Journal of Econometrics 210 (2019) 268–290

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Asymptotic theory for clustered samples[☆]

Bruce E. Hansen[a,*,1], Seojeong Lee[b,2]

[a] University of Wisconsin, United States
[b] University of New South Wales, Australia

ARTICLE INFO

Article history:
Received 17 May 2018
Received in revised form 29 November 2018
Accepted 2 February 2019
Available online 10 February 2019

ABSTRACT

We provide a complete asymptotic distribution theory for clustered data with a large number of independent groups, generalizing the classic laws of large numbers, uniform laws, central limit theory, and clustered covariance matrix estimation. Our theory allows for clustered observations with heterogeneous and unbounded cluster sizes. Our conditions cleanly nest the classical results for i.n.i.d. observations, in the sense that our conditions specialize to the classical conditions under independent sampling. We use this theory to develop a full asymptotic distribution theory for estimation based on linear least-squares, 2SLS, nonlinear MLE, and nonlinear GMM.

1. Introduction

Clustered samples are widely used in current applied econometric practice. Despite this dominance, there is little formal large-sample theory for estimation and inference. This paper provides such a foundation. We develop a complete, rigorous, and easily-interpretable asymptotic distribution theory for the "large number of clusters" framework. Our theory allows heterogeneous and growing cluster sizes, but requires that the number of clusters G grows with sample size n. Our core theory provides a weak law of large numbers (WLLN), central limit theorem (CLT), and consistent clustered variance estimation for clustered sample means. We also provide uniform laws of large numbers and uniform consistent clustered variance estimation appropriate for the distribution theory of nonlinear econometric estimators.

We apply this core theory to develop large sample distribution theory for standard econometric estimators: linear least-squares, 2SLS, MLE, and GMM. For each, we provide conditions for consistent estimation, asymptotic normality, consistent covariance matrix estimation, and asymptotic distributions for t-ratios and Wald statistics. The theory provided in this paper is the first formal theory for such econometric estimators allowing for clustered dependence.

Our assumptions are minimal, requiring only uniform integrability for the WLLN and squared uniform integrability for the CLT and clustered covariance matrix estimators, plus the requirement that individual clusters are asymptotically negligible. Our results show that there are inherent trade-offs in the conditions between the allowed degree of heterogeneity in cluster sizes and the number of finite moments. These trade-offs are least restrictive for the WLLN, are more restrictive for the CLT and consistent cluster covariance matrix estimation, and are strongest for CLTs applied to clustered second moments. These trade-offs do not arise in the independent sampling context.

[☆] We thank the Co-Editor Han Hong and two referees for helpful comments on a previous version, and Morten Nielsen and James MacKinnon for valuable conversations and suggestions.

[*] Corresponding author.
E-mail address: bruce.hansen@wisc.edu (B.E. Hansen).

[1] Hansen thanks the National Science Foundation and the Phipps Chair for research support.

[2] Lee acknowledges that this research was supported under the Australian Research Council Discovery Early Career Researcher Award (DECRA) funding scheme (project number DE170100787).

https://doi.org/10.1016/j.jeconom.2019.02.001
0304-4076/© 2019 Elsevier B.V. All rights reserved.

We show that under clustering the convergence rate depends on the degree of clustered dependence. Convergence rates may equal the square root of the sample size, the square root of the number of clusters, be a rate in between these two, or even slower than both. Our assumptions and theory allow for these possibilities. This is in contrast to the existing literature, which imposes specific rate assumptions. One useful finding is that the rate does not need to be known by the user; the asymptotic distribution of t-ratios and Wald statistics does not depend on the underlying rate of convergence. This generalizes similar results in Hansen (2007) and related results in Tabord-Meehan (2018).

This paper makes the following technical contributions. We show that the key to extending the classical WLLN and CLT to cluster-level data is developing uniform integrability bounds for cluster sums. To allow for arbitrary within-cluster dependence, this means that such bounds will be scaled by cluster sizes. This leads to bounds on the degree of cluster size heterogeneity which can be allowed under cluster dependence. Some of the most difficult technical work presented here is the extension of classical results to clustered covariance matrix estimators. These are not sample averages, but rather average across clusters of squared cluster sums. Handling such estimators requires a new technical treatment.

Clustered dependence in econometrics dates to the work of Moulton (1986, 1990), Liang and Zeger (1986), and in particular Arellano (1987), who proposed the popular cluster-robust covariance matrix estimator. The method was popularized by the implementation in Stata by Rogers (1993) and the widely-cited paper of Bertrand et al. (2004). Surveys can be found in Wooldridge (2003), Cameron and Miller (2011, 2015), MacKinnon (2012, 2016), and textbook treatments in Angrist and Pischke (2009) and Wooldridge (2010).

The "large G" asymptotic theory develops normal approximations under the assumption that $G \to \infty$. The earliest treatment appears in White (1984). Wooldridge (2010) asserts a distribution theory under the assumption that the cluster sizes are fixed. C. Hansen (2007) provides two sets of asymptotic results, including both $\sqrt{G}$ and $\sqrt{n}$ convergence rates under two distinct assumptions on the rate of convergence of the estimation variance. His results are derived under the assumption that all clusters are identical in size. Carter et al. (2017) provided asymptotic results allowing for heterogeneous clusters, but their results are limited by atypical regularity conditions. Independently of this paper, Djogbenou et al. (2018) have provided a rigorous asymptotic theory for heterogeneous clusters, with similar but stronger regularity conditions than ours. Their primary focus is theory for regression wild bootstrap, while our focus is regularity conditions for general econometric estimators.

An alternative to the "large G" asymptotic is the "fixed G" framework, which leads to a non-normal inference theory. Contributions to this literature include C. Hansen (2007), Bester et al. (2011), and Ibragimov and Müller (2010, 2016). A related paper is Conley and Taber (2011) which provide an asymptotic theory under the assumption of a small number of groups with policy changes. Canay et al. (2017) provide approximate randomization tests.

Small sample approaches to cluster robust inference include Donald and Lang (2007), Imbens and Kolesár (2016), and Young (2016). Bootstrap approaches are provided by Cameron et al. (2008), and MacKinnon and Webb (2017, 2018).

A recent contribution which develops cluster-robust inference for GMM is Hwang (2017).

The organization of the paper is as follows. After Section 2, which introduces cluster sampling, Sections 3–8 cover the core asymptotic theory, providing rigorous conditions for the WLLN (Section 3), rates of convergence (Section 4), the CLT (Section 5), cluster-robust covariance matrix estimation (Section 6), the ULLN (Section 7), and the CLT for clustered second moments (Section 8). Following this, we provide the distribution theory for the core econometric estimators, specifically linear regression and 2SLS (Section 9), Maximum Likelihood (Section 10), and GMM (Section 11). Each of these latter sections is written self-sufficiently, so they can be used directly by readers. Proofs of the core theorems are provided in the Appendix, and proofs for the applications are provided in the Supplemental Appendix.

2. Cluster sampling

The observations are $X_i \in \mathbb{R}^p$ for $i = 1, \ldots, n$. They are grouped into G mutually independent known clusters, indexed $g = 1, \ldots, G$, where the gth cluster has n_g observations. The clustering can be due to the sampling scheme, or done by the researcher due to known correlation structures. The number of observations n_g per cluster (the "cluster sizes") may vary across clusters. The total number of observations are $n = \sum_{g=1}^{G} n_g$. It will also be convenient to double-index the observations as X_{gj} for $g = 1, \ldots, G$ and $j = 1, \ldots, n_g$.

As is conventional in the clustering literature, the only dependence assumption we make is that the observations are independent across clusters, while the dependence within each cluster is unrestricted. Furthermore, we do not require that the observations or clusters come from identical distributions. Thus our framework includes i.n.i.d (independent, not necessarily identically distributed) as the special case $n_g = 1$.

The notation and assumptions allow for linear panel data models with cluster-specific fixed effects. In this case the observations X_{gj} should be viewed as clustered-demeaned observations. Another common application is linear panel data models with both cluster-specific and time-specific fixed effects. Our assumptions do not cover this case as removing the time effects will induce cross-cluster correlations. This is essentially "multiway" clustering and requires different methods. See MacKinnon et al. (2017).

Our distributional framework is asymptotic as n and G simultaneously diverge to infinity. This is typically referred to as the "large G" framework. Our assumptions, however, will allow G to diverge at a rate slower than n, by allowing the cluster sizes n_g to diverge. This is in contrast to the early asymptotic theory for clustering, which implicitly assumed that the cluster sizes were bounded.

Our theory assumes that the clusters are known, and observations are independent across clusters. This is a substantive restriction. Alternatively, it may be possible to develop a distribution theory which allows weak dependence across clusters, but we do not do so here.

A word on notation. For a vector a let $\|a\| = \left(a'a\right)^{1/2}$ denote the Euclidean norm. For a positive semi-definite matrix A let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote its smallest and largest eigenvalue, respectively. For a general matrix A let $\|A\| = \sqrt{\lambda_{\max}\left(A'A\right)}$ denote the spectral norm. For a positive semi-definite matrix A let $A^{1/2}$ denote the symmetric square root matrix such that $A^{1/2}A^{1/2} = A$. We let C denote a generic positive constant, that may be different in different uses.

3. Weak law of large numbers

For our core theory (WLLN & CLT), we focus on the sample mean $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ as an estimator of $E\overline{X}_n$. It will be convenient to define the cluster sums

$$\widetilde{X}_g = \sum_{j=1}^{n_g} X_{gj}$$

which are mutually independent under clustered sampling. The sample mean can then be written as

$$\overline{X}_n = \frac{1}{n}\sum_{g=1}^{G} \widetilde{X}_g.$$

We use the following regularity condition.

Assumption 1. As $n \to \infty$

$$\max_{g \le G} \frac{n_g}{n} \to 0. \tag{1}$$

Theorem 1 (*WLLN for Clustered Means*)**.** *If Assumption 1 holds and*

$$\lim_{M\to\infty} \sup_i \left(E\left\|X_i\right\| 1\left(\left\|X_i\right\| > M\right)\right) = 0 \tag{2}$$

then as $n \to \infty$,

$$\left\|\overline{X}_n - E\overline{X}_n\right\| \xrightarrow{p} 0. \tag{3}$$

The condition (2) states that X_i is uniformly integrable.[3] This condition is identical to the standard condition for the WLLN for independent heterogeneous observations, and thus Theorem 1 is a direct generalization of the WLLN for i.n.i.d. samples. (2) simplifies to $E\left\|X_i\right\| < \infty$ when the observations have identical marginal distributions. A sufficient condition allowing for distributional heterogeneity is $\sup_i E\left\|X_i\right\|^r < \infty$ for some $r > 1$.

Assumption 1 states that each cluster size n_g is asymptotically negligible. This implies $G \to \infty$, so we do not explicitly need to list the latter as an assumption. Assumption 1 allows for considerable heterogeneity in cluster sizes. It allows the cluster sizes to grow with sample size, so long as the growth is not proportional. For example, it allows clusters to grow at the rate $n_g = n^{\alpha}$ for $0 \le \alpha < 1$.

Assumption 1 is necessary for parameter estimation consistency while allowing arbitrary within-cluster dependence. Otherwise a single cluster could dominate the sample average. To see this, suppose that there is a cluster ℓ such that all observations within the cluster are identical, so that $X_{\ell j} = Z_{\ell}$ for some non-degenerate random variable Z_{ℓ}, and that this cluster violates Assumption 1, so that $n_{\ell}/n \to c > 0$. Suppose for all other clusters that $EX_{gj} = 0$ and $n_g/n \to 0$. Then $\overline{X}_n \xrightarrow{p} Z_{\ell}$ and is inconsistent. Thus Assumption 1 is necessary for the WLLN (3) if we allow for unstructured cluster heterogeneity.

Assumption 1 is equivalent to the condition

$$\frac{\sum_{g=1}^{G} n_g^2}{n^2} \to 0. \tag{4}$$

To see this, first observe that since $\sum_{g=1}^{G} n_g = n$, the left-hand-side of (4) is smaller than $\max_{g\le G} n_g/n \to 0$ under Assumption 1. Thus Assumption 1 implies (4). Second,

$$\max_{g\le G}\frac{n_g}{n} = \left(\max_{g\le G}\frac{n_g^2}{n^2}\right)^{1/2} \le \left(\sum_{g=1}^{G}\frac{n_g^2}{n^2}\right)^{1/2} \to 0$$

under (4). Thus (4) implies Assumption 1, so the two are equivalent.

[3] A referee points out that the sup in (2) could be weakened to an average. However our later results will use uniform integrability conditions similar to (2) so we state all results in this format.

4. Rate of convergence

Under i.i.d. sampling the rate of convergence of the sample mean is $n^{-1/2}$. Clustering can alter the rate of convergence. In this section we explore possible rates of convergence. From the work of Hansen (2007) it has been understood that if the dependence within each cluster is weak then the rate of convergence would be the i.i.d. rate $n^{-1/2}$ but if the dependence within each cluster is strong then the rate of convergence would be determined by the number of clusters: $G^{-1/2}$. What we now show is that the rate of convergence can be in between or even slower than these rates.

