}^G\kappa_g\right)^2/\sum_{g=1}^G\kappa_g^2-p+1}\\
&=\frac{\sum_{g=1}^G\kappa_g^2}{\left(\sum_{g=1}^G\kappa_g\right)^2-(p-1)\sum_{g=1}^G\kappa_g^2}\\
&=\sum_{g=1}^{G}\lambda_g^2+o\left(G^{-1}\right).
\end{aligned}
$$

Combining the result in (48) with Theorem 3, we have

$$
P\left(\frac{\mu(G^*-p+1)}{G^*}\mathbb{F}_\infty\leq x\right)=P\left(\mathcal{F}_{p,G^*-p+1}\leq x\right)+o(G^{-1}),
$$

■

Proof of Proposition 5. Note that the cluster-specific component γ_g can also be written as

$$
\gamma_g=R(X'X)^{-1}Var(X_g'\epsilon_g|X)(X'X)^{-1}R'.
$$

Under Assumption 4, we have.

$$
n_g^{-1}X_g'X_g\xrightarrow{p}\Gamma_g \text{ and } \Gamma_1=\ldots=\Gamma_G=\Gamma,
$$

which implies

$$
n^{-1}X'X=\sum_{g=1}^{G}\left(\frac{n_g}{n}\right)\left(n_g^{-1}X_g'X_g\right)\xrightarrow{p}\sum_{g=1}^{G}\lambda_g\Gamma_g=\Gamma.
$$

Also, Assumption 5 leads us to get

$$n_g^{-1} Var(X_g' \epsilon_g | X) \xrightarrow{p} \Omega_g, \Omega_1 = \ldots = \Omega_G = \Omega.$$

Thus, we have

$$\begin{aligned} n \times \gamma_g &= R(X'X)^{-1} Var(X_g' \epsilon_g | X)(X'X)^{-1} R' \\ &= R(n^{-1}X'X)^{-1} n^{-1} Var(X_g' \epsilon_g | X)(n^{-1}X'X)^{-1} R' \\ &\xrightarrow{p} \lambda_g \times R(\Gamma^{-1}\Omega\Gamma^{-1})R', \end{aligned}$$

which implies

$$\frac{\gamma_g}{\bar{\gamma}} \xrightarrow{p} \frac{\lambda_g \times \left(R(\Gamma^{-1}\Omega\Gamma^{-1})R'\right)}{G^{-1}\sum_{g=1}^{G} \lambda_g \times (R(\Gamma^{-1}\Omega\Gamma^{-1})R')} = \frac{\lambda_g}{G^{-1}}.$$

Using this result, we have the fixed-G probability limit of the effective number of clusters as

$$G_{\text{eff}} = \frac{G}{1 + \frac{1}{G}\sum_{g=1}^{G} \frac{(\gamma_g - \bar{\gamma})^2}{\bar{\gamma}^2}} \xrightarrow{p} \frac{G}{1 + \frac{1}{G}\sum_{g=1}^{G} (G\lambda_g - 1)^2},$$

where the right hand side is equal to

$$\frac{G}{1 + G\sum_{g=1}^{G} \lambda_g^2 - 2\sum \lambda_g + 1} = \frac{1}{\sum_{g=1}^{G} \lambda_g^2}.$$

■

Proof of Theorem 6. We first show that $\hat{V}_{R,JK} \xrightarrow{d} \tilde{\Lambda}\tilde{\mathbb{S}}_{pp}\tilde{\Lambda}'$. For each $g = 1, \ldots, G$, we have

$$\begin{aligned} (I_{n_g} - P_{gg})^{-1}\hat{\epsilon}_g &= (I_{n_g} - P_{gg})^{-1}(y_g - X_g\hat{\beta}) \\ &= (I_{n_g} - P_{gg})^{-1}(y_g - X_g(X'X)^{-1}X'y) \\ &= (I_{n_g} - P_{gg})^{-1}\left\{ \left(I_{n_g} - P_{gg}\right) y_g - X_g(X'X)^{-1} \sum_{h \neq g} X_h' y_h \right\} \\ &= y_g - (I_{n_g} - P_{gg})^{-1} X_g (X'X)^{-1} \sum_{h \neq g} X_h' y_h, \end{aligned} \tag{49}$$

where the third equation follows from $X'y = X_g'y_g + \sum_{h \neq g} X_h'y_h$. By Sherman-Morrison-Woodbury formula and the definition of $P_{gg} = X_g(X'X)^{-1}X_g'$,[2] we have

