Business School / School of Economics

UNSW Business School Working Paper

UNSW Business School Research Paper No. 2017 ECON 18

Asymptotic Theory for Clustered Samples

Bruce E Hansen
Seojeong Jay Lee

Asymptotic Theory for Clustered Samples

Bruce E. Hansen*
University of Wisconsin

Seojeong Lee†
University of New South Wales

December 2017‡

Abstract

We provide a complete asymptotic distribution theory for clustered data with a large number of 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.

*Hansen thanks the National Science Foundation and the Phipps Chair for research support.

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

‡We thank Morten Nielsen and James MacKinnon for valuable conversations and suggestions.

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 no individual cluster dominates asymptotically. 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 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. Our 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.

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 (1994) and the widely-cited paper of Bertrand, Duflo and Mullainathan (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, Schnepel and Steigerwald (2017) provided asymptotic results allowing for heterogeneous clusters, but their results are limited by atypical regularity conditions. Independently of this paper, Djogbenou, Nielsen and MacKinnon (2017) have provided a rigorous asymptotic theory for heterogeneous clusters, with similar but somewhat 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, Conley and C. Hansen (2011), and Ibragimov and Müeller (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, Romano, and Shaikh (2017) provide approximate randomization tests.

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

A recent contribution which develops cluster-robust inference for GMM is Hwang (2016).

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 (Section 9), 2SLS (Section 10), Maximum Likelihood (Section 11), and GMM (Section 12). Each of these latter sections are written self-sufficiently, so they can be used directly by readers. Technical details of the proofs are provided in the Appendix.

2 Cluster Sampling

The observations are $X_i \in \mathbb{R}^p$ for $i = 1, ..., n$, obtained by cluster sampling. They are grouped into G mutually independent known clusters, indexed $g = 1, ..., G$, where the g^{th} cluster has n_g observations. 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 $g = 1, ..., G$ and $j = 1, ..., n_g$.

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 necessarily require that the observations or clusters come from homogeneous distributions. Thus our framework includes i.n.i.d (independent, not necessarily identically distributed) as the special case $n_g = 1$.

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.

A word on notation. For a vector a let $\|a\| = (a'a)^{1/2}$ 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 $\|A\| = \sqrt{\lambda_{\max}(A'A)}$ 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 estimate of $E\overline{X}_n$.

Theorem 1. *(WLLN for clustered means). If*

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

and

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

as $n \to \infty$*, then*

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

The assumption (1) states that X_i is uniformly integrable. 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. Assumption (1) simplifies to $E\|X_i\| < \infty$ when the observations have identical marginal distributions. A sufficient condition allowing for distributional heterogeneity is $\sup_i E\|X_i\|^r < \infty$ for some $r > 1$.

The assumption (2) states that each cluster size n_g is asymptotically negligible. This implies $G \to \infty$, so we do not explicitly need to add this assumption. Assumption (2) 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 (2) is necessary for parameter estimation consistency while allowing arbitrary within-cluster dependence, as otherwise a single cluster could dominate the sample average.

Assumption (2) 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 (2). Thus (2) implies (4). Second,

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

under (4). Thus (4) implies (2), so the two are equivalent.

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.

The convergence rate can be calculated as the standard deviation of the sample mean. To calculate the latter it is 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 $\overline{X}_n = \frac{1}{n}\sum_{g=1}^{G} \widetilde{X}_g$.

For simplicity consider the scalar case $p = 1$. The standard deviation of $\overline{X}_n$ is then

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

Under i.i.d. sampling $\mathrm{sd}\left(\overline{X}_n\right) = O\left(n^{-1/2}\right)$. We now illustrate possible rates of convergence under cluster sampling in five examples.

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

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

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

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

Example 4. The observations follow random walks: $X_{gj} = X_{gj-1} + \varepsilon_{gj}$ with ε_{gj} i.i.d. $(0,1)$, $X_{g0} = 0$. 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 $G^{-1/2}$.

We now consider an example with heterogeneous cluster sizes.

Example 5. The clusters are of two sizes, $n_g = n^{\alpha_1}$ and $n_g = n^{\alpha_2}$ for $0 \leq \alpha_1 < \alpha_2 < 1$. There are $G_1 = 2^{-1}n^{1-\alpha_1}$ of the first type and $G_2 = 2^{-1}n^{1-\alpha_2}$ of the second type. (So $G = G_1 + G_2 = O\left(n^{1-\alpha_1}\right)$.) Within each cluster the observations are identical and have unit variances. The variances for the two types of clusters are $n^{2\alpha_1}$ and $n^{2\alpha_2}$. Then $\text{sd}\left(\overline{X}_n\right) = \left(\frac{n^2}{G_1 n^{2\alpha_1} + G_2 n^{2\alpha_2}}\right)^{-1/2} = \left(\frac{2n}{n^{\alpha_1}+n^{\alpha_2}}\right)^{-1/2} = O\left(n^{-(1-\alpha_2)/2}\right)$. Thus $\text{sd}\left(\overline{X}_n\right)$ converges at at a rate slower than both $n^{-1/2}$ and $G^{-1/2}$.

What we have seen is that the convergence rate $\text{sd}\left(\overline{X}_n\right)$ 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

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

Furthermore, as we discussed in the previous section, the elements of $\overline{X}_n$ may converge at different rates. It is therefore important to develop the CLT for both individual coefficients and sub-groups of coefficients. We do this by presenting results for linear combinations $R_n'\left(\overline{X}_n - E\overline{X}_n\right)$ where R_n is a non-random sequence of $p \times q$ matrices.

Theorem 2. *(CLT) If for some* $2 \leq r < \infty$

$$\lim_{M\to\infty} \sup_i \left(E\left\|X_i\right\|^r \mathbb{1}\left(\left\|X_i\right\| > M\right)\right) = 0, \tag{5}$$

$$\frac{\left(\sum_{g=1}^{G} n_g^r\right)^{2/r}}{n} \leq C < \infty, \tag{6}$$

$$\max_{g \leq G} \frac{n_g^2}{n} \to 0, \tag{7}$$

and

$$\lambda_{\min}(\Omega_n) \geq \lambda > 0, \tag{8}$$

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

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

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 $G = n$.

The assumptions (5)-(7) are stronger versions of (1)-(2), and thus the conditions of Theorem 2 imply those of Theorem 1.

Assumption (5) states that $\|X_i\|^r$ is uniformly integrable. When $r = 2$ this is similar to the Lindeberg condition for the CLT under independent heterogeneous sampling. Assumption (5) simplifies to $E\|X_i\|^r < \infty$ when the observations have identical marginal distributions. A sufficient condition allowing for distributional heterogeneity is $\sup_i E\|X_i\|^s < \infty$ for some $s > r \geq 2$.

Assumption (6) 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 $\max_{g \leq G} n_g^2/n = O(1)$, which is implied by (7).

Assumptions (6)-(7) allow for growing and heterogeneous cluster sizes. For example, it allows clusters to grow uniformly at the rate $n_g = n^\alpha$ for $0 \leq \alpha \leq (r-2)/2(r-1)$. (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 (6)-(7) hold for any $\alpha < 1$ and $r \geq 2$.

Assumption (6) is implied by

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

and they are equivalent when the cluster sizes are homogeneous. In general, however, (6) is less restrictive than (10). For example, when $r = 2$ (10) requires the cluster sizes to be bounded, while (6) does not. (Consider the heterogeneous example given in the previous paragraph. This satisfies (6) but not (10) when $r = 2$.)

Assumption (8) specifies that $\operatorname{var}\left(\sqrt{n}\alpha'\overline{X}_n\right)$ does not vanish for any conformable vector $\alpha \neq 0$. This excludes degenerate cases and perfect negative within-cluster correlation.

To compare our conditions with those of Djogbenou, MacKinnon, and Nielsen (2017), a small modification of our proof shows that assumptions (6)-(8) could be replaced by the single condition

$$\max_{g \leq G} \frac{n_g}{n^{(r-2)/2(r-1)} \lambda_{\min}(\Omega_n)^{r/2(r-1)}} \leq C, \tag{11}$$

which is that used in Theorem 1 of Djogbenou, MacKinnon, and Nielsen (2017), except they require the right-hand-side of (11) to be $o(1)$ rather than $O(1)$ and require $r \geq 4$. Assumption (11) is less restrictive than (10) when $\lambda_{\min}(\Omega_n)$ diverges to infinity (which means that all components of $\overline{X}_n$ converge at a rate slower than $\sqrt{n}$). While the examples in Section 4 show that this is possible, the actual rate of $\lambda_{\min}(\Omega_n)$ is unknown in any given application, meaning that the condition (11) is difficult to interpret and impossible to empirically verify. Hence we prefer the simpler bound (6) with (8).

Theorem 2 shows that the CLT holds for the sample mean $\overline{X}_n$ and linear combinations $R_n'\overline{X}_n$, and thus includes inference on sub-components.