The convergence rate can be calculated as the standard deviation of the sample mean. For simplicity we focus on the scalar case $p = 1$. The standard deviation of $\overline{X}_n$ is

$$\mathrm{sd}\left(\overline{X}_n\right) = \frac{1}{n}\left(\sum_{g=1}^{G} \mathrm{var}(\widetilde{X}_g)\right)^{1/2}.$$

We now consider several examples. For our first four we take the case where the clusters are all the same size: $n_g = n^{\alpha}$ for $0 < \alpha < 1$. In this case the number of clusters is $G = n^{1-\alpha}$.

We first consider a case where the convergence is the i.i.d. rate $n^{-1/2}$.

Example 1. The observations are independent within each cluster and $\mathrm{var}(X_i) = 1$. Then

$$\mathrm{var}(\widetilde{X}_g) = n_g = n^{\alpha}$$

and

$$\mathrm{sd}\left(\overline{X}_n\right) = n^{-1/2}.$$

The $n^{-1/2}$ rate extends to any case where the within-cluster dependence is weak, including autoregressive and moving average dependence.

Our second example is a case where the convergence is determined by the number of clusters.

Example 2. The observations are identical within each cluster (e.g. perfectly correlated) and $\mathrm{var}(X_i) = 1$. Then

$$\mathrm{var}(\widetilde{X}_g) = n_g^2 = n^{2\alpha}$$

and

$$\mathrm{sd}\left(\overline{X}_n\right) = n^{-(1-\alpha)/2} = G^{-1/2}.$$

The assumption that the observations are perfectly correlated is not essential to obtain the $G^{-1/2}$ rate. What is important is that there is a common component to the observations within a cluster.

Our third example is a case where the convergence rate is in between the above two cases. Not surprisingly, it can obtained by constructing strong but decaying within-cluster dependence.

Example 3. The observations are correlated within each cluster with $\mathrm{var}(X_i) = 1$ and $\mathrm{cov}(X_{gj}, X_{gl}) = 1/|j - l|$. Then

$$\mathrm{var}(\widetilde{X}_g) \sim n_g \log n_g \sim n^{\alpha} \log n$$

and

$$\mathrm{sd}\left(\overline{X}_n\right) \sim \sqrt{\log n/n}.$$

Furthermore, $G\mathrm{var}\left(\overline{X}_n\right) \to 0$. Thus $\mathrm{sd}\left(\overline{X}_n\right)$ converges at a rate in between $n^{-1/2}$ and $G^{-1/2}$.

Our next two examples are somewhat surprising. They are cases where the convergence rate is slower than both $n^{-1/2}$ and $G^{-1/2}$.

Example 4. The observations follow random walks within each cluster: $X_{gj} = X_{gj-1} + \varepsilon_{gj}$ with ε_{gj} i.i.d. $(0, 1)$ and $X_{g0} = 0$. Then

$$\mathrm{var}(\widetilde{X}_g) \sim n_g^3$$

and

$$\mathrm{sd}\left(\overline{X}_n\right) \sim n^{\alpha-1/2}.$$

Thus $\mathrm{sd}\left(\overline{X}_n\right)$ converges at a rate slower than both $n^{-1/2}$ and $G^{-1/2}$.

Example 5. The clusters are of two sizes, $n_g = 1$ and $n_g = n^\alpha$. There are $G_1 = n/2$ of the first type and $G_2 = n^{1-\alpha}/2$ of the second type. (So $G = G_1 + G_2 = O(n)$.) Within each cluster the observations are identical and have unit variances. var$(\widetilde{X}_g)$ for the two types of clusters are 1 and $n^{2\alpha}$, respectively. Then

$$\text{sd}\left(\overline{X}_n\right) = \left(\frac{G_1 + G_2 n^{2\alpha}}{n^2}\right)^{1/2} = \left(\frac{1 + n^\alpha}{2n}\right)^{1/2} = O\left(n^{-(1-\alpha)/2}\right).$$

Thus sd $\left(\overline{X}_n\right)$ converges at a rate slower than both $n^{-1/2}$ and $G^{-1/2}$.

The final example illustrates the importance of considering heterogeneous cluster sizes. The reason why the convergence rate is slower than both $n^{-1/2}$ and $G^{-1/2}$ is because the number of clusters is determined by the large number of small clusters, but the convergence rate is determined by the (relatively) small number of large clusters.

What we have seen is that the convergence rate sd $\left(\overline{X}_n\right)$ can equal the square root of sample size $n^{-1/2}$, can equal the square root of the number of groups $G^{-1/2}$, can be in between $G^{-1/2}$ and $n^{-1/2}$, or can be slower than both $n^{-1/2}$ and $G^{-1/2}$.

When $\overline{X}_n$ is a vector, it is likely that its elements converge at different rates since they can have different within-cluster correlation structures. For example, some variables could be independent within clusters while others could be identical within clusters.

These examples show that under cluster dependence the convergence rate is context-dependent and variable-dependent, and it is therefore important to allow for general rates of convergence and to not impose arbitrary rates in asymptotic analysis.

5. Central limit theory

Under i.i.d. sampling the standard deviation of the sample mean is of order $O(n^{-1/2})$, so $\sqrt{n}$ is the appropriate scaling to obtain the central limit theorem (CLT). As discussed in the previous section, clustering can alter the rate of convergence, so it is essential to standardize the sample mean by the actual variance rather than an assumed rate. The variance matrix of $\sqrt{n}\overline{X}_n$ is

$$\begin{aligned}\Omega_n &= E\left(n\left(\overline{X}_n - E\overline{X}_n\right)\left(\overline{X}_n - E\overline{X}_n\right)'\right)\\ &= \frac{1}{n}\sum_{g=1}^{G} E\left(\left(\widetilde{X}_g - E\widetilde{X}_g\right)\left(\widetilde{X}_g - E\widetilde{X}_g\right)'\right).\end{aligned}$$

We use the following regularity condition.

Assumption 2. For some $2 \le r < \infty$

$$\frac{\left(\sum_{g=1}^{G} n_g^r\right)^{2/r}}{n} \le C < \infty, \tag{5}$$

$$\max_{g \le G} \frac{n_g^2}{n} \to 0, \tag{6}$$

as $n \to \infty$.

Theorem 2 (*CLT*). *If for some $2 \le r < \infty$ Assumption 2 holds,*

$$\lim_{M\to\infty} \sup_i \left(E\left\|X_i\right\|^r 1\left(\left\|X_i\right\| > M\right)\right) = 0, \tag{7}$$

and

$$\lambda_n = \lambda_{\min}\left(\Omega_n\right) \ge \lambda > 0, \tag{8}$$

then as $n \to \infty$

$$\Omega_n^{-1/2}\sqrt{n}\left(\overline{X}_n - E\overline{X}_n\right) \xrightarrow{d} N\left(\mathbf{0}, I_p\right). \tag{9}$$

Theorem 2 provides a CLT for cluster samples which generalizes the classic CLT for independent heterogeneous samples. The latter holds with $r = 2$, $n_g = 1$ and $G = n$.

Assumption 2 and (7) are stronger than Assumption 1 and (2), and thus the conditions for the CLT imply those for the WLLN.

The condition (7) states that $\left\|X_i\right\|^r$ is uniformly integrable. When $r = 2$ this is similar to the Lindeberg condition for the CLT under independent heterogeneous sampling. (7) simplifies to $E\left\|X_i\right\|^r < \infty$ when the observations have identical marginal distributions. A sufficient condition allowing for distributional heterogeneity is $\sup_i E\left\|X_i\right\|^s < \infty$ for some $s > r \ge 2$.

Assumption 2 (5) is a restriction on the cluster sizes. It involves a trade-off with the number of moments r. It is least restrictive for large r, and more restrictive for small r. As $r \to \infty$ it approaches $\max_{g \le G} n_g^2/n = O(1)$, which is implied by Assumption 2 (6).

Assumption 2 allows for growing and heterogeneous cluster sizes. For example, it allows clusters to grow uniformly at the rate $n_g = n^\alpha$ for $0 \le \alpha \le (r-2)/2(r-1)$. (Note that this requires the cluster sizes to be bounded if $r = 2$.) It also allows for only a small number of clusters to grow. For example, suppose that $n_g = \overline{n}$ (bounded) for $G - K$ clusters and $n_g = G^{\alpha/2}$ for K clusters, with K fixed. Then Assumption 2 holds for any $\alpha < 1$ and $r \ge 2$.

Assumption 2 (5) is implied by

$$\max_{g \le G} \frac{n_g}{n^{(r-2)/2(r-1)}} \le C \tag{10}$$

and they are equivalent when the cluster sizes are homogeneous. In general, however, (5) is less restrictive than (10). For example, when $r = 2$, (10) requires the cluster sizes to be bounded, while (5) does not. (Consider the heterogeneous example given in the previous paragraph. This satisfies (5) but not (10) when $r = 2$.)

The condition (8) specifies that var $\left(\sqrt{n}\alpha'\overline{X}_n\right)$ does not vanish for any conformable vector $\alpha \neq 0$. This excludes degenerate cases and perfect negative within-cluster correlation. In general, if X_i is non-degenerate then (8) is not restrictive as there is no reasonable setting where it will be violated. If $\overline{X}_n$ converges at rate $n^{-1/2}$ then $\lambda_n = O(1)$ but when $\overline{X}_n$ converges at rate slower than $n^{-1/2}$ then λ_n will actually diverge with n. It should also be mentioned that condition (8) allows the components of Ω_n to converge at different rates.

Our proof of Theorem 2 actually uses the conditions

$$\frac{\left(\sum_{g=1}^G n_g^r\right)^{2/r}}{n\lambda_n} \le C < \infty \tag{11}$$

and

$$\max_{g \le G} \frac{n_g^2}{n\lambda_n} \to 0 \tag{12}$$

instead of (5)–(8). (11)–(12) is weaker than (5)–(8) when λ_n diverges to infinity (which occurs when $\overline{X}_n$ converges at a rate slower than $n^{-1/2}$). Since the sequence λ_n is unknown in an application it is difficult to interpret the assumptions (11)–(12). Hence we prefer the assumptions (5)–(8).

The conditions (11)–(12) may be stronger than necessary when within-cluster dependence is weak, but are necessary under strong within-cluster dependence. To see this, suppose that all observations within a cluster are identical, so that $X_{gj} = Z_g$ and Z_g has a finite variance but no higher moments. Then the Lindeberg condition for the CLT can be simplified to

$$\sum_{g=1}^G \frac{n_g^2}{n\lambda_n} E\left(\left\|Z_g\right\|^2 1\left(\left\|Z_g\right\|^2 \ge \frac{n\lambda_n\varepsilon}{n_g^2}\right)\right) \to 0$$

for all $\varepsilon > 0$. Each term in the sum must limit to zero, which requires (11)–(12) with $r = 2$.

We now compare our conditions with those of Djogbenou et al. (2018). Their Assumption 3 states (in our notation) for $r \ge 4$

$$\max_{g \le G} \frac{n_g}{n^{(r-2)/2(r-1)}\lambda_n^{r/2(r-1)}} = o(1). \tag{13}$$

Eq. (13) implies and is stronger than (11). Calculations similar to those in our appendix show that $\lambda_n \le O\left(\max_g n_g\right) = O(n)$. So (13) also implies

$$\left(\max_{g \le G} \frac{n_g^2}{n\lambda_n}\right)^{1/2} = \max_{g \le G} \frac{n_g}{n^{(r-2)/2(r-1)}\lambda_n^{r/2(r-1)}} \left(\frac{\lambda_n}{n}\right)^{1/2(r-1)} = o\left(1\right)$$

which is (12). Thus our conditions (11)–(12) are less restrictive than their condition (13), and do not require $r \ge 4$.

6. Cluster-robust variance matrix estimation

We now discuss cluster-robust covariance matrix estimation.

We first consider the case where X_i is mean zero (or equivalently that the mean is known). In this case the covariance matrix equals

$$\Omega_n = \frac{1}{n}\sum_{g=1}^G E\left(\widetilde{X}_g\widetilde{X}_g'\right).$$

In this case a natural estimator is

$$\widetilde{\Omega}_n = \frac{1}{n}\sum_{g=1}^{G} \widetilde{X}_g \widetilde{X}_g'.$$

Theorem 3. *Under the assumptions of Theorem 2, if in addition $EX_i = 0$ then as $n \to \infty$*

$$\Omega_n^{-1/2}\widetilde{\Omega}_n\Omega_n^{-1/2} \xrightarrow{p} I_p \tag{14}$$

and

$$\widetilde{\Omega}_n^{-1/2}\sqrt{n}X_n \xrightarrow{d} N\left(\mathbf{0}, I_p\right). \tag{15}$$

Theorem 3 shows that the cluster-robust covariance matrix estimator is consistent, and replacing the covariance matrix in the CLT with the estimated covariance matrix does not affect the asymptotic distribution. Implications of (15) are that cluster-robust t-ratios are asymptotically standard normal, and that cluster-robust Wald statistics are asymptotically chi-square distributed with p degrees of freedom.