$$(I_{n_g} - P_{gg})^{-1} = I_{n_g} + X_g(X'X - X_g'X_g)^{-1}X_g',$$

and

$$\begin{aligned} (I_{n_g} - P_{gg})^{-1} X_g (X'X)^{-1} &= X_g(X'X)^{-1} + X_g(X'X - X_g'X_g)^{-1}X_g'X_g(X'X)^{-1} \\ &= X_g \left\{ (X'X)^{-1} + \left(\sum_{h \neq g} X_h'X_h \right)^{-1} X_g'X_g(X'X)^{-1} \right\} \\ &= X_g \left(\sum_{h \neq g} X_h'X_h \right)^{-1}. \end{aligned} \tag{50}$$

[2] $(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$,

The last equation holds because

$$
\begin{aligned}
&(X'X)^{-1} + \left(\sum_{h\neq g} X_h'X_h\right)^{-1} X_g'X_g(X'X)^{-1} \\
&= \left(\sum_{h\neq g} X_h'X_h\right)^{-1} \left(X'X - X_g'X_g\right)(X'X)^{-1} + \left(\sum_{h\neq g} X_h'X_h\right)^{-1} X_g'X_g(X'X)^{-1} \\
&= \left(\sum_{h\neq g} X_h'X_h\right)^{-1}.
\end{aligned}
$$

Combining (50) into (49), we have

$$
\begin{aligned}
(I_{n_g} - P_{gg})^{-1}\hat{\epsilon}_g &= y_g - X_g\left(\sum_{h\neq g} X_h'X_h\right)^{-1} \sum_{h\neq g} X_h'y_h \\
&= y_g - X_g\hat{\beta}_{-g},
\end{aligned}
$$

where $\hat{\beta}_{-g}$ is the leave-cluster g-out OLS estimator. This result further implies that

$$
\begin{aligned}
(I_{n_g} - P_{gg})^{-1}\hat{\epsilon}_g &= y_g - X_g\hat{\beta}_{-g} \\
&= X_g(\beta - \hat{\beta}_{-g}) + \epsilon_g \\
&= \epsilon_g - X_g\left(\sum_{h\neq g} X_h'X_h\right)^{-1} \sum_{h\neq g} X_h'\epsilon_h.
\end{aligned}
$$

Now, for each $g = 1, \ldots, G$, we have

$$
\begin{aligned}
\frac{1}{\sqrt{n}} X_g^{'}\hat{\epsilon}^*_{g,JK} &= \sqrt{\frac{n-n_g}{n}} \times \left(\frac{1}{\sqrt{n}} X_g^{'}(I_{n_g} - P_{gg})^{-1}\hat{\epsilon}_g\right) \\
&= \sqrt{1-\lambda_g} \times \left\{\frac{1}{\sqrt{n}} X_g^{'}\epsilon_g - \frac{1}{n} X_g'X_g \left(\frac{1}{n}\sum_{h\neq g} X_h'X_h\right)^{-1} \sum_{h\neq g} \frac{1}{\sqrt{n}} X_h'\epsilon_h\right\}(1 + o_p(1)).
\end{aligned}
$$

Let $Z_g \overset{\text{i.i.d}}{\sim} N(0, I_{d_\theta})$. By Assumptions 1, 3–5,

$$
\frac{1}{\sqrt{n}} X_g^{'}\epsilon_g \xrightarrow{d} \sqrt{\lambda_g}\Lambda Z_g \text{ and } \sum_{h\neq g} \frac{1}{\sqrt{n}} X_h'\epsilon_h \xrightarrow{d} \sum_{h\neq g} \Lambda\sqrt{\lambda_h} Z_h
$$

holds jointly, and

$$
\frac{1}{n} X_g'X_g \xrightarrow{p} \lambda_g\Gamma \text{ and } \frac{1}{n}\sum_{h\neq g} X_h'X_h \xrightarrow{p} \left(\sum_{h\neq g} \lambda_h\right)\Gamma.
$$