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). This does not have practical application but is the foundation for all covariance matrix estimation theory. When $EX_i = 0$ 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 $n \to \infty$*

$$\left(R_n'\Omega_n R_n\right)^{-1/2}\left(R_n'\widetilde{\Omega}_n R_n\right)\left(R_n'\Omega_n R_n\right)^{-1/2} \xrightarrow{p} I_q \tag{12}$$

and

$$\left(R_n'\widetilde{\Omega}_n R_n\right)^{-1/2} R_n'\sqrt{n}\overline{X}_n \xrightarrow{d} N\left(\mathbf{0}, I_q\right). \tag{13}$$

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 (13) are that cluster-robust t-ratios are asymptotically standard normal, and that cluster-robust Wald statistics are asymptotically chi-square distributed with q degrees of freedom.

Construction of practical covariance matrix estimators are 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 $n \to \infty$*

$$\left(R_n'\Omega_n R_n\right)^{-1/2}\left(R_n'\widehat{\Omega}_n R_n\right)\left(R_n'\Omega_n R_n\right)^{-1/2} \xrightarrow{p} I_q \tag{14}$$

and

$$\left(R_n'\widehat{\Omega}_n R_n\right)^{-1/2} R_n'\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{d} N\left(\mathbf{0}, I_q\right). \tag{15}$$

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

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

Theorem 5. *(ULLN for clustered means). If (2) holds, for each $\theta \in \Theta$*

$$\lim_{M\to\infty} \sup_i \left(E\left\|f(X_i,\theta)\right\| 1\left(\left\|f(X_i,\theta)\right\| > M\right)\right) = 0, \tag{16}$$

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

$$\left\|f(x,\theta_1) - f(x,\theta_2)\right\| \leq A(x)h\left(\left\|\theta_1 - \theta_2\right\|\right) \tag{17}$$

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 $n \to \infty$

$$\sup_{\theta\in\Theta}\left\|\overline{f}_n(\theta) - E\overline{f}_n(\theta)\right\| \xrightarrow{p} 0. \tag{18}$$

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

$$\begin{aligned}\Omega_n(\theta) &= E\left(n\left(\overline{f}_n(\theta) - E\overline{f}_n(\theta)\right)\right)\left(\overline{f}_n(\theta) - E\overline{f}_n(\theta)\right)'\Big) \\ &= \frac{1}{n}\sum_{g=1}^{G} E\left(\left(\widetilde{f}_g(\theta) - n_g\mu(\theta)\right)\left(\widetilde{f}_g(\theta) - n_g\mu(\theta)\right)'\right).\end{aligned} \tag{19}$$

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

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

Imposing a restriction $\mu(\theta) = 0$, a natural estimator is

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

Theorem 6. *(ULLN for clustered variance). If $\mu(\theta) = Ef(X_i,\theta)$ does not vary across i, for some $2 \le r < \infty$ (6) holds, (7) holds, and for each $\theta \in \Theta$,*

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

and

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

and for each $\theta_1, \theta_2 \in \Theta$ (17) holds with $\sup_i EA(X_i)^2 \le C$, then as $n \to \infty$

$$\sup_{\theta\in\Theta}\left\|\Omega_n(\theta)^{-1/2}\widehat{\Omega}_n(\theta)\Omega_n(\theta)^{-1/2} - I_k\right\| \xrightarrow{p} 0. \tag{20}$$

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

$$\sup_{\theta\in\Theta}\left\|\Omega_n(\theta)^{-1/2}\widetilde{\Omega}_n(\theta)\Omega_n(\theta)^{-1/2} - I_k\right\| \xrightarrow{p} 0. \tag{21}$$

8 Central Limit Theorem for Clustered Second Moments

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

$$\overline{f}_G = \frac{1}{n}\sum_{g=1}^{G}\widetilde{f}_g \tag{22}$$

where

$$\widetilde{f}_g = \widetilde{X}_g \otimes \widetilde{X}_g \tag{23}$$

or

$$\widetilde{f}_g = \left(\widetilde{X}_g - n_g\overline{X}_n\right)\otimes\left(\widetilde{X}_g - n_g\overline{X}_n\right). \tag{24}$$

The WLLN for $\overline{f}_G$ holds by Theorem 3 (12) and Theorem 4 (14), 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, 2017).

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

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

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

Like the sample mean case, different rates of convergence of $\overline{f}_G$ can be handled by considering linear combinations $R_n'(\overline{f}_G - E\overline{f}_G)$ where R_n is a non-random sequence of $p \times q$ matrices. Define the variance matrix of $\sqrt{n}\overline{f}_G$ as

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

Theorem 7. *(CLT for clustered variance) If for some* $2 \le r < \infty$

$$\lim_{M\to\infty}\sup_i \left(E\left\|X_i\right\|^{2r} 1\left(\left\|X_i\right\| > M\right)\right) = 0, \tag{25}$$

$$\frac{\left(\sum_{g=1}^{G} n_g^{2r}\right)^{2/r}}{n} \le C < \infty, \tag{26}$$

$$\max_{g\le G} \frac{n_g^4}{n} \to 0, \tag{27}$$

and

$$\lambda_{\min}\left(\Omega_n\right) \ge \lambda > 0, \tag{28}$$

then as $n \to \infty$*, for any sequence of full-rank* $p^2 \times q$ *matrices* R_n*,*

$$\left(R_n'\Omega_n R_n\right)^{-1/2} R_n'\sqrt{n}\left(\overline{f}_G - E\overline{f}_G\right) \xrightarrow{d} N\left(\mathbf{0}, I_q\right). \tag{29}$$

Note that the conditions of Theorem 7 imply those of Theorem 2.

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

$$\widetilde{\psi}_g = \begin{pmatrix} \widetilde{Y}_g \\ \widetilde{X}_g \\ \widetilde{X}_g \otimes \widetilde{X}_g \end{pmatrix} \tag{30}$$

or

$$\widetilde{\psi}_g = \begin{pmatrix} \widetilde{Y}_g \\ \widetilde{X}_g \\ \left(\widetilde{X}_g - n_g \overline{X}_n\right) \otimes \left(\widetilde{X}_g - n_g \overline{X}_n\right) \end{pmatrix} \tag{31}$$

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* $2 \leq r < \infty$

$$\lim_{M\to\infty} \sup_i \left(E\left\|Y_i\right\|^r 1\left(\left\|X_i\right\| > M\right)\right) = 0, \tag{32}$$

$$\lim_{M\to\infty} \sup_i \left(E\left\|X_i\right\|^{2r} 1\left(\left\|X_i\right\| > M\right)\right) = 0, \tag{33}$$

$$\frac{\left(\sum_{g=1}^{G} n_g^{2r}\right)^{2/r}}{n} \leq C < \infty, \tag{34}$$

$$\max_{g\leq G} \frac{n_g^4}{n} \to 0, \tag{35}$$

and

$$\lambda_{\min}\left(\Omega_n\right) \geq \lambda > 0, \tag{36}$$

then as $n \to \infty$, *for any sequence of full-rank* $(s+p+p^2) \times q$ *matrices* R_n,

$$\left(R_n' \Omega_n R_n\right)^{-1/2} R_n' \sqrt{n}\left(\overline{\psi}_G - E\overline{\psi}_G\right) \xrightarrow{d} N\left(\mathbf{0}, I_q\right). \tag{37}$$

9 Linear Regression

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

$$\begin{aligned} \boldsymbol{y}_g &= \boldsymbol{X}_g \boldsymbol{\beta} + \boldsymbol{e}_g, \\ E\left(\boldsymbol{X}_g' \boldsymbol{e}_g\right) &= 0 \end{aligned} \tag{38}$$

where $\boldsymbol{e}_g$ is a $n_g \times 1$ error vector.

The OLS estimator for $\boldsymbol{\beta}$ can be written as

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{X}_g\right)^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{y}_g\right). \tag{39}$$

We start by providing conditions for consistent estimation. Define

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

Theorem 8. *If (2) holds,* $\lambda_{\min}(Q_n) \geq C > 0$, *and either*

1. $(\boldsymbol{x}_i, y_i)$ *have identical marginal distributions,* $E\|\boldsymbol{x}_i\|^2 < \infty$ *and* $E|y_i|^2 < \infty$;

   or

2. *For some* $r > 2$, $\sup_i E\|\boldsymbol{x}_i\|^r < \infty$ *and* $\sup_i E|y_i|^r < \infty$;

then as $n \to \infty$

$$\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}.$$

The assumptions are quite minimal.