Construction of practical covariance matrix estimators is context-specific, depending on the mean structure. For example, suppose that $\mu = EX_i$ does not vary across observations. In this case we can write

$$\Omega_n = \frac{1}{n}\sum_{g=1}^{G} E\left(\widetilde{X}_g \widetilde{X}_g'\right) - \frac{1}{n}\sum_{g=1}^{G} n_g^2 \mu\mu'.$$

The natural estimator for μ is $\overline{X}_n$ and that for Ω_n is

$$\widehat{\Omega}_n = \frac{1}{n}\sum_{g=1}^{G} \widetilde{X}_g \widetilde{X}_g' - \frac{1}{n}\sum_{g=1}^{G} n_g^2 \overline{X}_n \overline{X}_n'.$$

Theorem 4. *Under the assumptions of Theorem 2, if in addition $\mu = EX_i$ does not vary across observations, then as $n \to \infty$*

$$\Omega_n^{-1/2}\widehat{\Omega}_n\Omega_n^{-1/2} \xrightarrow{p} I_p \tag{16}$$

and

$$\widehat{\Omega}_n^{-1/2}\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{d} N\left(\mathbf{0}, I_p\right). \tag{17}$$

7. Uniform laws of large numbers

Now consider a uniform WLLN. Consider functions $f(x,\theta) \in \mathbb{R}^k$ indexed on $\theta \in \Theta$ where Θ is compact. Define the sample mean

$$\overline{f}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} f(X_i,\theta).$$

The following result is an application of Theorem 3 of Andrews (1992).

Theorem 5 (*ULLN for Clustered Means*)*. Suppose that Assumption 1 holds and for each $\theta \in \Theta$*

$$\lim_{M\to\infty} \sup_i \left(E\left\|f(X_i,\theta)\right\| 1\left(\left\|f(X_i,\theta)\right\| > M\right)\right) = 0. \tag{18}$$

Suppose as well that for each $\theta_1, \theta_2 \in \Theta$

$$\left\|f(x,\theta_1) - f(x,\theta_2)\right\| \leq A(x)h\left(\left\|\theta_1 - \theta_2\right\|\right) \tag{19}$$

where $h(u) \downarrow 0$ as $u \downarrow 0$ and $\sup_i EA(X_i) \leq C$. Then $E\overline{f}_n(\theta)$ is continuous in θ uniformly over $\theta \in \Theta$ and $n \geq 1$, and as $n \to \infty$

$$\sup_{\theta\in\Theta}\left\|\overline{f}_n(\theta) - E\overline{f}_n(\theta)\right\| \xrightarrow{p} 0. \tag{20}$$

We also consider a uniform law for the clustered variance. Set $\mu(\theta) = Ef(X_i,\theta)$ so that it does not vary across observations. The variance of $\sqrt{n}\overline{f}_n(\theta)$ is

$$\begin{aligned}\Omega_n(\theta) &= E\left(n\left(\overline{f}_n(\theta) - E\overline{f}_n(\theta)\right)\left(\overline{f}_n(\theta) - E\overline{f}_n(\theta)\right)'\right)\\ &= \frac{1}{n}\sum_{g=1}^{G} E\widetilde{f}_g(\theta)\widetilde{f}_g(\theta) - \frac{1}{n}\sum_{g=1}^{G} n_g^2\mu(\theta)\mu(\theta)'\end{aligned}$$

where $\widetilde{f}_g(\theta) = \sum_{j=1}^{n_g} f(X_{gj}, \theta)$ are the cluster sums. An appropriate estimator for $\Omega_n(\theta)$ is

$$\widehat{\Omega}_n(\theta) = \frac{1}{n}\sum_{g=1}^{G} \widetilde{f}_g(\theta)\widetilde{f}_g(\theta) - \frac{1}{n}\sum_{g=1}^{G} n_g^2 \overline{f}_n(\theta)\overline{f}_n(\theta)'.$$

In practice, a simpler estimator

$$\widetilde{\Omega}_n(\theta) = \frac{1}{n}\sum_{g=1}^{G} \widetilde{f}_g(\theta)\widetilde{f}_g(\theta)'$$

is often used if $\mu(\theta_0) = 0$ for $\theta_0 \in$ *interior* (Θ) and $\widehat{\theta} \xrightarrow{p} \theta_0$ for some estimator $\widehat{\theta}$.

The following result is an extension of Theorem 5 to the case of clustered variance estimators. It also relies on Theorem 3 of Andrews (1992).

Theorem 6 (*ULLN for Clustered Variance*).. *Suppose that Assumption 2 holds with $r = 2$, $\mu(\theta) = Ef(X_i, \theta)$ does not vary across i, for each $\theta \in \Theta$,*

$$\lim_{M\to\infty} \sup_i \left(E\left\|f(X_i,\theta)\right\|^2 1\left(\left\|f(X_i,\theta)\right\| > M\right)\right) = 0, \tag{21}$$

and for each $\theta_1, \theta_2 \in \Theta$ (19) holds with $\sup_i EA(X_i)^2 \leq C$. Then as $n \to \infty$

$$\sup_{\theta\in\Theta} \left\|\widehat{\Omega}_n(\theta) - \Omega_n(\theta)\right\| \xrightarrow{p} 0. \tag{22}$$

If $\mu(\theta) = 0$, then as $n \to \infty$

$$\sup_{\theta\in\Theta} \left\|\widetilde{\Omega}_n(\theta) - \Omega_n(\theta)\right\| \xrightarrow{p} 0. \tag{23}$$

8. Central limit theorem for clustered second moments

Although our primary focus is the sample mean, the core theory can be extended to statistics which are not sample means. In this section, we focus on the vectorized variance estimators

$$\overline{f}_G = \frac{1}{n}\sum_{g=1}^{G} \widetilde{f}_g$$

where

$$\widetilde{f}_g = \widetilde{X}_g \otimes \widetilde{X}_g$$

or

$$\widetilde{f}_g = \left(\widetilde{X}_g - n_g\overline{X}_n\right) \otimes \left(\widetilde{X}_g - n_g\overline{X}_n\right).$$

The WLLN for $\overline{f}_G$ holds by Theorem 3 (14) and Theorem 4 (16), and the ULLN for $\overline{f}_G$ holds by Theorem 6. However, the CLT given in Theorem 2 cannot be applied to $\overline{f}_G$ because $\overline{f}_G$ cannot be written as the sample mean over *i*. We provide the CLT for $\overline{f}_G$ below. This is useful to establish asymptotic distributions of estimators in a non-standard setting. For example, the asymptotic distribution of the generalized method of moments (GMM) estimators depends on the limiting distribution of the weight matrix when the moment condition is misspecified (Hall and Inoue, 2003; Lee, 2014; Hansen and Lee, 2018).

Similar to the sample mean, the convergence rate of $\overline{f}_G$ can vary under cluster dependence. Consider $\widetilde{f}_g = \widetilde{X}_g \otimes \widetilde{X}_g$ and assume $p = 1$ for simplicity. The standard deviation of $\overline{f}_G$ is

$$\operatorname{sd}\left(\overline{f}_G\right) = \frac{1}{n}\left(\sum_{g=1}^{G}\operatorname{var}\left(\widetilde{X}_g\widetilde{X}_g\right)\right)^{1/2} = \frac{1}{n}\left(\sum_{g=1}^{G}\sum_{j=1}^{n_g}\sum_{l=1}^{n_g}\operatorname{var}\left(X_{gj}X_{gl}\right)\right)^{1/2}.$$

Under i.i.d. sampling $\operatorname{sd}\left(\overline{f}_G\right) = O\left(n^{-1/2}\right)$. Under the Examples 1 and 2 in Section 4, the convergence rate is $G^{-1/2}$.

Define the variance matrix of $\sqrt{n}\overline{f}_G$ as

$$\begin{aligned}\Omega_n &= E\left(n\left(\overline{f}_G - E\overline{f}_G\right)\left(\overline{f}_G - E\overline{f}_G\right)'\right) \\ &= \frac{1}{n}\sum_{g=1}^{G} E\left(\left(\widetilde{f}_g - E\widetilde{f}_g\right)\left(\widetilde{f}_g - E\widetilde{f}_g\right)'\right).\end{aligned}$$

We use the following regularity condition.

Assumption 3. For some $2 \le r < \infty$

$$\frac{\left(\sum_{g=1}^{G} n_g^{2r}\right)^{2/r}}{n} \le C < \infty, \tag{24}$$

$$\max_{g \le G} \frac{n_g^4}{n} \to 0, \tag{25}$$

as $n \to \infty$.

Note that Assumption 3 is a strengthening of Assumption 2.

Theorem 7 (*CLT for Clustered Variance*)**.** *For some $2 \le r < \infty$ Assumption 3 holds,*

$$\lim_{M\to\infty} \sup_i \left(E\left\|X_i\right\|^{2r} 1\left(\left\|X_i\right\| > M\right)\right) = 0, \tag{26}$$

and

$$\lambda_n = \lambda_{\min}\left(\Omega_n\right) \ge \lambda > 0 \tag{27}$$

then as $n \to \infty$

$$\Omega_n^{-1/2}\sqrt{n}\left(\overline{f}_G - E\overline{f}_G\right) \xrightarrow{d} N\left(\mathbf{0}, I_q\right) \tag{28}$$

where $q = p^2$.

Finally we provide a CLT combining the previous results. For $Y_i \in \mathbb{R}^s$, $i = 1, \ldots, n$, obtained by cluster sampling, let $\widetilde{\psi}_g$ be the stacked vector

$$\widetilde{\psi}_g = \begin{pmatrix} \widetilde{Y}_g \\ \widetilde{X}_g \\ \widetilde{X}_g \otimes \widetilde{X}_g \end{pmatrix}$$

or

$$\widetilde{\psi}_g = \begin{pmatrix} \widetilde{Y}_g \\ \widetilde{X}_g \\ \left(\widetilde{X}_g - n_g\overline{X}_n\right) \otimes \left(\widetilde{X}_g - n_g\overline{X}_n\right) \end{pmatrix}$$

and $\overline{\psi}_G = n^{-1}\sum_{g=1}^{G}\widetilde{\psi}_g$. Let the variance matrix of $\sqrt{n}\overline{\psi}_G$ be

$$\Omega_n = E\left(n\left(\overline{\psi}_G - E\overline{\psi}_G\right)\left(\overline{\psi}_G - E\overline{\psi}_G\right)'\right).$$

The following Corollary provides the CLT for the joint process. Since it immediately follows from Theorems 2 and 7, the proof is omitted.

Corollary 1. *If for some $2 \le r < \infty$ Assumption 3 holds,*

$$\lim_{M\to\infty} \sup_i \left(E\left\|Y_i\right\|^{r} 1\left(\left\|Y_i\right\| > M\right)\right) = 0,$$

$$\lim_{M\to\infty} \sup_i \left(E\left\|X_i\right\|^{2r} 1\left(\left\|X_i\right\| > M\right)\right) = 0,$$

and

$$\lambda_{\min}\left(\Omega_n\right) \ge \lambda > 0,$$

then as $n \to \infty$

$$\Omega_n^{-1/2}\sqrt{n}\left(\overline{\psi}_G - E\overline{\psi}_G\right) \xrightarrow{d} N\left(\mathbf{0}, I_q\right)$$

where $q = s + p + p^2$.

9. Linear regression and two-stage least squares

It is useful to use cluster-level notation. Let $\boldsymbol{y}_g = (y_{g1}, \ldots, y_{gn_g})'$, $\boldsymbol{X}_g = (\boldsymbol{x}_{g1}, \ldots, \boldsymbol{x}_{gn_g})'$ and $\boldsymbol{Z}_g = (\boldsymbol{z}_{g1}, \ldots, \boldsymbol{z}_{gn_g})'$ denote an $n_g \times 1$ vector of dependent variables, $n_g \times k$ matrix of regressors, and $n_g \times l$ matrix of instruments for the gth cluster. A linear model can be written using cluster notation as

$$\boldsymbol{y}_g = \boldsymbol{X}_g\boldsymbol{\beta} + \boldsymbol{e}_g, \tag{29}$$

$$\begin{aligned} \boldsymbol{X}_g &= \boldsymbol{Z}_g \boldsymbol{\gamma} + \boldsymbol{u}_g, \\ E\left(\boldsymbol{Z}_g' \boldsymbol{e}_g\right) &= 0 \end{aligned} \tag{30}$$

where $\boldsymbol{e}_g$ is a $n_g \times 1$ error vector. The case of linear regression holds as the special case where $\boldsymbol{Z}_g = \boldsymbol{X}_g$ and $l = k$ (so that (30) becomes identity). Assume $l \geq k$. (29) is the structural equation and (30) is the first-stage equation.