Using these results and $\sum_{h\neq g} \lambda_h = 1 - \lambda_g$, we obtain that

$$
\frac{1}{\sqrt{n}} X_g^{'}\hat{\epsilon}^*_{g,JK} \xrightarrow{d} \sqrt{1-\lambda_g} \times \Lambda \left\{\sqrt{\lambda_g} Z_g - \left(\frac{\lambda_g}{1-\lambda_g}\right)\sum_{h\neq g} \sqrt{\lambda_h} Z_h\right\}
$$

by Slutsky's theorem. Note that

$$\sqrt{\lambda_g}Z_g - \left(\frac{\lambda_g}{1-\lambda_g}\right)\sum_{h\neq g}\sqrt{\lambda_h}Z_h = \frac{1}{1-\lambda_g}\left\{(1-\lambda_g)\sqrt{\lambda_g}Z_g - \lambda_g\sum_{h\neq g}\sqrt{\lambda_h}Z_h\right\}$$
$$= \frac{1}{1-\lambda_g}\left(\sqrt{\lambda_g}Z_g - \lambda_g\sum_{h=1}^{G}Z_h\right),$$

which implies that

$$\frac{1}{\sqrt{n}}X_g^{'}\hat{\epsilon}^*_{g,JK} \xrightarrow{d} \frac{1}{\sqrt{1-\lambda_g}}\left(\sqrt{\lambda_g}Z_g - \lambda_g\sum_{h=1}^{G}Z_h\right).$$

Thus, the scaled CCE matrix $\hat{V}_{R,JK}$ converges in distribution to a random matrix:

$$\hat{V}_{R,JK} = R(n^{-1}X'X)^{-1}\left[\frac{1}{n}\sum_{g=1}^{G}X_g'\hat{\epsilon}^*_{g,JK}\hat{\epsilon}^{*\prime}_{g,JK}X_g\right](n^{-1}X'X)^{-1}R'$$
$$\xrightarrow{d} R\Gamma^{-1}\left\{\sum_{g=1}^{G}\frac{1}{(1-\lambda_g)}\left(\sqrt{\lambda_g}Z_{p,g} - \lambda_g\sum_{g=1}^{G}\sqrt{\lambda_g}Z_{p,g}\right)\right.$$
$$\left.\times\left(\sqrt{\lambda_g}Z_{p,g} - \lambda_g\sum_{g=1}^{G}\sqrt{\lambda_g}Z_{p,g}\right)\right\}\Gamma^{-1}R'$$
$$\stackrel{d}{=} \tilde{\Lambda}\tilde{\mathbb{S}}_{pp}\tilde{\Lambda}',$$

which is the desired result. ■

Proof of Theorem 7. The proof is basically the same as the proof for the F approximation to $\mathbb{F}_\infty$ in Theorem **??**. Recall that Theorem 6 states

$$F_{JK}\left(\hat{\theta}\right) \xrightarrow{d} \mathbb{F}_{JK,\infty} = \left(\sum_{g=1}^{G}\sqrt{\lambda_g}Z_{p,g}\right)'\tilde{\mathbb{S}}_{pp}^{-1}\left(\sum_{g=1}^{G}\sqrt{\lambda_g}Z_{p,g}\right).$$

Using the same steps in the proof of Lemma 1 (ii), we have the following distributional equivalence:

$$\tilde{\mathbb{S}}_{pp} \stackrel{d}{=} \sum_{g=1}^{G}\nu_g\tilde{Z}_{p,g}\tilde{Z}_{p,g},$$

where $\tilde{Z}_{p,g} \overset{\text{i.i.d}}{\sim} N(0, I_p)$ and $\nu_g's$ are G eigenvalues of $D_{JK}D_{JK}'$ where $D = diag\left(\sqrt{\lambda}\right)\left[diag\left(1/\sqrt{1-\lambda}\right) - 11'diag\left(\right.\right.$

We first define the effective degree of freedom

$$G^*_{JK} = \frac{\left(\sum_{g=1}^{G}\nu_g\right)^2}{\sum_{g=1}^{G}\nu_g^2}.$$

Due to the identity between the trace and sum of eigenvalues of a matrix, simple calculation yields

$$\sum_{g=1}^{G}\nu_g = 1 \text{ and } \sum_{g=1}^{G}\nu_g^2 = \sum_{g=1}^{G}\frac{\lambda_g^2}{(1-\lambda_g)^2} - 2\sum_{g=1}^{G}\frac{\lambda_g^3}{(1-\lambda_g)^2} + \left(\sum_{g=1}^{G}\frac{\lambda_g^2}{1-\lambda_g}\right)^2.$$

Therefore, we can write

$$G^*_{JK} = \frac{1}{\sum_{g=1}^{G} \frac{\lambda_g^2}{(1-\lambda_g)^2} + \left(\sum_{g=1}^{G} \frac{\lambda_g^2}{1-\lambda_g}\right)^2 - 2\sum_{g=1}^{G} \frac{\lambda_g^3}{(1-\lambda_g)^2}}.$$