We now provide a distributional theory. Define

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

The residuals for the g^{th} cluster are

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

The variance estimator is

$$\widehat{V}_n = d_n \left(\frac{1}{n} \sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{X}_g\right)^{-1} \left(\frac{1}{n} \sum_{g=1}^{G} \boldsymbol{X}_g' \widehat{\boldsymbol{e}}_g \widehat{\boldsymbol{e}}_g' \boldsymbol{X}_g\right) \left(\frac{1}{n} \sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{X}_g\right)^{-1}.$$

with d_n a possible finite-sample degree-of-freedom adjustment which converges to one. For example, C. Hansen (2007) proposed $d_n = G/(G-1)$ and Stata sets

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

Theorem 9. *Suppose*

$$\lambda_{\min}(Q_n) \geq C > 0, \tag{40}$$

$$\lambda_{\min}(\Omega_n) \geq \lambda > 0, \tag{41}$$

for some $2 \leq r \leq s < \infty$ *(6)-(7) hold,* $\sup_i E\left\|\boldsymbol{x}_i\right\|^{2s} < \infty$ *and* $\sup_i E\left|y_i\right|^{2s} < \infty$*, and either (i)* $(\boldsymbol{x}_i, y_i)$ *have identical marginal distributions; or (ii)* $r < s$*. Then, for any sequence of full-rank* $k \times q$ *matrices* R_n*, as* $n \to \infty$

$$\left(R_n' V_n R_n\right)^{-1/2} R_n' \sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} N\left(\mathbf{0}, I_q\right), \tag{42}$$

$$\left(R_n' V_n R_n\right)^{-1/2} R_n' \widehat{V}_n R_n \left(R_n' V_n R_n\right)^{-1/2} \xrightarrow{p} I_q, \tag{43}$$

and

$$\left(R_n' \widehat{V}_n R_n\right)^{-1/2} R_n' \sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} N\left(\mathbf{0}, I_q\right). \tag{44}$$

Theorem 9 shows that the OLS coefficient estimates are asymptotically normal under our large-G asymptotic framework, and that the asymptotic distribution is unaffected if the variance matrix is replaced with a cluster-robust estimate. The latter implies that cluster-robust t-ratios are asymptotically normal and cluster-robust Wald statistics are asymptotically χ_q^2. The standard errors for $R_n'\widehat{\boldsymbol{\beta}}$ can be obtained by taking the square roots of the diagonal elements of $n^{-1}R_n'\widehat{V}_n R_n$. Another implication is that the conventional t-ratio diverges if the coefficient converges slower than $\sqrt{n}$.

The assumptions for Theorem 9 are quite minimal. The eigenvalue conditions (40)-(41) state that the design and covariances matrices are bounded away from singularity. This allows for considerable heterogeneity and do not require that these matrices converge as n increases. The moment conditions are also minimal, and are identical to the conditions used for i.i.d. and i.n.i.d samples. The bounds (6)-(7) on the cluster sizes allow for heterogeneous and growing cluster sizes, with a trade-off with the number of finite moments.

10 Two-Stage Least Squares

We continue to use the same notations $\boldsymbol{y}_g$ and $\boldsymbol{X}_g$ from the previous section. The standard two-stage least squares model in cluster notation is

$$\boldsymbol{y}_g = \boldsymbol{X}_g \boldsymbol{\beta} + \boldsymbol{e}_g, \tag{45}$$

$$\boldsymbol{X}_g = \boldsymbol{Z}_g \gamma + \boldsymbol{u}_g, \tag{46}$$

$$E\left(\boldsymbol{Z}_g' \boldsymbol{e}_g\right) = 0$$

where $\boldsymbol{Z}_g = (\boldsymbol{z}_{g1}, ..., \boldsymbol{z}_{gn_g})'$ is an $n_g \times l$ matrix of instruments for the g^{th} cluster. Assume $l \geq k$. (45) is the structural equation and (46) is the first-stage equation.

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

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{Z}_g \left(\sum_{g=1}^{G} \boldsymbol{Z}_g' \boldsymbol{Z}_g \right)^{-1} \sum_{g=1}^{G} \boldsymbol{Z}_g' \boldsymbol{X}_g \right)^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{Z}_g \left(\sum_{g=1}^{G} \boldsymbol{Z}_g' \boldsymbol{Z}_g \right)^{-1} \sum_{g=1}^{G} \boldsymbol{Z}_g' \boldsymbol{y}_g \right). \tag{47}$$

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

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

$$W_n = \frac{1}{n} \sum_{g=1}^{G} E\left(\boldsymbol{Z}_g' \boldsymbol{Z}_g\right).$$

Theorem 10. *If (2) 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 $r > 2$, $\sup_i E\left|y_i\right|^r < \infty$, $\sup_i E\left\|\boldsymbol{x}_i\right\|^r < \infty$, and $\sup_i E\left\|\boldsymbol{z}_i\right\|^r < \infty$;*

then as $n \to \infty$

$$\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}.$$

Next we provide the asymptotic distribution. Define

$$\Omega_n = \frac{1}{n} \sum_{g=1}^{G} E\left(\boldsymbol{Z}_g' \boldsymbol{e}_g \boldsymbol{e}_g' \boldsymbol{Z}_g\right),$$

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

The residuals for the g^{th} cluster are

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

Define

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

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. Stata sets

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

for 2SLS estimators with *cluster* option, identical to the OLS case.

Theorem 11. *Suppose Q_n has full rank k, $\lambda_{\min}(W_n) \geq C > 0$, $\lambda_{\min}(\Omega_n) \geq \lambda > 0$, (7) holds, for some $2 \leq r \leq s < \infty$, (6) holds, $\sup_i E|y_i|^{2s} < \infty$, $\sup_i E\|\boldsymbol{x}_i\|^{2s} < \infty$, 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 $n \to \infty$

$$\left(R_n' V_n R_n\right)^{-1/2} R_n' \sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} N\left(\mathbf{0}, I_q\right), \tag{48}$$

$$\left(R_n' V_n R_n\right)^{-1/2} R_n' \widehat{V}_n R_n \left(R_n' V_n R_n\right)^{-1/2} \xrightarrow{p} I_q, \tag{49}$$

and

$$\left(R_n' \widehat{V}_n R_n\right)^{-1/2} R_n' \sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} N\left(\mathbf{0}, I_q\right). \tag{50}$$

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 (Pseudo) Maximum Likelihood

Suppose that we observe a sequence of random vectors $X_i \in \mathbb{R}^p$, $i = 1, ..., n$ with the same marginal distributions from a density $f(x, \boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \mathbb{R}^k$. Define $\boldsymbol{X}_g = (X_{g1}, ..., X_{gn_g})'$ 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

$$\sum_{g=1}^{G} \log f_g(\boldsymbol{X}_g, \boldsymbol{\theta}). \tag{51}$$

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

$$L_n(\theta) = \sum_{g=1}^{G} \sum_{j=1}^{n_g} \log f(X_{gj}, \boldsymbol{\theta}). \tag{52}$$

Define the pseudo-MLE as

$$\widehat{\boldsymbol{\theta}} = \underset{\theta \in \Theta}{\arg\max}\, L_n(\boldsymbol{\theta}). \tag{53}$$

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

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

Theorem 12. *If (2) 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

$$\begin{aligned}
H_n(\boldsymbol{\theta}) &= \frac{1}{n} \sum_{i=1}^{n} E\left[\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \log f(X_i, \boldsymbol{\theta})\right], \\
\Omega_n(\boldsymbol{\theta}) &= \frac{1}{n} \sum_{g=1}^{G} E\left(\sum_{j=1}^{n_g} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(X_{gj}, \boldsymbol{\theta})\right)\left(\sum_{j=1}^{n_g} \frac{\partial}{\partial \boldsymbol{\theta}'} \log f(X_{gj}, \boldsymbol{\theta})\right), \\
V_n &= H_n(\boldsymbol{\theta}_0)^{-1} \Omega_n(\boldsymbol{\theta}_0) H_n(\boldsymbol{\theta}_0)^{-1}.
\end{aligned}$$

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 13. *In addition to the assumptions of Theorem 12, if (6) holds for some $r \geq 2$, (7) holds,*

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) $\int \sup_{\boldsymbol{\theta}\in\mathcal{N}} \left\| \frac{\partial}{\partial\boldsymbol{\theta}} f(x,\boldsymbol{\theta})\right\| dx < \infty$,
   (c) $E\left\|\frac{\partial}{\partial\boldsymbol{\theta}}\log f(X_i,\boldsymbol{\theta})\right\|^r < \infty$,
   (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 $EA(X_i) \leq C$,
   (f) $\lambda_{\min}(H_n(\boldsymbol{\theta})) \geq C > 0$,
   (g) $\lambda_{\min}(\Omega_n(\boldsymbol{\theta})) \geq \lambda > 0$,

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

$$\left(R_n'V_nR_n\right)^{-1/2}R_n'\sqrt{n}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(\mathbf{0},I_q\right), \tag{54}$$

$$\left(R_n'V_nR_n\right)^{-1/2}R_n'\widehat{V}_nR_n\left(R_n'V_nR_n\right)^{-1/2} \xrightarrow{p} I_q, \tag{55}$$

and

$$\left(R_n'\widehat{V}_nR_n\right)^{-1/2}R_n'\sqrt{n}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(\mathbf{0},I_q\right). \tag{56}$$

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

12 Generalized Method of Moments

Suppose that we observe a sequence of random vectors $X_i \in \mathbb{R}^p$, $i = 1, ..., n$ 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

$$\widetilde{m}_g(\boldsymbol{\theta}) = \sum_{j=1}^{n_g} m(X_{gj}, \boldsymbol{\theta}). \tag{57}$$

An unconditional moment model in cluster notation is given by

$$E\widetilde{m}_g(\boldsymbol{\theta}_0) = 0. \tag{58}$$

We assume the $l \geq k$ so the moment model is over-identified. Write the sample mean of the moment function as

$$\overline{m}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} m(X_i, \boldsymbol{\theta}). \tag{59}$$

Since (58) holds for all $g = 1, ..., G$, the usual unconditional moment condition $E\overline{m}_n(\boldsymbol{\theta}_0) = 0$ follows. The generalized method of moments (GMM) estimator is given by

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \Theta} n \cdot \overline{m}_n(\boldsymbol{\theta})' \widehat{W}_n^{-1} \overline{m}_n(\boldsymbol{\theta}) \tag{60}$$

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

$$\widehat{W}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} v(X_i, \boldsymbol{\theta}) v(X_i, \boldsymbol{\theta})' \tag{61}$$

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 (61) 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

$$\begin{aligned}\widehat{W}_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{g=1}^{G}\left(\widetilde{m}_g(\boldsymbol{\theta}) - n_g\overline{m}_n(\boldsymbol{\theta})\right)\left(\widetilde{m}_g(\boldsymbol{\theta}) - n_g\overline{m}_n(\boldsymbol{\theta})\right)' \\ &= \frac{1}{n}\sum_{g=1}^{G}\widetilde{m}_g(\boldsymbol{\theta})\widetilde{m}_g(\boldsymbol{\theta})' - \frac{1}{n}\sum_{g=1}^{G}n_g^2\overline{m}_n(\boldsymbol{\theta})\overline{m}_n(\boldsymbol{\theta})'. \qquad (62)\end{aligned}$$

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.