The two-stage least squares (2SLS) estimator for $\boldsymbol{\beta}$ can be written as

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{Z}_g \left(\sum_{g=1}^{G} \boldsymbol{Z}_g' \boldsymbol{Z}_g \right)^{-1} \sum_{g=1}^{G} \boldsymbol{Z}_g' \boldsymbol{X}_g \right)^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{Z}_g \left(\sum_{g=1}^{G} \boldsymbol{Z}_g' \boldsymbol{Z}_g \right)^{-1} \sum_{g=1}^{G} \boldsymbol{Z}_g' \boldsymbol{y}_g \right).$$

We first show consistency of $\widehat{\boldsymbol{\beta}}$. Define

$$\begin{aligned} Q_n &= \frac{1}{n} \sum_{g=1}^{G} E\left(\boldsymbol{Z}_g' \boldsymbol{X}_g\right), \\ W_n &= \frac{1}{n} \sum_{g=1}^{G} E\left(\boldsymbol{Z}_g' \boldsymbol{Z}_g\right). \end{aligned}$$

Theorem 8. *If Assumption 1 holds, Q_n has full rank k, $\lambda_{\min}(W_n) \geq C > 0$, and either*

1. *$(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ have identical marginal distributions with finite second moments;*
   or
2. *For some $r > 2$, $\sup_i E\,|y_i|^r < \infty$, $\sup_i E\,\|\boldsymbol{x}_i\|^r < \infty$, and $\sup_i E\,\|\boldsymbol{z}_i\|^r < \infty$;*
   then as $n \to \infty$, $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$.

Next we provide the asymptotic distribution. Define

$$\begin{aligned} \Omega_n &= \frac{1}{n} \sum_{g=1}^{G} E\left(\boldsymbol{Z}_g' \boldsymbol{e}_g \boldsymbol{e}_g' \boldsymbol{Z}_g\right), \\ V_n &= \left(Q_n' W_n^{-1} Q_n\right)^{-1} Q_n' W_n^{-1} \Omega_n W_n^{-1} Q_n \left(Q_n' W_n^{-1} Q_n\right)^{-1}. \end{aligned}$$

The residuals for the gth cluster are

$$\widehat{\boldsymbol{e}}_g = \boldsymbol{y}_g - \boldsymbol{X}_g \widehat{\boldsymbol{\beta}}.$$

Define

$$\begin{aligned} \widehat{\Omega}_n &= \frac{1}{n} \sum_{g=1}^{G} \boldsymbol{Z}_g' \widehat{\boldsymbol{e}}_g \widehat{\boldsymbol{e}}_g' \boldsymbol{Z}_g, \\ \widehat{Q}_n &= \frac{1}{n} \sum_{g=1}^{G} \boldsymbol{Z}_g' \boldsymbol{X}_g, \\ \widehat{W}_n &= \frac{1}{n} \sum_{g=1}^{G} \boldsymbol{Z}_g' \boldsymbol{Z}_g. \end{aligned}$$

The variance estimator is

$$\widehat{V}_n = d_n \left(\widehat{Q}_n' \widehat{W}_n^{-1} \widehat{Q}_n\right)^{-1} \widehat{Q}_n' \widehat{W}_n^{-1} \widehat{\Omega}_n \widehat{W}_n^{-1} \widehat{Q}_n \left(\widehat{Q}_n' \widehat{W}_n^{-1} \widehat{Q}_n\right)^{-1}.$$

with d_n a possible finite-sample degree-of-freedom adjustment. For example, Hansen (2007) proposed $d_n = G/(G-1)$ for the regression case (under homogeneous cluster sizes), and Stata sets

$$d_n = \left(\frac{n-1}{n-k}\right)\left(\frac{G}{G-1}\right)$$

for the OLS and 2SLS estimators under *cluster* option.

Theorem 9. *Suppose that Assumption 2 holds for some $2 \leq r \leq s < \infty$, Q_n has full rank k, $\lambda_{\min}(W_n) \geq C > 0$, $\lambda_{\min}(\Omega_n) \geq \lambda > 0$, $\sup_i E\,|y_i|^{2s} < \infty$, $\sup_i E\,\|\boldsymbol{x}_i\|^{2s} < \infty$, and $\sup_i E\,\|\boldsymbol{z}_i\|^{2s} < \infty$, and either*

1. *$(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ have identical marginal distributions; or*
2. *$r < s$;*

then, for any sequence of full-rank $k \times q$ matrices R_n, as $n \to \infty$

$$\left(R_n' V_n R_n\right)^{-1/2} R_n' \sqrt{n} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} N\left(\mathbf{0}, I_q\right), \tag{31}$$

$$\left(R_n' V_n R_n\right)^{-1/2} R_n' \widehat{V}_n R_n \left(R_n' V_n R_n\right)^{-1/2} \xrightarrow{p} I_q, \tag{32}$$

and

$$\left(R_n' \widehat{V}_n R_n\right)^{-1/2} R_n' \sqrt{n} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} N\left(\mathbf{0}, I_q\right). \tag{33}$$

The standard errors for $R_n'\widehat{\boldsymbol{\beta}}$ can be obtained by taking the square roots of the diagonal elements of $n^{-1}R_n'\widehat{V}_nR_n$.

10. (Pseudo) maximum likelihood

Suppose that we observe a sequence of random vectors $X_i \in \mathbb{R}^p$, $i = 1, \ldots, n$ with the same marginal distributions from a density $f(x, \boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \mathbb{R}^k$. Let $\boldsymbol{X}_g = (X_{g1}, \ldots, X_{gn_g})'$ be a $n_g \times p$ matrix for each cluster. For the observations in the cluster g, let $f_g(\boldsymbol{X}_g, \boldsymbol{\theta}_0)$ be the joint density. Since the observations within the same cluster need not be independent, $f_g(\boldsymbol{X}_g, \boldsymbol{\theta}_0) \neq \prod_{i=1}^{n_g} f(X_{gi}, \boldsymbol{\theta}_0)$ in general. This also implies that $f_g(\boldsymbol{X}_g, \boldsymbol{\theta}_0) \neq f_h(\boldsymbol{X}_h, \boldsymbol{\theta}_0)$ for $g \neq h$. Given specification of $f_g(\boldsymbol{X}_g, \boldsymbol{\theta}_0)$, the maximum likelihood estimator (MLE) can be obtained as the maximizer of

$$\sum_{g=1}^{G} \log f_g(\boldsymbol{X}_g, \boldsymbol{\theta}).$$

However, the joint density $f_g(\boldsymbol{X}_g, \boldsymbol{\theta})$ may be difficult to specify in practice. A simpler alternative is to use a pseudo-likelihood $\prod_{i=1}^{n_g} f(X_{gi}, \boldsymbol{\theta}_0)$ for the joint density $f_g(\boldsymbol{X}_g, \boldsymbol{\theta}_0)$, and specify the log likelihood function as

$$L_n(\theta) = \sum_{g=1}^{G} \sum_{j=1}^{n_g} \log f(X_{gj}, \boldsymbol{\theta}).$$

Define the pseudo-MLE as

$$\widehat{\boldsymbol{\theta}} = \underset{\theta \in \Theta}{\operatorname{argmax}} \, L_n(\boldsymbol{\theta}).$$

This estimator is also called the partial (or pooled) MLE (Wooldridge, 2010).

This estimator is the standard implementation of MLE under clustered dependence. To our knowledge there is no existing distribution theory for this standard estimator.

We first show consistency of $\widehat{\boldsymbol{\theta}}$. The following is based on Theorem 2.1 of Newey and McFadden (1994).

Theorem 10. *If Assumption 1 holds,*

1. *X_i have identical marginal distributions with the density $f(x, \boldsymbol{\theta}_0)$ and $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, which is compact,*
2. *if $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ then $f(x, \boldsymbol{\theta}) \neq f(x, \boldsymbol{\theta}_0)$,*
3. *$E[\sup_{\boldsymbol{\theta} \in \Theta} |\log f(X_i, \boldsymbol{\theta})|] < \infty$,*
4. *for each $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$,*

$$\|\log f(x, \boldsymbol{\theta}_1) - \log f(x, \boldsymbol{\theta}_2)\| \leq A(x) h\left(\|\theta_1 - \theta_2\|\right)$$

where $h(u) \downarrow 0$ as $u \downarrow 0$ and $EA(X_i) \leq C$,

Then as $n \to \infty$, $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$.

Next we show the asymptotic distribution. Define

$$H_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} E\left[\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \log f(X_i, \boldsymbol{\theta})\right],$$

$$\Omega_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{g=1}^{G} E\left(\sum_{j=1}^{n_g} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(X_{gj}, \boldsymbol{\theta})\right)\left(\sum_{j=1}^{n_g} \frac{\partial}{\partial \boldsymbol{\theta}'} \log f(X_{gj}, \boldsymbol{\theta})\right),$$

$$V_n = H_n(\boldsymbol{\theta}_0)^{-1} \Omega_n(\boldsymbol{\theta}_0) H_n(\boldsymbol{\theta}_0)^{-1}.$$

Define the sample versions

$$\widehat{H}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \log f(X_i, \boldsymbol{\theta}),$$

$$\widehat{\Omega}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{g=1}^{G} \left(\sum_{j=1}^{n_g} \frac{\partial}{\partial\boldsymbol{\theta}} \log f(X_{gj}, \boldsymbol{\theta})\right)\left(\sum_{j=1}^{n_g} \frac{\partial}{\partial\boldsymbol{\theta}'} \log f(X_{gj}, \boldsymbol{\theta})\right).$$

The variance estimator is

$$\widehat{V}_n = \widehat{H}_n(\widehat{\boldsymbol{\theta}})^{-1}\widehat{\Omega}_n(\widehat{\boldsymbol{\theta}})\widehat{H}_n(\widehat{\boldsymbol{\theta}})^{-1}.$$

Note that the information matrix equality does not hold because $\sum_{j=1}^{n_g} \log f(X_{gj}, \boldsymbol{\theta}_0) \neq f_g(\boldsymbol{X}_g, \boldsymbol{\theta}_0)$ in general.

Theorem 11. *In addition to the assumptions of Theorem 10, Assumption 2 holds with* $r = 2$,

1. $\boldsymbol{\theta}_0 \in$ *interior*$(\boldsymbol{\Theta})$,
2. *for some neighborhood* $\mathcal{N}$ *of* $\boldsymbol{\theta}_0$,
   - (a) $f(x, \boldsymbol{\theta})$ *is twice continuously differentiable and* $f(x, \boldsymbol{\theta}) > 0$,
   - (b) $\int \sup_{\boldsymbol{\theta}\in\mathcal{N}} \left\| \frac{\partial}{\partial\boldsymbol{\theta}} \log f(x, \boldsymbol{\theta})\right\| dx < \infty$,
   - (c) $E\left\| \frac{\partial}{\partial\boldsymbol{\theta}} \log f(X_i, \boldsymbol{\theta})\right\|^2 < \infty$,
   - (d) $E \sup_{\boldsymbol{\theta}\in\mathcal{N}} \left\| \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \log f(X_i, \boldsymbol{\theta})\right\|^2 < \infty$,
   - (e) *and for each* $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{N}$,

     $$\left\| \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \log f(x, \boldsymbol{\theta}_1) - \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \log f(x, \boldsymbol{\theta}_2)\right\| \leq A(x)h\left(\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|\right)$$

     where $h(u) \downarrow 0$ *as* $u \downarrow 0$ *and* $EA(X_i) \leq C$,
3. $\lambda_{\min}(H_n(\boldsymbol{\theta}_0)) \geq C > 0$,
4. $\lambda_{\min}(\Omega_n(\boldsymbol{\theta}_0)) \geq \lambda > 0$,

then for any sequence of full-rank $k \times q$ *matrices* R_n, *as* $n \to \infty$

$$\left(R_n' V_n R_n\right)^{-1/2} R_n' \sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(\mathbf{0}, I_q\right), \tag{34}$$

$$\left(R_n' V_n R_n\right)^{-1/2} R_n' \widehat{V}_n R_n \left(R_n' V_n R_n\right)^{-1/2} \xrightarrow{p} I_q, \tag{35}$$

and

$$\left(R_n' \widehat{V}_n R_n\right)^{-1/2} R_n' \sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(\mathbf{0}, I_q\right). \tag{36}$$

The standard errors for $R_n'\widehat{\boldsymbol{\beta}}$ can be obtained by taking the square roots of the diagonal elements of $n^{-1}R_n'\widehat{V}_n R_n$.

11. Generalized method of moments

Suppose that we observe a sequence of random vectors $X_i \in \mathbb{R}^p$, $i = 1, \ldots, n$ from cluster sampling. A known moment function is given by $m(X_i, \boldsymbol{\theta})$ where $m(\cdot, \cdot)$ is $l \times 1$ and $\boldsymbol{\theta}$ is $k \times 1$. Define the cluster sum as

$$\widetilde{m}_g(\boldsymbol{\theta}) = \sum_{j=1}^{n_g} m(X_{gj}, \boldsymbol{\theta}).$$

An unconditional moment model in cluster notation is given by

$$E\widetilde{m}_g(\boldsymbol{\theta}_0) = 0. \tag{37}$$

We assume that $\boldsymbol{\theta}_0$ is identified and $l > k$ so the moment model is over-identified. Write the sample mean of the moment function as

$$\overline{m}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} m(X_i, \boldsymbol{\theta}).$$

Since (37) holds for all $g = 1, \ldots, G$, the usual unconditional moment condition $E\overline{m}_n(\boldsymbol{\theta}_0) = 0$ follows. The generalized method of moments (GMM) estimator is given by

$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} n \cdot \overline{m}_n(\boldsymbol{\theta})' \widehat{W}_n^{-1} \overline{m}_n(\boldsymbol{\theta}) \tag{38}$$

where $\widehat{W}_n^{-1}$ is an $l \times l$ positive definite weight matrix, which may or may not depend on an estimated parameter. Typically, the weight matrix is obtained by plugging in a preliminary consistent estimator, $\widetilde{\boldsymbol{\theta}}$, so that $\widehat{W}_n^{-1} = \widehat{W}_n(\widetilde{\boldsymbol{\theta}})^{-1}$.