Next, note that

$$\begin{aligned}
\frac{1}{\sqrt{n}} X_g^{'} \hat{\epsilon}^*_{g,(2)} &= \frac{1}{\sqrt{n}} X_g^{'} (I_{n_g} - P_{gg})^{-1/2} \hat{\epsilon}_g \\
&= \frac{1}{\sqrt{n}} \left\{ X_g^{'} (I_{n_g} - P_{gg})^{1/2} \times \left(\sqrt{\frac{n}{n-n_g}} \right) \right\} \times \sqrt{\frac{n-n_g}{n}} (I_{n_g} - P_{gg})^{-1} \hat{\epsilon}_g \\
&= \frac{1}{\sqrt{n}} \left\{ X_g^{'} (I_{n_g} - P_{gg})^{1/2} \times \left(\sqrt{\frac{n}{n-n_g}} \right) \right\} \hat{\epsilon}^*_{g,(3)}. \qquad (51)
\end{aligned}$$

For simplicity, let $d_\theta = 1$. Then,

$$\left\| \frac{1}{\sqrt{n}} X_g^{'} (I_{n_g} - P_{gg})^{1/2} \left(\sqrt{\frac{n}{n-n_g}} \right) - \frac{1}{\sqrt{n}} X_g^{'} \right\|^2 = \frac{X_g^{'} (I_{n_g} - P_{gg}) X_g}{n} \times \left(\frac{n}{n-n_g} \right) - \frac{1}{n} X_g' X_g \qquad (52)$$

$$- \frac{2}{n} X_g^{'} (I_{n_g} - P_{gg})^{1/2} X_g \left(\sqrt{\frac{n}{n-n_g}} \right) \qquad (53)$$

For the right-hand-side term (52),

$$\begin{aligned}
&\frac{X_g^{'} (I_{n_g} - P_{gg}) X_g}{n} \times \left(\frac{n}{n-n_g} \right) - \frac{1}{n} X_g' X_g \\
&= \left(\frac{n}{n-n_g} \right) \times \left(\frac{X_g^{'} X_g}{n} - \frac{X_g^{'} X_g (X'X)^{-1} X_g' X_g}{n} \right) - \frac{1}{n} X_g' X_g \\
&= \left\{ \frac{1}{1-\lambda_g} \times \left(\lambda_g \Gamma - \lambda_g^2 \Gamma \right) - \lambda_g \Gamma \right\} + o_p(1) \\
&= o_p(1).
\end{aligned}$$

For (53),

$$\frac{2}{n} X_g^{'} (I_{n_g} - P_{gg})^{1/2} X_g \left(\sqrt{\frac{n}{n-n_g}} \right) \leq \left\| \frac{2}{n} X_g^{'} (I_{n_g} - P_{gg})^{1/2} \right\| \left\| X_g \sqrt{\frac{n}{n-n_g}} \right\|,$$

and

$$\begin{aligned}
\left\| \frac{2}{n} X_g^{'} (I_{n_g} - P_{gg})^{1/2} \right\|^2 &= \frac{4}{n} \times \frac{X_g^{'} (I_{n_g} - P_{gg}) X_g}{n} \\
&= \frac{4}{n} \times \lambda_g (1-\lambda_g) \Gamma + o_p(1),
\end{aligned}$$

Thus, we have

$$\left\| \frac{2}{n} X_g^{'} (I_{n_g} - P_{gg})^{1/2} \right\|^2 \left\| X_g \sqrt{\frac{n}{n-n_g}} \right\|^2 = O_p\left(\frac{1}{n} \right),$$

which shows that

$$\frac{1}{\sqrt{n}} X_g^{'} (I_{n_g} - P_{gg})^{1/2} \left(\sqrt{\frac{n}{n - n_g}} \right) = \frac{1}{\sqrt{n}} X_g^{'} + o_p(1).$$

From this result and (51), we have that

$$\frac{1}{\sqrt{n}} X_g^{'} \hat{\epsilon}_{g,(2)}^{*} = \frac{1}{\sqrt{n}} X_g^{'} \hat{\epsilon}_{g,(3)}^{*} + o_p(1),$$

which shows the asymptotic equivalence $\hat{V}_{R,(3)} = \hat{V}_{R,(2)} + o_p(1)$. This completes the proof. ■