Theorem 14. *If (2) 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 14 holds given the choice of the weight matrix can be listed, but these are also required to have the asymptotic distribution. For this reason, we do not separately list those conditions. In addition, we assume that the conventional weight matrix is constructed using either $v(X_i, \boldsymbol{\theta}) = m(X_i, \boldsymbol{\theta})$ or $v(X_i, \boldsymbol{\theta}) = m(X_i, \boldsymbol{\theta}) - \overline{m}_n(\boldsymbol{\theta})$ for simplicity.

To show the asymptotic distribution of the GMM estimator, define

$$
\begin{aligned}
Q_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} E\left[\frac{\partial}{\partial \boldsymbol{\theta}'} m(X_i, \boldsymbol{\theta})\right], \\
\Omega_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{g=1}^{G} E\left(\widetilde{m}_g(\boldsymbol{\theta}) - n_g \overline{m}_n(\boldsymbol{\theta})\right)\left(\widetilde{m}_g(\boldsymbol{\theta}) - n_g \overline{m}_n(\boldsymbol{\theta})\right)', \\
V_n &= (Q_n' W_n^{-1} Q_n)^{-1} Q_n' W_n^{-1} \Omega_n W_n^{-1} Q_n (Q_n' W_n^{-1} Q_n)^{-1},
\end{aligned}
$$

where $Q_n = Q_n(\boldsymbol{\theta}_0)$ and $\Omega_n = \Omega_n(\boldsymbol{\theta}_0)$. If the clustered efficient weight matrix (62) is used, then the asymptotic variance matrix simplifies to

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

Define the sample versions as

$$
\begin{aligned}
\widehat{Q}_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}'} m(X_i, \boldsymbol{\theta}), \\
\widehat{\Omega}_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{g=1}^{G} \left(\widetilde{m}_g(\boldsymbol{\theta}) - n_g \overline{m}_n(\boldsymbol{\theta})\right)\left(\widetilde{m}_g(\boldsymbol{\theta}) - n_g \overline{m}_n(\boldsymbol{\theta})\right)'
\end{aligned}
$$

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

$$
\widehat{V}_n = (\widehat{Q}_n' \widehat{W}_n^{-1} \widehat{Q}_n)^{-1} \widehat{Q}_n' \widehat{W}_n^{-1} \widehat{\Omega}_n \widehat{W}_n^{-1} \widehat{Q}_n (\widehat{Q}_n' \widehat{W}_n^{-1} \widehat{Q}_n)^{-1}, \tag{63}
$$

if $\widehat{W_n}$ is given by (61) and

$$
\widehat{V}_n = (\widehat{Q}_n' \widehat{\Omega}_n^{-1} \widehat{Q}_n)^{-1}, \tag{64}
$$

if $\widehat{W_n}$ is given by (62), 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 is not consistent because the conventional efficient weight matrix is not consistent for the inverse of the variance matrix of the moment function. The GMM criterion (60) based on the clustered efficient weight matrix (62) 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 15. *In addition to the assumptions of Theorem 14, if*

1. $\boldsymbol{\theta}_0 \in$ *interior*$(\boldsymbol{\Theta})$,

2. *for some neighborhood* $\mathcal{N}$ *of* $\boldsymbol{\theta}_0$,

(a) $m(x, \boldsymbol{\theta})$ *is continuously differentiable with probability approaching one,*

(b) either X_i *have identical marginal distributions with* $E \sup_{\boldsymbol{\theta} \in \mathcal{N}} \|m(X_i, \boldsymbol{\theta})\|^r < \infty$ *for some* $r \geq 2$ *and* $E \sup_{\boldsymbol{\theta} \in \mathcal{N}} \left\| \frac{\partial}{\partial \boldsymbol{\theta}'} m(X_i, \boldsymbol{\theta}) \right\|^2 < \infty$*;*
or $E \sup_i \sup_{\boldsymbol{\theta} \in \mathcal{N}} \|m(X_i, \boldsymbol{\theta})\|^r < \infty$ *and* $E \sup_i \sup_{\boldsymbol{\theta} \in \mathcal{N}} \left\| \frac{\partial}{\partial \boldsymbol{\theta}'} m(X_i, \boldsymbol{\theta}) \right\|^r < \infty$ *for some* $r > 2$*,*

(c) for each $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{N}$

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

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

(d) $\lambda_{\min}(W_n(\boldsymbol{\theta})) \geq C > 0$*,*

(e) $\lambda_{\min}(\Omega_n(\boldsymbol{\theta})) \geq \lambda > 0$*,*

3. Q_n *is full column rank,*

4. (6) holds for the same r *as in Condition 2(b),*

5. (7) holds,

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

$$\left(R_n' V_n R_n\right)^{-1/2} R_n' \sqrt{n} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(\mathbf{0}, I_q\right), \tag{65}$$

$$\left(R_n' V_n R_n\right)^{-1/2} R_n' \widehat{V}_n R_n \left(R_n' V_n R_n\right)^{-1/2} \xrightarrow{p} I_q, \tag{66}$$

$$\left(R_n' \widehat{V}_n R_n\right)^{-1/2} R_n' \sqrt{n} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(\mathbf{0}, I_q\right), \tag{67}$$

and

$$J_n(\widehat{\boldsymbol{\theta}}) \xrightarrow{d} \chi^2_{l-k}. \tag{68}$$

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

13 Appendix

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*

$$\lim_{B\to\infty} \sup_i E\left(\|X_i\|^r 1\left(\|X_i\| > B\right)\right) = 0, \tag{69}$$

then

$$\lim_{B\to\infty} \sup_m E\left(\left\|m^{-1}\widetilde{X}_m\right\|^r 1\left(\left\|m^{-1}\widetilde{X}_m\right\| > B\right)\right) = 0. \tag{70}$$

Proof of Lemma 1: The proof is based on the proof of Theorem 1 of Etemadi (2006). Equation (69) implies that $\sup_i E\|X_i\|^r \leq C$ for some $C < \infty$. By the C_r inequality

$$\left\|m^{-1}\widetilde{X}_m\right\|^r = \frac{1}{m^r}\left\|\sum_{i=1}^m X_i\right\|^r \leq \frac{1}{m}\sum_{i=1}^m \|X_i\|^r \tag{71}$$

and hence

$$E\left\|m^{-1}\widetilde{X}_m\right\|^r \leq C. \tag{72}$$

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

$$\sup_i E\left(\|X_i\|^r 1\left(\|X_i\| > \sqrt{B}\right)\right) \leq \varepsilon, \tag{73}$$

which is feasible under (69). Using (71),

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

by (73), Markov's inequality, (72), and $B^{r/2} \geq C/\varepsilon$. Since ε is arbitrary this implies (70). ■

The next Lemma is useful to establish 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 (69) holds then*

$$\lim_{B\to\infty} \sup_m E\left(\left\|m^{-2}\left(\widetilde{f}_m - E\widetilde{f}_m\right)\right\|^{r/2} 1\left(\left\|m^{-2}\left(\widetilde{f}_m - E\widetilde{f}_m\right)\right\| > B\right)\right) = 0. \tag{74}$$

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 (72),

$$\begin{aligned}
\left\|m^{-2}\left(\widetilde{f}_m - E\widetilde{f}_m\right)\right\|^{r/2} &\leq \left(\left\|m^{-2}\widetilde{f}_m\right\| + \left\|m^{-2}E\widetilde{f}_m\right\|\right)^{r/2} &(75)\\
&\leq 2^{r/2-1}\left(\left\|m^{-2}\widetilde{f}_m\right\|^{r/2} + E\left\|m^{-2}\widetilde{f}_m\right\|^{r/2}\right) &(76)\\
&\leq 2^{r/2-1}\left(\left\|m^{-1}\widetilde{X}_m\right\|^{r} + E\left\|m^{-1}\widetilde{X}_m\right\|^{r}\right) &(77)\\
&\leq 2^{r/2-1}\left(\left\|m^{-1}\widetilde{X}_m\right\|^{r} + C\right). &(78)
\end{aligned}$$

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

$$\sup_i E\left(\|X_i\|^r 1\left(\|X_i\| > B^{1/4}\right)\right) \leq \frac{\varepsilon}{2^{r/2-1}}, \tag{79}$$

which is feasible under (69). Using (78) and (71),

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

by (79), Markov's inequality, (72), and $2^{r-1}(B^{r/4}+C)C/B^{r/2} \leq \varepsilon$ using the discriminant. Since ε is arbitrary this implies (74).