We consider two forms of GMM estimator. The first one is based on a non-clustered weight matrix, which takes the form of

$$\widehat{W}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} v(X_i, \boldsymbol{\theta}) v(X_i, \boldsymbol{\theta})' \tag{39}$$

for some $l \times 1$ vector $v(x, \boldsymbol{\theta})$. This includes the conventional one-step and two-step GMM estimators. For 2SLS, $v(X_i, \boldsymbol{\theta}) = Z_i$ where Z_i is an $l \times 1$ vector of instruments. The efficient two-step GMM uses $v(X_i, \boldsymbol{\theta}) = m(X_i, \boldsymbol{\theta})$ or $v(X_i, \boldsymbol{\theta}) = m(X_i, \boldsymbol{\theta}) - \overline{m}_n(\boldsymbol{\theta})$. The conventional efficient weight matrix, however, does not provide efficiency anymore under cluster sampling because a weight matrix of the form of (39) is not consistent for the variance matrix of $\sqrt{n}(\overline{m}_n(\boldsymbol{\theta}) - E\overline{m}_n(\boldsymbol{\theta}))$ in general.

The second is based on the clustered efficient weight matrix, which leads to the two-step efficient GMM under cluster sampling. The weight matrix takes the form of

$$\widehat{W}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{g=1}^{G} \widetilde{m}_g(\boldsymbol{\theta}) \widetilde{m}_g(\boldsymbol{\theta})' - \frac{1}{n} \sum_{g=1}^{G} n_g^2 \overline{m}_n(\boldsymbol{\theta}) \overline{m}_n(\boldsymbol{\theta})'. \tag{40}$$

Alternatively, the uncentered version of $\widehat{W}_n(\boldsymbol{\theta})$ and $\widehat{\Omega}_n(\boldsymbol{\theta})$ can be used to obtain the efficient two-step GMM estimator but the centered version is generally recommended. For more discussion, see Hansen (2018).

Since we assume that the weight matrix depends on a consistent preliminary estimator, we exclude the continuously updating (CU) GMM estimator in our analysis. Whenever possible, we omit the dependence of the weight matrices on $\widetilde{\boldsymbol{\theta}}$ and write $\widehat{W}_n = \widehat{W}_n(\widetilde{\boldsymbol{\theta}})$. Define $W_n = E\widehat{W}_n(\boldsymbol{\theta}_0)$.

We first show consistency of the GMM estimator. The following is based on Theorem 2.1 of Newey and McFadden (1994).

Theorem 12. *If Assumption 1 holds,*

1. *Θ is compact,*
2. *$\boldsymbol{\theta}_0$ is the unique solution to $E\overline{m}_n(\boldsymbol{\theta}) = 0$,*
3. *for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, either X_i have identical marginal distributions with $E\left\|m(X_i, \boldsymbol{\theta})\right\| < \infty$, or $\sup_i E\left\|m(X_i, \boldsymbol{\theta})\right\|^r < \infty$ for some $r > 1$,*
4. *for each $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$*

   $$\left\|m(x, \boldsymbol{\theta}_1) - m(x, \boldsymbol{\theta}_2)\right\| \leq A(x) h\left(\left\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\right\|\right)$$

   where $h(u) \downarrow 0$ as $u \downarrow 0$ and $EA(X_i) \leq C$,
5. *$\lambda_{\min}(W_n) \geq C > 0$,*
6. *$\widehat{W}_n^{-1} - W_n^{-1} \xrightarrow{p} 0$,*

   then as $n \to \infty$, $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$.

Primitive conditions under which Condition 6 of Theorem 12 holds can be found given the choice of the weight matrix. For simplicity, we assume that if the conventional weight matrix is used then either $v(X_i, \boldsymbol{\theta}) = m(X_i, \boldsymbol{\theta})$ or $v(X_i, \boldsymbol{\theta}) = m(X_i, \boldsymbol{\theta}) - \overline{m}_n(\boldsymbol{\theta})$. If the clustered weight matrix is used then it takes the form of (40). The conditions of Theorem 13 are sufficient for Condition 6 of Theorem 12 to hold.

To show the asymptotic distribution of the GMM estimator, define

$$Q_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} E\left[\frac{\partial}{\partial \boldsymbol{\theta}'} m(X_i, \boldsymbol{\theta})\right],$$

$$\Omega_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{g=1}^{G} E\widetilde{m}_g(\boldsymbol{\theta}) \widetilde{m}_g(\boldsymbol{\theta})',$$

$$V_n = (Q_n' W_n^{-1} Q_n)^{-1} Q_n' W_n^{-1} \Omega_n W_n^{-1} Q_n (Q_n' W_n^{-1} Q_n)^{-1},$$

where $Q_n = Q_n(\boldsymbol{\theta}_0)$ and $\Omega_n = \Omega_n(\boldsymbol{\theta}_0)$. If the clustered efficient weight matrix (40) is used, then the asymptotic variance matrix simplifies to

$$V_n = (Q_n' \Omega_n^{-1} Q_n)^{-1}.$$

Define the sample versions as

$$\widehat{Q}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}'} m(X_i, \boldsymbol{\theta}),$$

$$\widehat{\Omega}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{g=1}^{G} \widetilde{m}_g(\boldsymbol{\theta})\widetilde{m}_g(\boldsymbol{\theta})' - \frac{1}{n}\sum_{g=1}^{G} n_g^2 \overline{m}_n(\boldsymbol{\theta})\overline{m}_n(\boldsymbol{\theta})'$$

and let $\widehat{Q}_n = \widehat{Q}_n(\widehat{\boldsymbol{\theta}})$ and $\widehat{\Omega}_n = \widehat{\Omega}_n(\widehat{\boldsymbol{\theta}})$. The variance estimator is

$$\widehat{V}_n = (\widehat{Q}_n'\widehat{W}_n^{-1}\widehat{Q}_n)^{-1}\widehat{Q}_n'\widehat{W}_n^{-1}\widehat{\Omega}_n\widehat{W}_n^{-1}\widehat{Q}_n(\widehat{Q}_n'\widehat{W}_n^{-1}\widehat{Q}_n)^{-1},$$

if $\widehat{W}_n$ is given by (39) and

$$\widehat{V}_n = (\widehat{Q}_n'\widehat{\Omega}_n^{-1}\widehat{Q}_n)^{-1},$$

if $\widehat{W}_n$ is given by (40), i.e., $\widehat{W}_n = \widehat{\Omega}_n$.

The over-identifying restrictions test (the J test, hereinafter) is a test based on the GMM criterion to test whether the moment model is correctly specified or not, i.e., $E\widetilde{m}_g(\boldsymbol{\theta}_0) = 0$. An implication of cluster sampling is that the conventional J test statistic will not have a standard chi-square asymptotic distribution because the conventional efficient weight matrix is not consistent for the inverse of the variance matrix of the moment function. The GMM criterion (38) based on the clustered efficient weight matrix (40) evaluated at the estimator is the robust J test statistic. Define

$$J_n(\widehat{\boldsymbol{\theta}}) = n \cdot \overline{m}_n(\widehat{\boldsymbol{\theta}})'\widehat{W}_n^{-1}\overline{m}_n(\widehat{\boldsymbol{\theta}}).$$

Theorem 13. *In addition to the assumptions of Theorem 12, if Assumption 2 holds with $r = 2$,*

1. $\boldsymbol{\theta}_0 \in$ *interior*$(\boldsymbol{\Theta})$,
2. *for some neighborhood $\mathcal{N}$ of $\boldsymbol{\theta}_0$,*
   - (a) $m(X_i, \boldsymbol{\theta})$ *is continuously differentiable with probability approaching one,*
   - (b) *either X_i have identical marginal distributions with $E\sup_{\boldsymbol{\theta}\in\mathcal{N}} \|m(X_i, \boldsymbol{\theta})\|^2 < \infty$; or $E\sup_i \sup_{\boldsymbol{\theta}\in\mathcal{N}} \|m(X_i, \boldsymbol{\theta})\|^r < \infty$ for some $r > 2$,*
   - (c) $E\sup_i \sup_{\boldsymbol{\theta}\in\mathcal{N}} \left\| \frac{\partial}{\partial\boldsymbol{\theta}'} m(X_i, \boldsymbol{\theta})\right\|^2 < \infty$
   - (d) *for each $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{N}$*

     $$\left\| \frac{\partial}{\partial \boldsymbol{\theta}} m(x, \boldsymbol{\theta}_1) - \frac{\partial}{\partial \boldsymbol{\theta}} m(x, \boldsymbol{\theta}_2)\right\| \le A(x) h\left(\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|\right)$$

     where $h(u) \downarrow 0$ as $u \downarrow 0$ and $\sup_i EA(X_i) \le C$,
3. $\lambda_{\min}(W_n(\boldsymbol{\theta}_0)) \ge C > 0$,
4. $\lambda_{\min}(\Omega_n(\boldsymbol{\theta}_0)) \ge \lambda > 0$,
5. Q_n *is full column rank,*

then for any sequence of full-rank $k \times q$ matrices R_n, as $n \to \infty$

$$\left(R_n' V_n R_n\right)^{-1/2} R_n' \sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(\mathbf{0}, I_q\right), \tag{41}$$

$$\left(R_n' V_n R_n\right)^{-1/2} R_n' \widehat{V}_n R_n \left(R_n' V_n R_n\right)^{-1/2} \xrightarrow{p} I_q, \tag{42}$$

$$\left(R_n' \widehat{V}_n R_n\right)^{-1/2} R_n' \sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(\mathbf{0}, I_q\right), \tag{43}$$

and

$$J_n(\widehat{\boldsymbol{\theta}}) \xrightarrow{d} \chi^2_{l-k}. \tag{44}$$

The standard errors for $R_n'\widehat{\boldsymbol{\beta}}$ can be obtained by taking the square roots of the diagonal elements of $n^{-1}R_n'\widehat{V}_n R_n$.

Appendix A

We start with a useful technical result which states that if random variables are uniformly integrable then so are their cluster averages, regardless of their joint dependence.

Lemma 1. *For random vectors X_i set $\widetilde{X}_m = \sum_{i=1}^m X_i$. For $r \geq 1$, if*

$$\lim_{B\to\infty} \sup_i E\left(\|X_i\|^r \, 1\left(\|X_i\| > B\right)\right) = 0, \tag{45}$$

then

$$\lim_{B\to\infty} \sup_m E\left(\left\|m^{-1}\widetilde{X}_m\right\|^r 1\left(\left\|m^{-1}\widetilde{X}_m\right\| > B\right)\right) = 0. \tag{46}$$

Proof of Lemma 1. The proof is based on the proof of Theorem 1 of Etemadi (2006). Eq. (45) implies that $\sup_i E\,\|X_i\|^r \leq C$ for some $C < \infty$. By the C_r inequality

$$\left\|m^{-1}\widetilde{X}_m\right\|^r = \frac{1}{m^r}\left\|\sum_{i=1}^m X_i\right\|^r \leq \frac{1}{m}\sum_{i=1}^m \|X_i\|^r \tag{47}$$

and hence

$$E\left\|m^{-1}\widetilde{X}_m\right\|^r \leq C. \tag{48}$$

Fix $\varepsilon > 0$. Find $B \geq (C/\varepsilon)^{2/r}$ sufficiently large such that

$$\sup_i E\left(\|X_i\|^r \, 1\left(\|X_i\| > \sqrt{B}\right)\right) \leq \varepsilon, \tag{49}$$

which is feasible under (45). Using (47),

$$\begin{aligned}
&E\left(\left\|m^{-1}\widetilde{X}_m\right\|^r 1\left(\left\|m^{-1}\widetilde{X}_m\right\| > B\right)\right)\\
&\leq \frac{1}{m}\sum_{i=1}^m E\left(\|X_i\|^r \, 1\left(\left\|m^{-1}\widetilde{X}_m\right\| > B\right)\right)\\
&= \frac{1}{m}\sum_{i=1}^m E\left(\|X_i\|^r \, 1\left(\left\|m^{-1}\widetilde{X}_m\right\| > B\right) 1\left(\|X_i\| > \sqrt{B}\right)\right)\\
&+ \frac{1}{m}\sum_{i=1}^m E\left(\|X_i\|^r \, 1\left(\left\|m^{-1}\widetilde{X}_m\right\| > B\right) 1\left(\|X_i\| \leq \sqrt{B}\right)\right)\\
&\leq \frac{1}{m}\sum_{i=1}^m E\left(\|X_i\|^r \, 1\left(\|X_i\| > \sqrt{B}\right)\right) + B^{r/2}E1\left(\left\|m^{-1}\widetilde{X}_m\right\| > B\right)\\
&\leq \varepsilon + \frac{E\left\|m^{-1}\widetilde{X}_m\right\|^r}{B^{r/2}}\\
&\leq 2\varepsilon
\end{aligned}$$

by (49), Markov's inequality, (48), and $B^{r/2} \geq C/\varepsilon$. Since ε is arbitrary this implies (46). ■

The next Lemma is useful for establishing the WLLN and CLT for the vectorized clustered second moments.