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, (71), and (72),

$$
\begin{aligned}
E\left\|m^{-1}\left(\widetilde{X}_m - m\overline{X}_n\right)\right\|^r &= E\left\|m^{-1}\sum_{i=1}^{m} X_i - n^{-1}\sum_{i=1}^{n} X_i\right\|^r && (80) \\
&\leq E\left(\left\|m^{-1}\sum_{i=1}^{m} X_i\right\| + \left\|n^{-1}\sum_{i=1}^{n} X_i\right\|\right)^r && (81) \\
&\leq 2^r C && (82)
\end{aligned}
$$

and

$$
\begin{aligned}
\left\|m^{-2}\left(\widetilde{f}_m - E\widetilde{f}_m\right)\right\|^{r/2} &\leq \left(\left\|m^{-2}\widetilde{f}_m\right\| + \left\|m^{-2}E\widetilde{f}_m\right\|\right)^{r/2} \\
&\leq 2^{r/2-1}\left(\left\|m^{-1}\left(\widetilde{X}_m - m\overline{X}_n\right)\right\|^r + E\left\|m^{-1}\left(\widetilde{X}_m - m\overline{X}_n\right)\right\|^r\right) \\
&\leq 2^{3r/2-1}\left(2^{-1}\left(m^{-1}\sum_{i=1}^{m}\|X\|^r + n^{-1}\sum_{i=1}^{n}\|X_i\|^r\right) + C\right). && (83)
\end{aligned}
$$

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

$$\sup_i E\left(\|X_i\|^r 1\left(\|X_i\| > B^{1/4}\right)\right) \leq \frac{\varepsilon}{2^{3r/2-1}}, \tag{84}$$

and proceed as above to show (74). This completes the proof. ■

Proof of Theorem 1: Without loss of generality assume $EX_i = 0$. Define the cluster sums $\widetilde{X}_g = \sum_{j=1}^{n_g} X_{gj}$ so that $\overline{X}_n = \frac{1}{n}\sum_{g=1}^{G} \widetilde{X}_g$ where $\widetilde{X}_g$ are mutually independent and mean zero.

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

$$\sup_g E\left\|\left(n_g^{-1}\widetilde{X}_g 1\left(\left\|n_g^{-1}\widetilde{X}_g\right\| > B\right)\right) - E\left(n_g^{-1}\widetilde{X}_g 1\left(\left\|n_g^{-1}\widetilde{X}_g\right\| > B\right)\right)\right\| \leq \varepsilon \tag{85}$$

which is feasible by Lemma 1 with $r = 1$ under (1). Using the triangle inequality, Jensen's inequality and (85),

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

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

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

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

Proof of Theorem 2: Without loss of generality we assume $EX_i = 0$. We start by rewriting the statistic to simplify the analysis. Define

$$X_i^* = \Omega_n^{-1/2} X_i \tag{86}$$

$$\widetilde{X}_g^* = \sum_{j=1}^{n_g} X_{gj}^* \tag{87}$$

$$\overline{X}_n^* = \frac{1}{n} \sum_{g=1}^{G} \widetilde{X}_g^* \tag{88}$$

$$R_n^* = \Omega_n^{1/2} R_n \left(R_n' \Omega_n R_n\right)^{-1/2}. \tag{89}$$

Then

$$\left(R_n' \Omega_n R_n\right)^{-1/2} R_n' \sqrt{n} \overline{X}_n = R_n^{*\prime} \sqrt{n} \overline{X}_n^*$$

where $R_n^{*\prime} R_n^* = I_q$ and $n \text{var}\left(\overline{X}_n^*\right) = I_p$.

It is useful to observe that X_i^* satisfies assumption (5). Indeed, the matrix norm inequality and assumption (8) imply

$$\|X_i^*\| \leq \left\|\Omega_n^{-1/2} X_i\right\| \leq \left\|\Omega_n^{-1/2}\right\| \|X_i\| \leq \lambda^{-1/2} \|X_i\|.$$

Thus

$$\lim_{M \to \infty} \sup_i \left(E \|X_i^*\|^r \mathbb{1}\left(\|X_i^*\| > M\right)\right) \leq \lambda^{-r/2} \lim_{M \to \infty} \sup_i \left(E \|X_i\|^r \mathbb{1}\left(\|X_i\| > \lambda^{1/2} M\right)\right) = 0 \tag{90}$$

as stated.

We apply the multivariate Lindeberg-Feller central limit theorem (e.g. Hansen (2018) Theorem 6.9.2) since $\widetilde{X}_g^*$ are independent but not identically distributed. Since $\text{var}\left(R_n^{*\prime} \sqrt{n} \overline{X}_n^*\right) = I_q$ and $\left\|R_n^{*\prime} \widetilde{X}_g^*\right\| = \left\|\widetilde{X}_g^*\right\|$ a sufficient condition for the CLT (9) is that for all $\varepsilon > 0$

$$\frac{1}{n} \sum_{g=1}^{G} E\left(\left\|\widetilde{X}_g^*\right\|^2 \mathbb{1}\left(\left\|\widetilde{X}_g^*\right\|^2 \geq n\varepsilon\right)\right) \to 0 \tag{91}$$

as $n \to \infty$.

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

$$\sup_g E\left(\left\|n_g^{-1} \widetilde{X}_g^*\right\|^r \mathbb{1}\left(\left\|n_g^{-1} \widetilde{X}_g^*\right\| > B\right)\right) \leq \frac{\delta \varepsilon^{r/2-1}}{C^{r/2}}. \tag{92}$$

which is feasible by Lemma 1 under (90). Pick n large enough so that

$$\max_{g \leq G} \frac{n_g}{n^{1/2}} \leq \frac{\varepsilon^{1/2}}{B} \tag{93}$$

which is feasible by (7). Thus

$$\begin{aligned}
& \frac{1}{n}\sum_{g=1}^{G} E\left(\left\|\widetilde{X}_g^*\right\|^2 1\left(\left\|\widetilde{X}_g^*\right\|^2 \geq n\varepsilon\right)\right) && (94)\\
= & \frac{1}{n}\sum_{g=1}^{G} E\left(\frac{\left\|\widetilde{X}_g^*\right\|^r}{\left\|\widetilde{X}_g^*\right\|^{r-2}} 1\left(\left\|\widetilde{X}_g^*\right\| \geq (n\varepsilon)^{1/2}\right)\right) \\
\leq & \frac{1}{\varepsilon^{r/2-1} n^{r/2}} \sum_{g=1}^{G} E\left(\left\|\widetilde{X}_g^*\right\|^r 1\left(\left\|\widetilde{X}_g^*\right\| \geq (n\varepsilon)^{1/2}\right)\right) \\
\leq & \frac{1}{\varepsilon^{r/2-1} n^{r/2}} \sum_{g=1}^{G} n_g^r E\left(\left\|n_g^{-1}\widetilde{X}_g^*\right\|^r 1\left(\left\|n_g^{-1}\widetilde{X}_g^*\right\| \geq B\right)\right) \\
\leq & \frac{\delta}{C^{r/2}} \frac{\sum_{g=1}^{G} n_g^r}{n^{r/2}} \\
\leq & \delta.
\end{aligned}$$

The second inequality is (93), the third is (92), and the final is (6). Since ε and δ are arbitrary we have established (91) and hence (9). ■

Proof of Theorem 3: Make the transformations (86)-(89) and define

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

Then (12) is equivalent to

$$\left\|R_n^{*\prime}\widetilde{\Omega}_n^* R_n^* - I_q\right\| = \left\|R_n^{*\prime}\left(\widetilde{\Omega}_n^* - I_p\right) R_n^*\right\| = \left\|\widetilde{\Omega}_n^* - I_p\right\| \xrightarrow{p} 0 \tag{95}$$

the second equality using the fact that $R_n^{*\prime} R_n^* = I_q$. We now show (95).

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

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

By the triangle inequality,

$$\begin{aligned}
E\left\|\widetilde{\Omega}_n^* - I_p\right\| &= \frac{1}{n} E\left\|\sum_{g=1}^{G}\left(\widetilde{Y}_g \widetilde{Y}_g' - E\left(\widetilde{Y}_g \widetilde{Y}_g'\right)\right)\right\| && (96)\\
&\quad + \frac{2}{n}\sum_{g=1}^{G} E\left(\left\|\widetilde{X}_g^*\right\|^2 1\left(\left\|\widetilde{X}_g^*\right\|^2 > n\varepsilon\right)\right). && (97)
\end{aligned}$$

In (94) it was shown that for n sufficiently large (97) is bounded by 2δ. We now consider (96).

Using Jensen's inequality, the assumption that the clusters are independent and thus uncorrelated, the bounds $\left\|\widetilde{Y}_g\widetilde{Y}_g'\right\| \leq n\varepsilon$ and $\left\|\widetilde{Y}_g\widetilde{Y}_g'\right\| \leq \left\|\widetilde{X}_g^*\right\|^2$, and $\varepsilon = \delta^2/4p$, we obtain

$$\begin{aligned}
\frac{1}{n}E\left\|\sum_{g=1}^{G}\left(\widetilde{Y}_g\widetilde{Y}_g' - E\left(\widetilde{Y}_g\widetilde{Y}_g'\right)\right)\right\| &\leq \frac{1}{n}\left(E\left\|\sum_{g=1}^{G}\left(\widetilde{Y}_g\widetilde{Y}_g' - E\left(\widetilde{Y}_g\widetilde{Y}_g'\right)\right)\right\|^2\right)^{1/2} \\
&\leq \frac{1}{n}\left(\sum_{g=1}^{G}E\left\|\widetilde{Y}_g\widetilde{Y}_g' - E\left(\widetilde{Y}_g\widetilde{Y}_g'\right)\right\|^2\right)^{1/2} \\
&\leq 2\varepsilon^{1/2}\left(\frac{1}{n}\sum_{g=1}^{G}E\left\|\widetilde{X}_g^*\right\|^2\right)^{1/2} \\
&\leq 2\varepsilon^{1/2}\left(\mathrm{tr}\left(n\mathrm{var}\left(\overline{X}_n^*\right)\right)\right)^{1/2} \\
&= 2\varepsilon^{1/2}\left(\mathrm{tr}I_p\right)^{1/2} \\
&= \delta. \qquad (98)
\end{aligned}$$

Together, we have shown that for n sufficiently large,

$$E\left\|\widetilde{\Omega}_n^* - I_p\right\| \leq 3\delta \qquad (99)$$

and hence (95) by Markov's Inequality.

By the continuous mapping theorem

$$\left(R_n^{*\prime}\widetilde{\Omega}_n^*R_n^*\right)^{-1/2} \xrightarrow{p} I_q^{-1/2} = I_q.$$

Note that

$$R_n^{*\prime}\widetilde{\Omega}_n^*R_n^* = \left(R_n'\Omega_nR_n\right)^{-1/2}\left(R_n'\widetilde{\Omega}_nR_n\right)\left(R_n'\Omega_nR_n\right)^{-1/2}.$$

Combined with Theorem 2 we find

$$\begin{aligned}
&\left(R_n'\widetilde{\Omega}_nR_n\right)^{-1/2}R_n'\sqrt{n}\overline{X}_n \\
=\; &\left(R_n'\widetilde{\Omega}_nR_n\right)^{-1/2}\left(R_n'\Omega_nR_n\right)^{1/2}\left(R_n'\Omega_nR_n\right)^{-1/2}R_n'\sqrt{n}\overline{X}_n \\
=\; &\left(R_n'\Omega_nR_n\right)^{-1/4}\left(R_n^{*\prime}\widetilde{\Omega}_n^*R_n^*\right)^{-1/2}\left(R_n'\Omega_nR_n\right)^{1/4}\left(R_n'\Omega_nR_n\right)^{-1/2}R_n'\sqrt{n}\overline{X}_n \\
\xrightarrow{d}\; &N\left(\mathbf{0}, I_q\right)
\end{aligned}$$

because

$$\left\|\left(R_n'\Omega_nR_n\right)^{-1/4}\left(R_n^{*\prime}\widetilde{\Omega}_n^*R_n^*\right)^{-1/2}\left(R_n'\Omega_nR_n\right)^{1/4} - I_q\right\| = \left\|\left(R_n^{*\prime}\widetilde{\Omega}_n^*R_n^*\right)^{-1/2} - I_q\right\| \xrightarrow{p} 0.$$

This is (13). ■

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 (7),

$$\begin{aligned} & \left\| \left(R_n'\Omega_n R_n\right)^{-1/2} \left(R_n'\widehat{\Omega}_n R_n\right) \left(R_n'\Omega_n R_n\right)^{-1/2} - I_q \right\| \\ \leq & \left\| \left(R_n'\Omega_n R_n\right)^{-1/2} \left(R_n'\widetilde{\Omega}_n R_n\right) \left(R_n'\Omega_n R_n\right)^{-1/2} - I_q \right\| \\ & + \left(\frac{1}{n^2}\sum_{g=1}^{G} n_g^2\right) \left\| \left(R_n'\Omega_n R_n\right)^{-1/2} R_n'\sqrt{n}\overline{X}_n \right\|^2 \\ \leq & \ o_p(1). \end{aligned}$$

This is (14). Equation (15) follows as in the proof of (13). ■

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 (18) 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 $\theta_1, \theta_2 \in \Theta$,

$$\left\| \widetilde{f}_g(\theta_1) - \widetilde{f}_g(\theta_2) \right\| \leq A_g h\left(\|\theta_1 - \theta_2\|\right) \tag{100}$$

where $h(u) \downarrow 0$ as $u \downarrow 0$ and $\frac{1}{n}\sum_{g=1}^{G} E\left(A_g\right) \leq A < \infty$. The total boundedness condition holds by assumption and the WLLN holds by Theorem 1 under (2) and (16), so it only remains to establish the Lipschitz condition (100). Indeed, using the triangle inequality and (17)

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

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 (100) and hence (18) holds. ■

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

We first examine the case with no estimated mean (21). Define $\Omega_n^* = a_n \Omega_n(\theta)$ and $\widetilde{\Omega}_n^* = a_n \widetilde{\Omega}_n$ where

$$a_n = \left(\frac{\left(\sum_{g=1}^{G} n_g^r\right)^{2/r}}{n}\right)^{r/2} \frac{n}{\sum_{g=1}^{G} n_g^2}. \tag{101}$$

Note that $a_n n \to \infty$ and $a_n/n \to 0$. In addition,

$$a_n = a_n \frac{\sum_{g=1}^{G} n_g}{n} < a_n \frac{\sum_{g=1}^{G} n_g^2}{n} < \infty. \tag{102}$$

Set $W_g(\theta) = \sum_{j=1}^{n_g}\sum_{\ell=1}^{n_g} f(X_{gj},\theta) f(X_{g\ell},\theta)'$ so that $\widetilde{\Omega}_n^*(\theta) = \frac{a_n}{n}\sum_{g=1}^{G} W_g(\theta)$. Then, showing

$$\sup_{\theta\in\Theta}\left\|\Omega_n^*(\theta)^{-1/2}\widetilde{\Omega}_n^*(\theta)\Omega_n^*(\theta)^{-1/2} - I_k\right\| \to_p 0 \tag{103}$$

is equivalent to showing (21).

We check the condition of Theorem 3 of Andrews (1992). Notice that Θ is totally bounded and for each θ,

$$\left\|\Omega_n^*(\theta)^{-1/2}\widetilde{\Omega}_n^*(\theta)\Omega_n^*(\theta)^{-1/2} - I_k\right\| = \left\|\Omega_n(\theta)^{-1/2}\widetilde{\Omega}_n(\theta)\Omega_n(\theta)^{-1/2} - I_k\right\| \to_p 0$$

by Theorem 3 under the assumptions. Using Theorem 3 of Andrews (1992) it is sufficient to verify that

$$\|W_g(\theta_1) - W_g(\theta_2)\| \leq A_g h\left(\|\theta_1 - \theta_2\|\right) \tag{104}$$

with $\sup_{n\geq 1}\frac{a_n}{n}\sum_{g=1}^{G} E\left(A_g\right) \leq A < \infty$ under the assumption that $\lambda_{\min}(\Omega_n(\theta)) \geq \lambda > 0$ for each θ.

Indeed

$$\begin{aligned}
\|W_g(\theta_2) - W_g(\theta_1)\| &\leq \sum_{j=1}^{n_g}\sum_{\ell=1}^{n_g} \|f(X_{gj},\theta_2) - f(X_{gj},\theta_1)\| \, \|f(X_{g\ell},\theta_2) - f(X_{g\ell},\theta_1)\| \\
&\quad + \sum_{j=1}^{n_g}\sum_{\ell=1}^{n_g} \|f(X_{gj},\theta_2) - f(X_{gj},\theta_1)\| \, \|f(X_{g\ell},\theta_1)\| \\
&\quad + \sum_{j=1}^{n_g}\sum_{\ell=1}^{n_g} \|f(X_{gj},\theta_1)\| \, \|f(X_{g\ell},\theta_2) - f(X_{g\ell},\theta_1)\| \\
&\leq \sum_{j=1}^{n_g}\sum_{\ell=1}^{n_g} A(X_{gj})A(X_{g\ell}) h\left(\|\theta_1 - \theta_2\|\right)^2 \\
&\quad + \sum_{j=1}^{n_g}\sum_{\ell=1}^{n_g} A(X_{gj}) \, \|f(X_{g\ell},\theta_1)\| \, h\left(\|\theta_1 - \theta_2\|\right) \\
&\quad + \sum_{j=1}^{n_g}\sum_{\ell=1}^{n_g} A(X_{g\ell}) \, \|f(X_{gj},\theta_1)\| \, h\left(\|\theta_1 - \theta_2\|\right) \\
&\leq A_g h\left(\|\theta_1 - \theta_2\|\right) \qquad (105)
\end{aligned}$$

where

$$A_g = \sum_{j=1}^{n_g}\sum_{\ell=1}^{n_g} \left(A(X_{gj})A(X_{g\ell})\overline{h} + 2A(X_{gj}) \, \|f(X_{g\ell},\theta_1)\|\right)$$

and $\overline{h} = \sup_{\theta_1,\theta_2\in\Theta} h\left(\|\theta_1 - \theta_2\|\right)$. Recall that $\sup_i EA(X_i)^2 \leq C$, and $\sup_i E\,\|f(X_i,\theta_1)\|^r \leq C$ for $r > 2$ by assumption. This implies

$$\begin{aligned}
\frac{a_n}{n}\sum_{g=1}^{G} E\left(A_g\right) &= \frac{a_n}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g}\sum_{\ell=1}^{n_g} \left(E\left(A(X_{gj})A(X_{g\ell})\right)\overline{h} + 2E\left(A(X_{gj}) \, \|f(X_{g\ell},\theta_1)\|\right)\right) \\
&\leq \left(\frac{\left(\sum_{g=1}^{G} n_g^r\right)^{2/r}}{n}\right)^{r/2} \left(2 + \overline{h}\right) C, \qquad (106)
\end{aligned}$$

by Cauchy-Schwarz and Hölder's inequalities. This verifies (104) and proves (21).