Lemma 2. *For random vectors X_i set $\widetilde{X}_m = \sum_{i=1}^m X_i$ and $\widetilde{f}_m = \widetilde{X}_m \otimes \widetilde{X}_m$ or $\widetilde{f}_m = \left(\widetilde{X}_m - m\overline{X}_n\right) \otimes \left(\widetilde{X}_m - m\overline{X}_n\right)$ where $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$. For $r \geq 2$, if (45) holds then*

$$\lim_{B\to\infty} \sup_m E\left(\left\|m^{-2}\left(\widetilde{f}_m - E\widetilde{f}_m\right)\right\|^{r/2} 1\left(\left\|m^{-2}\left(\widetilde{f}_m - E\widetilde{f}_m\right)\right\| > B\right)\right) = 0. \tag{50}$$

Proof of Lemma 2. The proof proceeds similar to that of Lemma 1. First consider $\widetilde{f}_m = \widetilde{X}_m \otimes \widetilde{X}_m$. By the triangle inequality, the C_r inequality, the fact that $\|\widetilde{X}_m \otimes \widetilde{X}_m\|^{r/2} = \|\widetilde{X}_m\|^r$, and (48),

$$\begin{aligned}
\left\|m^{-2}\left(\widetilde{f}_m - E\widetilde{f}_m\right)\right\|^{r/2} &\leq \left(\left\|m^{-2}\widetilde{f}_m\right\| + \left\|m^{-2}E\widetilde{f}_m\right\|\right)^{r/2}\\
&\leq 2^{r/2-1}\left(\left\|m^{-2}\widetilde{f}_m\right\|^{r/2} + E\left\|m^{-2}\widetilde{f}_m\right\|^{r/2}\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq 2^{r/2-1}\left(\left\|m^{-1}\widetilde{X}_m\right\|^r + E\left\|m^{-1}\widetilde{X}_m\right\|^r\right)\\
&\leq 2^{r/2-1}\left(\left\|m^{-1}\widetilde{X}_m\right\|^r + C\right). \qquad (51)
\end{aligned}
$$

Fix $\varepsilon > 0$. Find $B \geq \left(2^{r-2}C(1+\sqrt{1+2^{3-r}\varepsilon})/\varepsilon\right)^{4/r}$ sufficiently large such that

$$
\sup_i E\left(\|X_i\|^r 1\left(\|X_i\| > B^{1/4}\right)\right) \leq \frac{\varepsilon}{2^{r/2-1}}, \qquad (52)
$$

which is feasible under (45). Using (47) and (51),

$$
\begin{aligned}
&E\left(\left\|m^{-2}\left(\widetilde{f}_m - E\widetilde{f}_m\right)\right\|^{r/2} 1\left(\left\|m^{-2}\left(\widetilde{f}_m - E\widetilde{f}_m\right)\right\| > B\right)\right)\\
&\leq 2^{r/2-1}E\left(\left(\left\|m^{-1}\widetilde{X}_m\right\|^r + C\right) 1\left(\left\|m^{-2}\left(\widetilde{f}_m - E\widetilde{f}_m\right)\right\| > B\right)\right)\\
&= 2^{r/2-1}\frac{1}{m}\sum_{i=1}^m E\left(\|X_i\|^r 1\left(\left\|m^{-2}\left(\widetilde{f}_m - E\widetilde{f}_m\right)\right\| > B\right) 1\left(\|X_i\| > B^{1/4}\right)\right)\\
&+ 2^{r/2-1}\frac{1}{m}\sum_{i=1}^m E\left(\|X_i\|^r 1\left(\left\|m^{-2}\left(\widetilde{f}_m - E\widetilde{f}_m\right)\right\| > B\right) 1\left(\|X_i\| \leq B^{1/4}\right)\right)\\
&+ 2^{r/2-1}CE\left(1\left(\left\|m^{-2}\left(\widetilde{f}_m - E\widetilde{f}_m\right)\right\| > B\right)\right)\\
&\leq 2^{r/2-1}\frac{1}{m}\sum_{i=1}^m E\left(\|X_i\|^r 1\left(\|X_i\| > B^{1/4}\right)\right)\\
&+ 2^{r/2-1}\left(B^{r/4} + C\right)E\left(1\left(\left\|m^{-2}\left(\widetilde{f}_m - E\widetilde{f}_m\right)\right\| > B\right)\right)\\
&\leq \varepsilon + 2^{r/2-1}\left(B^{r/4} + C\right)\frac{E\left\|m^{-2}\left(\widetilde{f}_m - E\widetilde{f}_m\right)\right\|^{r/2}}{B^{r/2}}\\
&\leq 2\varepsilon
\end{aligned}
$$

by (52), Markov's inequality, (48), and $2^{r-1}(B^{r/4}+C)C/B^{r/2} \leq \varepsilon$ using the discriminant. Since ε is arbitrary this implies (50).

Now consider $\widetilde{f}_m = \left(\widetilde{X}_m - m\overline{X}_n\right) \otimes \left(\widetilde{X}_m - m\overline{X}_n\right)$. By Minkowski's inequality, the C_r inequality, (47), and (48),

$$
\begin{aligned}
E\left\|m^{-1}\left(\widetilde{X}_m - m\overline{X}_n\right)\right\|^r &= E\left\|m^{-1}\sum_{i=1}^m X_i - n^{-1}\sum_{i=1}^n X_i\right\|^r\\
&\leq E\left(\left\|m^{-1}\sum_{i=1}^m X_i\right\| + \left\|n^{-1}\sum_{i=1}^n X_i\right\|\right)^r\\
&\leq 2^r C
\end{aligned}
$$

and

$$
\begin{aligned}
\left\|m^{-2}\left(\widetilde{f}_m - E\widetilde{f}_m\right)\right\|^{r/2} &\leq \left(\left\|m^{-2}\widetilde{f}_m\right\| + \left\|m^{-2}E\widetilde{f}_m\right\|\right)^{r/2}\\
&\leq 2^{r/2-1}\left(\left\|m^{-1}\left(\widetilde{X}_m - m\overline{X}_n\right)\right\|^r + E\left\|m^{-1}\left(\widetilde{X}_m - m\overline{X}_n\right)\right\|^r\right)\\
&\leq 2^{3r/2-1}\left(2^{-1}\left(m^{-1}\sum_{i=1}^m\|X\|^r + n^{-1}\sum_{i=1}^n\|X_i\|^r\right) + C\right).
\end{aligned}
$$

Given ε, find $B \geq \left(2^{3r-2}C(1+\sqrt{1+2^{3(1-r)}\varepsilon})/\varepsilon\right)^{4/r}$ sufficiently large such that

$$
\sup_i E\left(\|X_i\|^r 1\left(\|X_i\| > B^{1/4}\right)\right) \leq \frac{\varepsilon}{2^{3r/2-1}},
$$

and proceed as above to show (50). This completes the proof. ■

Proof of Theorem 1. Without loss of generality assume $EX_i = 0$. Fix $\varepsilon > 0$. Pick B sufficiently large so that

$$
\sup_g E\left\|\left(n_g^{-1}\widetilde{X}_g 1\left(\left\|n_g^{-1}\widetilde{X}_g\right\| > B\right)\right) - E\left(n_g^{-1}\widetilde{X}_g 1\left(\left\|n_g^{-1}\widetilde{X}_g\right\| > B\right)\right)\right\| \leq \varepsilon \qquad (53)
$$

which is feasible by Lemma 1 with $r = 1$ under (2). Using the triangle inequality, Jensen's inequality and (53),

$$\begin{aligned}
E\left\|\overline{X}_n\right\| &= E\left\|\frac{1}{n}\sum_{g=1}^{G}\widetilde{X}_g\right\| \\
&\le E\left\|\frac{1}{n}\sum_{g=1}^{G}\left(\widetilde{X}_g 1\left(\left\|n_g^{-1}\widetilde{X}_g\right\| \le B\right) - E\left(\widetilde{X}_g 1\left(\left\|n_g^{-1}\widetilde{X}_g\right\| \le B\right)\right)\right)\right\| \\
&+ \frac{1}{n}\sum_{g=1}^{G}E\left\|\left(\widetilde{X}_g 1\left(\left\|n_g^{-1}\widetilde{X}_g\right\| > B\right) - E\left(\widetilde{X}_g 1\left(\left\|n_g^{-1}\widetilde{X}_g\right\| > B\right)\right)\right)\right\| \\
&\le \left(E\left\|\frac{1}{n}\sum_{g=1}^{G}\left(\widetilde{X}_g 1\left(\left\|n_g^{-1}\widetilde{X}_g\right\| \le B\right) - E\left(\widetilde{X}_g 1\left(\left\|n_g^{-1}\widetilde{X}_g\right\| \le B\right)\right)\right)\right\|^2\right)^{1/2} + \frac{1}{n}\sum_{g=1}^{G}n_g\varepsilon \\
&= \left(\frac{1}{n^2}\sum_{g=1}^{G}E\left\|\widetilde{X}_g 1\left(\left\|n_g^{-1}\widetilde{X}_g\right\| \le B\right) - E\left(\widetilde{X}_g 1\left(\left\|n_g^{-1}\widetilde{X}_g\right\| \le B\right)\right)\right\|^2\right)^{1/2} + \varepsilon \\
&\le \left(\frac{4B^2}{n^2}\sum_{g=1}^{G}n_g^2\right)^{1/2} + \varepsilon \\
&\le o(1) + \varepsilon.
\end{aligned}$$

The equality uses the assumption that the clusters are independent and thus uncorrelated and the fact $\sum_{g=1}^{G} n_g = n$. The third inequality uses the bound

$$\left\|\widetilde{X}_g 1\left(\left\|n_g^{-1}\widetilde{X}_g\right\| \le B\right) - E\left(\widetilde{X}_g 1\left(\left\|n_g^{-1}\widetilde{X}_g\right\| \le B\right)\right)\right\| \le 2Bn_g.$$

The fourth inequality is (4). Since ε is arbitrary, $E\left\|\overline{X}_n\right\| \to 0$. By Markov's inequality, (3) follows. ■

Proof of Theorem 2. Without loss of generality we assume $EX_i = 0$. Note that

$$\Omega_n^{-1/2}\sqrt{n}\overline{X}_n = \Omega_n^{-1/2}\sum_{g=1}^{G}n^{-1/2}\widetilde{X}_g$$

We apply the multivariate Lindeberg–Feller central limit theorem (e.g. Hansen (2018) Theorem 6.15) since $\widetilde{X}_g$ are independent but not identically distributed. A sufficient condition for the CLT (9) is that for all $\varepsilon > 0$

$$\frac{1}{n\lambda_n}\sum_{g=1}^{G}E\left(\left\|\widetilde{X}_g\right\|^2 1\left(\left\|\widetilde{X}_g\right\|^2 \ge n\lambda_n\varepsilon\right)\right) \to 0 \tag{54}$$

as $n \to \infty$.