Next we show (20). Set

$$W_g(\theta) = \left(\sum_{j=1}^{n_g} f(X_{gj},\theta) - n_g\overline{f}_n(\theta)\right)\left(\sum_{j=1}^{n_g} f(X_{gj},\theta) - n_g\overline{f}_n(\theta)\right)'$$

so that $\widehat{\Omega}_n^*(\theta) = \frac{a_n}{n}\sum_{g=1}^{G} W_g(\theta)$. Notice that Θ is totally bounded and

$$\left\|\Omega_n^*(\theta)^{-1/2}\widehat{\Omega}_n^*(\theta)\Omega_n^*(\theta)^{-1/2} - I_k\right\| = \left\|\Omega_n(\theta)^{-1/2}\widehat{\Omega}_n(\theta)\Omega_n(\theta)^{-1/2} - I_k\right\| \to_p 0$$

by Theorem 4 under the assumptions. Using Theorem 3 of Andrews (1992) it is sufficient to verify that

$$\|W_g(\theta_1) - W_g(\theta_2)\| \leq (A_g + A_h + A_f)\, h\left(\|\theta_1 - \theta_2\|\right) \tag{107}$$

with $\frac{a_n}{n}\sum_{g=1}^{G} E\left(A_g + A_h + A_f\right) \leq A^* < \infty$.

Since we can write

$$W_g(\theta) = \sum_{j=1}^{n_g}\sum_{l=1}^{n_g} f(X_{gj},\theta)f(X_{gl},\theta)' - \sum_{j=1}^{n_g} f(X_{gj},\theta)n_g\overline{f}_n(\theta)' - n_g\overline{f}_n(\theta)\sum_{j=1}^{n_g} f(X_{gj},\theta)' + n_g^2\overline{f}_n(\theta)\overline{f}_n(\theta)',$$

by the triangle inequality

$$\begin{aligned}
\|W_g(\theta_2) - W_g(\theta_1)\| \leq\ & \sum_{j=1}^{n_g}\sum_{l=1}^{n_g}\left\|f(X_{gj},\theta_2)f(X_{gl},\theta_2)' - f(X_{gj},\theta_1)f(X_{gl},\theta_1)'\right\| && (108)\\
& + n_g\sum_{j=1}^{n_g}\left\|f(X_{gj},\theta_2)\overline{f}_n(\theta_2)' - f(X_{gj},\theta_1)\overline{f}_n(\theta_1)'\right\| && (109)\\
& + n_g\sum_{j=1}^{n_g}\left\|\overline{f}_n(\theta_2)f(X_{gj},\theta_2)' - \overline{f}_n(\theta_1)f(X_{gj},\theta_1)'\right\| && (110)\\
& + n_g^2\left\|\overline{f}_n(\theta_2)\overline{f}_n(\theta_2)' - \overline{f}_n(\theta_1)\overline{f}_n(\theta_1)'\right\| && (111)
\end{aligned}$$

First, take the RHS of (108). By (105),

$$\sum_{j=1}^{n_g}\sum_{l=1}^{n_g}\left\|f(X_{gj},\theta_2)f(X_{gl},\theta_2)' - f(X_{gj},\theta_1)f(X_{gl},\theta_1)'\right\| \leq A_g h\left(\|\theta_1 - \theta_2\|\right). \tag{112}$$

Next, take (109). Since

$$\begin{aligned}
\left\|\overline{f}_n(\theta_2) - \overline{f}_n(\theta_1)\right\| &\leq \frac{1}{n}\sum_{i=1}^{n}\|f(X_i,\theta_2) - f(X_i,\theta_1)\| \\
&\leq \overline{A}_n h\left(\|\theta_2 - \theta_1\|\right) && (113)
\end{aligned}$$

where $\overline{A}_n = n^{-1}\sum_{i=1}^{n} A(X_i)$, by the triangle inequality

$$\begin{aligned}
& \left\|f(X_{gj},\theta_2)\overline{f}_n(\theta_2)' - f(X_{gj},\theta_1)\overline{f}_n(\theta_1)'\right\| \\
\leq\ & \left\|\left(f(X_{gj},\theta_2) - f(X_{gj},\theta_1)\right)\left(\overline{f}_n(\theta_2) + \overline{f}_n(\theta_1)\right)\right\| \\
& + \left\|f(X_{gj},\theta_2)\left(\overline{f}_n(\theta_2) - \overline{f}_n(\theta_1)\right) - \left(f(X_{gj},\theta_2) - f(X_{gj},\theta_1)\right)\overline{f}_n(\theta_2)\right\| \\
\leq\ & \left\|\left(f(X_{gj},\theta_2) - f(X_{gj},\theta_1)\right)\right\|\left(\left\|\overline{f}_n(\theta_1)\right\| + 2\left\|\overline{f}_n(\theta_2)\right\|\right) \\
& + \|f(X_{gj},\theta_2)\|\frac{1}{n}\sum_{i=1}^{n}\|f(X_i,\theta_2) - f(X_i,\theta_1)\| \\
\leq\ & \left(A(X_{gj})\left(\left\|\overline{f}_n(\theta_1)\right\| + 2\left\|\overline{f}_n(\theta_2)\right\|\right) + \|f(X_{gj},\theta_2)\|\,\overline{A}_n\right) h\left(\|\theta_2 - \theta_1\|\right).
\end{aligned}$$

Thus,

$$n_g \sum_{j=1}^{n_g} \left\| f(X_{gj},\theta_2)\overline{f}_n(\theta_2)' - f(X_{gj},\theta_1)\overline{f}_n(\theta_1)' \right\| \le A_h h\left(\|\theta_2-\theta_1\|\right). \tag{114}$$

where

$$A_h = n_g \sum_{j=1}^{n_g} \left(A(X_{gj}) \left(\left\|\overline{f}_n(\theta_1)\right\| + 2\left\|\overline{f}_n(\theta_2)\right\| \right) + \|f(X_{gj},\theta_2)\| \overline{A}_n \right).$$

(110) has the identical bound to (114).

Lastly, take (111).

$$\begin{aligned} & n_g^2 \left\|\overline{f}_n(\theta_2)\overline{f}_n(\theta_2)' - \overline{f}_n(\theta_1)\overline{f}_n(\theta_1)'\right\| \\ \le\ & n_g^2 \left\|\overline{f}_n(\theta_2) - \overline{f}_n(\theta_1)\right\|^2 + 2n_g \left\|\overline{f}_n(\theta_2) - \overline{f}_n(\theta_1)\right\| \cdot \left\|\overline{f}_n(\theta_1)\right\| \\ \le\ & A_f h\left(\|\theta_2-\theta_1\|\right) \end{aligned} \tag{115}$$

where

$$A_f = n_g^2 \overline{A}_n^2 h\left(\|\theta_2-\theta_1\|\right) + 2n_g \overline{A}_n \left\|\overline{f}_n(\theta_1)\right\|.$$

By (112), (114), and (115),

$$\|W_g(\theta_2) - W_g(\theta_1)\| \le \left(A_g + 2A_h + A_f\right) h\left(\|\theta_2-\theta_1\|\right). \tag{116}$$

By Hölder's inequality,

$$\begin{aligned} \frac{a_n}{n}\sum_{g=1}^{G} E(A_h) \le\ & \frac{a_n}{n}\sum_{g=1}^{G} n_g \sum_{j=1}^{n_g} \left(\frac{3}{n}\sum_{i=1}^{n} \left(EA(X_{gj})^{r/(r-1)}\right)^{(r-1)/r} \left(\sup_{\theta\in\Theta} E\|f(X_i,\theta)\|^r\right)^{1/r} \right. \\ & \left. + \frac{1}{n}\sum_{i=1}^{n} \left(EA(X_i)^{r/(r-1)}\right)^{(r-1)/r} \left(\sup_{\theta\in\Theta} E\|f(X_{gj},\theta)\|^r\right)^{1/r} \right) \\ \le\ & \left(\frac{\left(\sum_{g=1}^{G} n_g^r\right)^{2/r}}{n} \right)^{r/2} 4C < \infty. \end{aligned} \tag{117}$$