Fix $\varepsilon > 0$ and $\delta > 0$. Pick B sufficiently large so that

$$\sup_g E\left(\left\|n_g^{-1}\widetilde{X}_g\right\|^r 1\left(\left\|n_g^{-1}\widetilde{X}_g\right\| > B\right)\right) \le \frac{\delta\varepsilon^{r/2-1}}{C^{r/2}}. \tag{55}$$

which is feasible by Lemma 1 under (7). Pick n large enough so that

$$\max_{g\le G}\frac{n_g}{(n\lambda_n\varepsilon)^{1/2}} \le \frac{1}{B} \tag{56}$$

which is feasible by (12). Thus

$$\begin{aligned}
&\frac{1}{n\lambda_n}\sum_{g=1}^{G}E\left(\left\|\widetilde{X}_g\right\|^2 1\left(\left\|\widetilde{X}_g\right\|^2 \ge n\lambda_n\varepsilon\right)\right) \\
&= \frac{1}{n\lambda_n}\sum_{g=1}^{G}E\left(\frac{\left\|\widetilde{X}_g\right\|^r}{\left\|\widetilde{X}_g\right\|^{r-2}}1\left(\left\|\widetilde{X}_g\right\| \ge (n\lambda_n\varepsilon)^{1/2}\right)\right)
\end{aligned} \tag{57}$$

$$
\begin{aligned}
&\leq \frac{1}{n\lambda_n\left(n\lambda_n\varepsilon\right)^{(r-2)/2}}\sum_{g=1}^{G} E\left(\left\|\widetilde{X}_g\right\|^r 1\left(\left\|\widetilde{X}_g\right\| \geq \left(n\lambda_n\varepsilon\right)^{1/2}\right)\right)\\
&\leq \frac{1}{\varepsilon^{r/2-1}\left(n\lambda_n\right)^{r/2}}\sum_{g=1}^{G} n_g^r E\left(\left\|n_g^{-1}\widetilde{X}_g\right\|^r 1\left(\left\|n_g^{-1}\widetilde{X}_g\right\| \geq B\right)\right)\\
&\leq \frac{\delta}{C^{r/2}}\frac{\sum_{g=1}^{G} n_g^r}{\left(n\lambda_n\right)^{r/2}}\\
&\leq \delta.
\end{aligned}
$$

The second inequality is (56), the third is (55), and the final is (11). Since ε and δ are arbitrary we have established (54) and hence (9). ■

Proof of Theorem 3. Fix $\delta > 0$. Set $\varepsilon = \delta^2/4p$. Define $\widetilde{X}_g^* = \Omega_n^{-1/2}\widetilde{X}_g$ and $\widetilde{Y}_g = \widetilde{X}_g^* 1\left(\left\|\widetilde{X}_g^*\right\|^2 \leq n\varepsilon\right)$. Then

$$
\begin{aligned}
\widetilde{\Omega}_n^* &= \frac{1}{n}\sum_{g=1}^{G}\widetilde{X}_g^*\widetilde{X}_g^{*\prime}\\
&= \frac{1}{n}\sum_{g=1}^{G}\widetilde{Y}_g\widetilde{Y}_g' + \frac{1}{n}\sum_{g=1}^{G}\widetilde{X}_g^*\widetilde{X}_g^{*\prime} 1\left(\left\|\widetilde{X}_g^*\right\|^2 > n\varepsilon\right).
\end{aligned}
$$

By the triangle inequality,

$$
E\left\|\widetilde{\Omega}_n^* - I_p\right\| \leq \frac{1}{n}E\left\|\sum_{g=1}^{G}\left(\widetilde{Y}_g\widetilde{Y}_g' - E\left(\widetilde{Y}_g\widetilde{Y}_g'\right)\right)\right\| \tag{58}
$$

$$
+\frac{2}{n}\sum_{g=1}^{G}E\left(\left\|\widetilde{X}_g^*\right\|^2 1\left(\left\|\widetilde{X}_g^*\right\|^2 > n\varepsilon\right)\right). \tag{59}
$$

An argument similar to (57) shows that for n sufficiently large (59) is bounded by 2δ. We now consider (58).

Using Jensen's inequality, the assumption that the clusters are independent and thus uncorrelated, and the triangle inequality, (58) is bounded by

$$
\begin{aligned}
\frac{1}{n}\left(E\left\|\sum_{g=1}^{G}\left(\widetilde{Y}_g\widetilde{Y}_g' - E\left(\widetilde{Y}_g\widetilde{Y}_g'\right)\right)\right\|^2\right)^{1/2} &= \frac{1}{n}\left(\sum_{g=1}^{G}E\left\|\widetilde{Y}_g\widetilde{Y}_g' - E\left(\widetilde{Y}_g\widetilde{Y}_g'\right)\right\|^2\right)^{1/2}\\
&\leq \frac{2}{n}\left(\sum_{g=1}^{G}E\left\|\widetilde{Y}_g\widetilde{Y}_g'\right\|^2\right)^{1/2}.
\end{aligned} \tag{60}
$$

Using the bounds $\left\|\widetilde{Y}_g\widetilde{Y}_g'\right\| \leq n\varepsilon$ and $\left\|\widetilde{Y}_g\widetilde{Y}_g'\right\| \leq \left\|\widetilde{X}_g^*\right\|^2$, we deduce $\left\|\widetilde{Y}_g\widetilde{Y}_g'\right\|^2 \leq n\varepsilon\left\|\widetilde{X}_g^*\right\|^2$. Thus (60) is bounded by

$$
\begin{aligned}
2\varepsilon^{1/2}\left(\frac{1}{n}\sum_{g=1}^{G}E\left\|\widetilde{X}_g^*\right\|^2\right)^{1/2} &= 2\varepsilon^{1/2}\left(\frac{1}{n}E\left\|\sum_{g=1}^{G}\widetilde{X}_g^*\right\|^2\right)^{1/2}\\
&= 2\varepsilon^{1/2}\left(n\operatorname{var}\left(\overline{X}_n^*\right)\right)^{1/2}\\
&= 2\varepsilon^{1/2}\left(\operatorname{tr}I_p\right)^{1/2}\\
&= \delta
\end{aligned}
$$

The first equality holds because $\widetilde{X}_g^*$ are independent and mean zero, and the second and third use the definition of $\overline{X}_n^*$. The final equality is $\varepsilon = \delta^2/4p$.

Together, we have shown that for n sufficiently large,

$$
E\left\|\widetilde{\Omega}_n^* - I_p\right\| \leq 3\delta
$$

and hence (14) by Markov's Inequality.

By the continuous mapping theorem

$$\widetilde{\Omega}_n^{*-1/2} \xrightarrow{p} I_p^{-1/2} = I_p$$

and

$$\left\| \Omega_n^{-1/4} \widetilde{\Omega}_n^{*-1/2} \Omega_n^{1/4} - I_p \right\| = \left\| \widetilde{\Omega}_n^{*-1/2} - I_p \right\| \xrightarrow{p} 0.$$

Combined with Theorem 2 we find

$$\begin{aligned} &\widetilde{\Omega}_n^{-1/2} \sqrt{n} X_n \\ &= \widetilde{\Omega}_n^{-1/2} \Omega_n^{1/2} \Omega_n^{-1/2} \sqrt{n} X_n \\ &= \Omega_n^{-1/4} \widetilde{\Omega}_n^{*-1/2} \Omega_n^{1/4} \Omega_n^{-1/2} \sqrt{n} X_n \\ &\xrightarrow{d} N\left(\mathbf{0}, I_p\right) \end{aligned}$$

This is (15). ■

Proof of Theorem 4. Since the estimator $\widehat{\Omega}_n$ is invariant to μ, without loss of generality we assume $\mu = 0$. In this case

$$\widehat{\Omega}_n = \widetilde{\Omega}_n - \frac{1}{n} \sum_{g=1}^{G} n_g^2 \overline{X}_n \overline{X}_n'.$$

Then by the triangle inequality, Theorem 3, Theorem 2, and (6),

$$\begin{aligned} &\left\| \Omega_n^{-1/2} \widehat{\Omega}_n \Omega_n^{-1/2} - I_p \right\| \\ &\le \left\| \Omega_n^{-1/2} \widetilde{\Omega}_n \Omega_n^{-1/2} - I_p \right\| \\ &+ \left(\frac{1}{n^2} \sum_{g=1}^{G} n_g^2 \right) \left\| \Omega_n^{-1/2} \sqrt{n} X_n \right\|^2 \\ &\le o_p(1). \end{aligned}$$

This is (16). Eq. (17) follows as in the proof of (15). ■

Proof of Theorem 5. Define the cluster sums $\widetilde{f}_g(\theta) = \sum_{i=1}^{n_g} f(X_{gi}, \theta)$ so that $\overline{f}_n(\theta) = \frac{1}{n} \sum_{g=1}^{G} \widetilde{f}_g(\theta)$ where $\widetilde{f}_g(\theta)$ are mutually independent.

Andrews (1992, Theorem 3) shows that (20) holds if Θ is totally bounded,

$$\left\| \frac{1}{n} \sum_{g=1}^{G} \left(\widetilde{f}_g(\theta) - E\widetilde{f}_g(\theta) \right) \right\| \to_p 0$$

and for all $\theta_1, \theta_2 \in \Theta$,

$$\left\| \widetilde{f}_g(\theta_1) - \widetilde{f}_g(\theta_2) \right\| \le A_g h \left(\|\theta_1 - \theta_2\| \right) \tag{61}$$

where $h(u) \downarrow 0$ as $u \downarrow 0$ and $\frac{1}{n} \sum_{g=1}^{G} E\left(A_g\right) \le A < \infty$. The total boundedness condition holds by assumption and the WLLN holds by Theorem 1 under Assumption 1 and (18), so it only remains to establish the Lipschitz condition (61). Indeed, using the triangle inequality and (19)

$$\begin{aligned} \left\| \widetilde{f}_g(\theta_2) - \widetilde{f}_g(\theta_1) \right\| &= \left\| \sum_{j=1}^{n_g} \left(f(X_{gj}, \theta_2) - f(X_{gj}, \theta_1) \right) \right\| \\ &\le \sum_{j=1}^{n_g} \left\| f(X_{gj}, \theta_2) - f(X_{gj}, \theta_1) \right\| \\ &\le \sum_{j=1}^{n_g} A(X_{gj}) h \left(\|\theta_1 - \theta_2\| \right) \\ &= A_g h \left(\|\theta_1 - \theta_2\| \right) \end{aligned}$$

where $A_g = \sum_{j=1}^{n_g} A(X_{gj})$. Notice that

$$\frac{1}{n}\sum_{g=1}^{G} E\left(A_g\right) = \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} EA(X_{gj}) \leq C$$

since $\sup_i EA(X_i) \leq C$. This verifies (61) and hence (20) holds. ■

Proof of Theorem 6. Without loss of generality, assume $\mu(\theta) = 0$.

We first examine the case with no estimated mean (23). Andrews (1992, Theorem 3) shows that (23) holds if for all $\theta \in \Theta$

$$\left\|\widetilde{\Omega}_n(\theta) - E\widetilde{\Omega}_n(\theta)\right\| \rightarrow_p 0, \tag{62}$$

and for all $\theta_1, \theta_2 \in \Theta$,

$$\left\|\widetilde{f}_g(\theta_1)\widetilde{f}_g(\theta_1)' - \widetilde{f}_g(\theta_2)\widetilde{f}(\theta_2)'\right\| \leq A_g h(\|\theta_1 - \theta_2\|) \tag{63}$$

with $h(u) \downarrow 0$ as $u \downarrow 0$ and $\frac{1}{n}\sum_{g=1}^{G} EA_g \leq A < \infty$. We now establish (62) and (63).

Take (62). Fix $\theta \in \Theta$. For brevity, suppress the dependence of $\widetilde{f}_g(\theta)$ on θ. Fix $\delta > 0$. Set $\varepsilon = (\delta/C)^2$. Define $\widetilde{h}_g = \widetilde{f}_g 1\left(\|\widetilde{f}_g\| \leq \sqrt{n\varepsilon}\right)$. Then

$$\widetilde{\Omega}_n(\theta) = \frac{1}{n}\sum_{g=1}^{G}\widetilde{h}_g\widetilde{h}_g' + \frac{1}{n}\sum_{g=1}^{G}\widetilde{f}_g\widetilde{f}_g' 1\left(\|\widetilde{f}_g\| > \sqrt{n\varepsilon}\right).$$

By the triangle inequality

$$E\left\|\widetilde{\Omega}_n(\theta) - E\widetilde{\Omega}_n(\theta)\right\| = \frac{1}{n}E\left\|\sum_{g=1}^{G}\left(\widetilde{h}_g\widetilde{h}_g' - E\widetilde{h}_g\widetilde{h}_g'\right)\right\| \tag{64}$$

$$+\frac{2}{n}\sum_{g=1}^{G} E\left(\left\|\widetilde{f}_g\right\|^2 1\left(\|\widetilde{f}_g\| > \sqrt{n\varepsilon}\right)\right). \tag{65}$$

Take (64). Assumption (21) and the C_r inequality allow us to deduce that

$$E\left\|\widetilde{f}_g\right\|^2 \leq Cn_g^2 \tag{66}$$

for some $C < \infty$. Using Jensen's inequality, the assumption the clusters are independent and thus uncorrelated, the bounds $\left\|\widetilde{h}_g\right\| \leq \sqrt{n\varepsilon}$ and $\left\|\widetilde{h}_g\right\| \leq \left\|\widetilde{f}_g\right\|$, (66), (5) with $r = 2$ and the definition of ε, we obtain that (64) is bounded by

$$\frac{1}{n}\left(E\left\|\sum_{g=1}^{G}\left(\widetilde{h}_g\widetilde{h}_g' - E\widetilde{h}_g\widetilde{h}_g'\right)\right\|^2\right)^{1/2} \leq \frac{1}{n}\left(\sum_{g=1}^{G} E\left\|\widetilde{h}_g\right\|^4\right)^{1/2}$$

$$\leq \varepsilon^{1/2}C^{1/2}\left(\frac{1}{n}\sum_{g=1}^{G} n_g^2\right)^{1/2} \leq \delta.$$

Take (65). Lemma 1 implies that $\left\|n_g^{-1}\widetilde{f}_g\right\|^2$ is uniformly integrable given Assumption (21). This means we can pick B sufficiently large so that

$$\sup_g E\left(\left\|n_g^{-1}\widetilde{f}_g\right\|^2 1\left(\left\|n_g^{-1}\widetilde{f}_g\right\| > B\right)\right) \leq \frac{\delta}{C} \tag{67}$$

Pick n large enough so that

$$\max_{g \leq G}\frac{n_g}{n^{1/2}} \leq \max_{g \leq G}\frac{n_g^2}{n^{1/2}} \leq \frac{\sqrt{\varepsilon}}{B}$$

which is feasible by (6). Then (65) is bounded by

$$\frac{2}{n}\sum_{g=1}^{G} E\left(\left\|\widetilde{f}_g\right\|^2 1\left(\left\|n_g^{-1}\widetilde{f}_g\right\| > B\right)\right) \leq \frac{2}{n}\sum_{g=1}^{G} n_g^2\frac{\delta}{C} \leq 2\delta,$$

using (5) and (67) with $r = 2$. We have shown that $E\left\|\widetilde{\Omega}_n(\theta) - E\widetilde{\Omega}_n(\theta)\right\| \leq 3\delta$. Since δ is arbitrary, by Markov's inequality, (62) is shown.

Take (63). Fix any $\theta_1, \theta_2 \in \Theta$. Set $\widetilde{f}_g = \sup_{\theta\in\Theta} \left\|\widetilde{f}_g(\theta)\right\|$. Using the triangle inequality and Assumption (19)

$$\left\|\widetilde{f}_g(\theta_2) - \widetilde{f}_g(\theta_1)\right\| \le \sum_{j=1}^{n_g} A(X_{gj}) h\left(\|\theta_1-\theta_2\|\right).$$

Then

$$\begin{aligned}\left\|\widetilde{f}_g(\theta_1)\widetilde{f}_g(\theta_1)' - \widetilde{f}_g(\theta_2)\widetilde{f}(\theta_2)'\right\| &\le 2\widetilde{f}_g \left\|\widetilde{f}_g(\theta_2)-\widetilde{f}(\theta_1)\right\| \\ &\le 2\widetilde{f}_g\left(\sum_{j=1}^{n_g} A(X_{gj})\right) h\left(\|\theta_1-\theta_2\|\right).\end{aligned}$$

Hence (63) holds with $A_g = 2\widetilde{f}_g \left(\sum_{j=1}^{n_g} A(X_{gj})\right)$.

It remains to show that $\frac{1}{n}\sum_{g=1}^G EA_g \le A < \infty$. Assumption (21) and the C_r inequality allow us to deduce that $E\widetilde{f}_g^2 \le Cn_g^2$. Applying Holder's inequality

$$EA_g \le 2\sum_{j=1}^{n_g}\left(E\widetilde{f}_g^2\right)^{1/2}\left(EA^2(X_{gj})\right)^{1/2} \le 2Cn_g^2.$$

Hence

$$\frac{1}{n}\sum_{g=1}^G EA_g \le 2C\frac{1}{n}\sum_{g=1}^G n_g^2 \le 2C^2$$

by Assumption (5) with $r = 2$. This establishes (63).

By showing (62) and (63) we have established (23).

The case with estimated mean (22) immediately follows from (23) and Theorem 5. ■

Proof of Theorem 7. Define

$$\widetilde{f}_g^* = \Omega_n^{-1/2}\widetilde{f}_g$$

$$\overline{f}_G^* = \frac{1}{n}\sum_{g=1}^G\widetilde{f}_g^*.$$

Then

$$\Omega_n^{-1/2}\sqrt{n}\left(\overline{f}_G - E\overline{f}_G\right) = \sqrt{n}\left(\overline{f}_G^* - E\overline{f}_G^*\right)$$

where $n\mathrm{var}\left(\overline{f}_G^*\right) = I_p$.

Since $\widetilde{f}_g^*$ are independent but not identically distributed, we apply the multivariate Lindeberg–Feller central limit theorem (e.g. Hansen (2018) Theorem 6.15). Since $\mathrm{var}\left(\sqrt{n}\overline{f}_G^*\right) = I_p$ a sufficient condition for the CLT (28) is that for all $\varepsilon > 0$

$$\begin{aligned}&\frac{1}{n}\sum_{g=1}^G E\left(\left\|\widetilde{f}_g^* - E\widetilde{f}_g^*\right\|^2 1\left(\left\|\widetilde{f}_g^* - E\widetilde{f}_g^*\right\|^2 \ge n\varepsilon\right)\right) \\ &\le \frac{1}{n\lambda}\sum_{g=1}^G E\left(\left\|\widetilde{f}_g - E\widetilde{f}_g\right\|^2 1\left(\left\|\widetilde{f}_g - E\widetilde{f}_g\right\|^2 \ge n\varepsilon\lambda\right)\right) \to 0\end{aligned} \tag{68}$$

as $n \to \infty$.

Fix $\varepsilon > 0$ and $\delta > 0$. Pick B sufficiently large so that

$$\sup_g E\left(\left\|n_g^{-2}\left(\widetilde{f}_g - E\widetilde{f}_g\right)\right\|^r 1\left(\left\|n_g^{-2}\left(\widetilde{f}_g - E\widetilde{f}_g\right)\right\| > B\right)\right) \le \frac{\delta\varepsilon^{r/2-1}\lambda^{r/2}}{C^{r/2}}. \tag{69}$$

which is feasible by Lemma 2 under (26). Pick n large enough so that

$$\max_{g\le G}\frac{n_g^2}{n^{1/2}} \le \frac{(\varepsilon\lambda)^{1/2}}{B} \tag{70}$$

which is feasible by (25). Thus

$$\begin{aligned}
&\frac{1}{n\lambda}\sum_{g=1}^{G} E\left(\left\|\widetilde{f}_g - E\widetilde{f}_g\right\|^2 1\left(\left\|\widetilde{f}_g - E\widetilde{f}_g\right\|^2 \geq n\varepsilon\lambda\right)\right) \qquad (71)\\
&= \frac{1}{n\lambda}\sum_{g=1}^{G} E\left(\frac{\left\|\widetilde{f}_g - E\widetilde{f}_g\right\|^r}{\left\|\widetilde{f}_g - E\widetilde{f}_g\right\|^{r-2}} 1\left(\left\|\widetilde{f}_g - E\widetilde{f}_g\right\| \geq (n\varepsilon\lambda)^{1/2}\right)\right)\\
&\leq \frac{1}{\varepsilon^{r/2-1}n^{r/2}\lambda^{r/2}}\sum_{g=1}^{G} E\left(\left\|\widetilde{f}_g - E\widetilde{f}_g\right\|^r 1\left(\left\|\widetilde{f}_g - E\widetilde{f}_g\right\| \geq (n\varepsilon\lambda)^{1/2}\right)\right)\\
&\leq \frac{1}{\varepsilon^{r/2-1}n^{r/2}\lambda^{r/2}}\sum_{g=1}^{G} n_g^{2r} E\left(\left\|n_g^{-2}\left(\widetilde{f}_g - E\widetilde{f}_g\right)\right\|^r 1\left(\left\|n_g^{-2}\left(\widetilde{f}_g - E\widetilde{f}_g\right)\right\| \geq B\right)\right)\\
&\leq \frac{\delta}{C^{r/2}}\frac{\sum_{g=1}^{G} n_g^{2r}}{n^{r/2}}\\
&\leq \delta.
\end{aligned}$$

The second inequality is (70), the third is (69), and the final is (24). Since ε and δ are arbitrary we have established (68) and hence (28). ■

The proofs of Theorems 8–Theorem 13 are presented in the Supplemental Appendix.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2019.02.001.

References

Andrews, Donald W.K., 1992. Generic uniform convergence. Econometric Theory 8 (2), 241–257.

Angrist, Joshua D., Pischke, J.S., 2009. , Mostly Harmless Econometrics: An Empiricist'S Companion. Princeton University Press, Princeton.

Arellano, Manuel, 1987. Computing robust standard errors for within-groups estimators. Oxford Bulletin of Economics and Statistics 49, 431–434.

Bertrand, Marianne, Duflo, Esther, Mullainathan, Sendhil, 2004. How much should we trust differences-in-differences estimates?. Q. J. Econ. 119 (1) 249–275.

Bester, C. Alan, Conley, Timothy G., Hansen, Christian B., 2011. Inference with dependent data using cluster covariance estimators. J. Econometrics 165 (2), 137–151.

Cameron, A. Colin, Gelbach, Jonah B., Miller, Douglas L., 2008. Bootstrap-based improvements for inference with clustered errors. Review of Economics and Statistics 90, 414–427.

Cameron, A. Colin, Miller, Douglas L., 2011. Robust inference with clustered data. In: Ullah, A., Giles, D.E. (Eds.), HandBook of Empirical Economics and Finance. CRC Press, New York, pp. 1–28.

Cameron, A. Colin, Miller, Douglas L., 2015. A practitioner's guide to cluster robust inference. J. Hum. Resour. 50, 317–372.

Canay, Ivan A., Romano, Joseph P., Shaikh, Azeem M., 2017. Randomization tests under an approximate symmetry assumption. Econometrica 85 (3) 1013–1030.

Carter, Andrew V., Schnepel, Kevin T., Steigerwald, Douglas G., 2017. Asymptotic behavior of a t-test robust to cluster heterogeneity. Rev. Econ. Stat. 99, 698–709.

Conley, Timothy G., Taber, Christopher R., 2011. Inference with 'difference in differences' with a small number of policy changes. Rev. Econ. Stat. 93, 113–125.

Djogbenou, Antoine A., MacKinnon, James G., Nielsen, Morten Ørregaard, 2018. Asymptotic Theory and Wild Bootstrap Inference with Clustered Errors. Queen's University, Working paper.

Donald, Stephen G., Lang, Kevin, 2007. Inference with difference in differences and other panel data. Rev. Econ. Stat. 89, 221–223.

Etemadi, Nasrollah, 2006. Convergence of weighted averages of random variables revisited. Proc. Amer. Math. Soc. 134 (9), 2739–2744.

Hall, Alastair R., Inoue, Atsushi, 2003. The large sample behaviour of the generalized method of moments estimator in misspecified models. J. Econometrics 114 (2), 361–394.

Hansen, Christian B., 2007. Asymptotic properties of a robust variance matrix estimator for panel data when T is large. J. Econometrics 141, 597–620.

Hansen, Bruce E., 2018. Econometrics http://www.ssc.wisc.edu/~bhansen/econometrics/.

Hansen, Bruce E., Lee, Seojeong, Inference for Iterated GMM under Misspecification and Clustering, Working paper.

Hwang, Jungbin, 2017. Simple and Trustworthy Cluster-Robust GMM Inference. University of Connecticut, Working paper.

Ibragimov, Rustam, Müller, Ulrich K., 2010. t-statistic based correlation and heterogeneity robust inference. J. Bus. Econom. Statist. 28, 453–468.

Ibragimov, Rustam, Müller, Ulrich K., 2016. Inference with few heterogeneous clusters. Rev. Econ. Stat. 98, 83–96.

Imbens, Guido W., Kolesár, Michal, 2016. Robust standard errors in small samples: Some practical advice. Rev. Econ. Stat. 98, 701–712.

Lee, Seojeong, 2014. Asymptotic refinements of a misspecification-robust bootstrap for generalized method of moments estimators. J. Econometrics 178, 398–413.

Liang, Kung-Yee, Zeger, Scott L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73, 13–22.

MacKinnon, James G., 2012. Thirty years of heteroskedasticity-robust inference. In: Chen, Xiaohong, Swanson, Norman R. (Eds.), Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis. Springer, New York, pp. 437–461.

MacKinnon, James G., 2016. Inference with large clustered datasets. L'Actualité Économique 92, 649–665.

MacKinnon, James G., Nielsen, Morten Ørregaard, Webb, Matthew D., 2017. Bootstrap and Asymptotic Inference with Multiway Clustering. Workking Paper. Queen's University.

MacKinnon, James G., Webb, Matthew D., 2017. Wild bootstrap inference for wildly different cluster sizes. J. Appl. Econometrics 32, 233–254.

MacKinnon, James G., Webb, Matthew D., 2018. The wild bootstrap for few (treated) clusters. Econom. J. 21, 114–135.

Moulton, Brent R., 1986. Random group effects and the precision of regression estimates. J. Econometrics 32, 385–397.

Moulton, Brent R., 1990. An illustration of a pitfall in estimating the effects of aggregate variables on micro units. Rev. Econ. Stat. 72, 334–338.

Newey, Whitney K., McFadden, Daniel, 1994. Large sample estimation and hypothesis testing. Handb. Econom. 4, 2111–2245.

Rogers, William, 1993. Regression standard errors in clustered samples. STATA Technical Bulletin 13, 19–23.

Tabord-Meehan, Max, 2018. Inference with dyadic data: Asymptotic behavior of the dyadic-robust t-statistic. J. Bus. Econom. Statist.

White, Halbert, 1984. Asymptotic Theory for Econometricians. Academic Press.

Wooldridge, Jeffrey M., 2003. Cluster-sample methods in applied econometrics. Amer. Econ. Rev. 93, 133–138.

Wooldridge, Jeffrey M., 2010. Econometric Analysis of Cross Section and Panel Data, second ed. MIT Press, Cambridge.

Young, Alwyn, 2016. Improved, Nearly Exact, Statistical Inference with Robust and Clustered Covariance Matrices using Effective Degrees of Freedom Corrections. London School of Economics, Working paper.