By (102), Jensen's, Cauchy-Schwarz, and Hölder's inequalities,

$$\begin{aligned} \frac{a_n}{n}\sum_{g=1}^{G} E(A_f) \le\ & \frac{a_n}{n}\sum_{g=1}^{G} \left(\frac{n_g^2 \overline{h}}{n}\sum_{i=1}^{n} EA(X_i)^2 \right. \\ & \left. + \frac{2n_g}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \left(EA(X_i)^{r/(r-1)}\right)^{(r-1)/r} \left(\sup_{\theta\in\Theta} E\|f(X_j,\theta_1)\|^r\right)^{1/r} \right) \\ \le\ & \left(\frac{\left(\sum_{g=1}^{G} n_g^r\right)^{2/r}}{n} \right)^{r/2} C\overline{h} + 2a_n C < \infty. \end{aligned} \tag{118}$$

Combining (106), (117), and (118),

$$\frac{a_n}{n}\sum_{g=1}^{G} E(A_g + 2A_h + A_f) \leq A^* < \infty. \tag{119}$$

This verifies (107) and proves (20). ■

Proof of Theorem 7: Define

$$\begin{aligned} \widetilde{f}_g^* &= \Omega_n^{-1/2}\widetilde{f}_g & (120)\\ \overline{f}_G^* &= \frac{1}{n}\sum_{g=1}^{G}\widetilde{f}_g^* & (121)\\ R_n^* &= \Omega_n^{1/2}R_n\left(R_n'\Omega_n R_n\right)^{-1/2}. & (122) \end{aligned}$$

Then

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

where $R_n^{*\prime}R_n^* = I_q$ and $n\text{var}\left(\overline{f}_G^*\right) = I_p$.

Since $\widetilde{f}_g^*$ are independent but not identically distributed, we apply the multivariate Lindeberg-Feller central limit theorem (e.g. Hansen (2018) Theorem 6.9.2). Since var $\left(R_n^{*\prime}\sqrt{n}\overline{f}_G^*\right) = I_q$ and $\left\|R_n^{*\prime}\widetilde{f}_g^*\right\| = \left\|\widetilde{f}_g^*\right\|$ a sufficient condition for the CLT (29) is that for all $\varepsilon > 0$

$$\begin{aligned} &\frac{1}{n}\sum_{g=1}^{G} E\left(\left\|\widetilde{f}_g^* - E\widetilde{f}_g^*\right\|^2 1\left(\left\|\widetilde{f}_g^* - E\widetilde{f}_g^*\right\|^2 \geq n\varepsilon\right)\right)\\ \leq\ &\frac{1}{n\lambda}\sum_{g=1}^{G} E\left(\left\|\widetilde{f}_g - E\widetilde{f}_g\right\|^2 1\left(\left\|\widetilde{f}_g - E\widetilde{f}_g\right\|^2 \geq n\varepsilon\lambda\right)\right) \to 0 \end{aligned} \tag{123}$$

as $n \to \infty$.

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

$$\sup_g E\left(\left\|n_g^{-2}\left(\widetilde{f}_g - E\widetilde{f}_g\right)\right\|^r 1\left(\left\|n_g^{-2}\left(\widetilde{f}_g - E\widetilde{f}_g\right)\right\| > B\right)\right) \leq \frac{\delta\varepsilon^{r/2-1}\lambda^{r/2}}{C^{r/2}}. \tag{124}$$

which is feasible by Lemma 2 under (25). Pick n large enough so that

$$\max_{g\leq G}\frac{n_g^2}{n^{1/2}} \leq \frac{(\varepsilon\lambda)^{1/2}}{B} \tag{125}$$

which is feasible by (27). Thus

$$\begin{aligned}
&\frac{1}{n\lambda}\sum_{g=1}^{G} E\left(\left\|\widetilde{f}_g - E\widetilde{f}_g\right\|^2 1\left(\left\|\widetilde{f}_g - E\widetilde{f}_g\right\|^2 \geq n\varepsilon\lambda\right)\right) \qquad (126)\\
= \;& \frac{1}{n\lambda}\sum_{g=1}^{G} E\left(\frac{\left\|\widetilde{f}_g - E\widetilde{f}_g\right\|^r}{\left\|\widetilde{f}_g - E\widetilde{f}_g\right\|^{r-2}} 1\left(\left\|\widetilde{f}_g - E\widetilde{f}_g\right\| \geq (n\varepsilon\lambda)^{1/2}\right)\right)\\
\leq \;& \frac{1}{\varepsilon^{r/2-1} n^{r/2}\lambda^{r/2}}\sum_{g=1}^{G} E\left(\left\|\widetilde{f}_g - E\widetilde{f}_g\right\|^r 1\left(\left\|\widetilde{f}_g - E\widetilde{f}_g\right\| \geq (n\varepsilon\lambda)^{1/2}\right)\right)\\
\leq \;& \frac{1}{\varepsilon^{r/2-1} n^{r/2}\lambda^{r/2}}\sum_{g=1}^{G} n_g^{2r} E\left(\left\|n_g^{-2}\left(\widetilde{f}_g - E\widetilde{f}_g\right)\right\|^r 1\left(\left\|n_g^{-2}\left(\widetilde{f}_g - E\widetilde{f}_g\right)\right\| \geq B\right)\right)\\
\leq \;& \frac{\delta}{C^{r/2}}\frac{\sum_{g=1}^{G} n_g^{2r}}{n^{r/2}}\\
\leq \;& \delta.
\end{aligned}$$

The second inequality is (125), the third is (124), and the final is (26). Since ε and δ are arbitrary we have established (123) and hence (29). ■

Proof of Theorem 8: We can write

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \widehat{Q}_n^{-1}\widehat{S}_n$$

where

$$\widehat{Q}_n = \frac{1}{n}\sum_{g=1}^{G} \boldsymbol{X}_g'\boldsymbol{X}_g = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'$$

and

$$\widehat{S}_n = \frac{1}{n}\sum_{g=1}^{G} \boldsymbol{X}_g'\boldsymbol{e}_g.$$

The random variables $(\boldsymbol{x}_i\boldsymbol{x}_i', \boldsymbol{x}_i e_i)$ are uniformly integrable and thus satisfy the conditions for Theorem 1. Hence

$$\left\|\widehat{S}_n\right\| \xrightarrow{p} 0$$

and

$$\left\|\widehat{Q}_n - Q_n\right\| \xrightarrow{p} 0.$$

The latter, plus $\lambda_{\min}(Q_n) \geq C$, imply that

$$\left\|Q_n^{-1/2}\widehat{Q}_n Q_n^{-1/2} - I_k\right\| = \left\|Q_n^{-1/2}\left(\widehat{Q}_n - Q_n\right)Q_n^{-1/2}\right\| \leq C^{-1}\left\|\widehat{Q}_n - Q_n\right\| \xrightarrow{p} 0.$$

Applying the continuous mapping theorem

$$\left\|Q_n^{1/2}\widehat{Q}_n^{-1}Q_n^{1/2} - I_k\right\| \xrightarrow{p} 0. \tag{127}$$

It follows that

$$\begin{aligned}\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\| &= \left\|Q_n^{-1/2}Q_n^{1/2}\widehat{Q}_n^{-1}Q_n^{1/2}Q_n^{-1/2}\widehat{S}_n\right\| \\ &\leq \left(\left\|Q_n^{-1}\right\| + \left\|Q_n^{1/2}\widehat{Q}_n^{-1}Q_n^{1/2} - I_k\right\|\right)\left\|\widehat{S}_n\right\| \xrightarrow{p} 0\end{aligned}$$

as stated. ■

Proof of Theorem 9: We start by defining

$$R_n^* = V_n^{1/2}R_n\left(R_n'V_nR_n\right)^{-1/2}$$

so that

$$\begin{aligned}\left(R_n'V_nR_n\right)^{-1/2}R_n'\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) &= R_n^{*\prime}V_n^{-1/2}Q_n^{-1}\sqrt{n}\widehat{S}_n \qquad (128) \\ &\quad + R_n^{*\prime}V_n^{-1/2}\left(\widehat{Q}_n^{-1} - Q_n^{-1}\right)\sqrt{n}\widehat{S}_n. \qquad (129)\end{aligned}$$

Notice that $\boldsymbol{x}_ie_i$ is uniformly integrable under the assumptions, and $n\text{var}\left(V_n^{-1/2}Q_n^{-1}\widehat{S}_n\right) = I_k$. So Theorem 2 can be applied to find

$$V_n^{-1/2}Q_n^{-1}\sqrt{n}\widehat{S}_n \xrightarrow{d} N\left(\mathbf{0}, I_k\right) \tag{130}$$

and (128) satisfies

$$R_n^{*\prime}V_n^{-1/2}Q_n^{-1}\sqrt{n}\widehat{S}_n \xrightarrow{d} N\left(\mathbf{0}, I_q\right). \tag{131}$$

Take (129). Using the fact that $\|AB\| = \|BA\|$ for positive definite A and B, and the matrix norm inequality, (130) and (127), (129) equals

$$\begin{aligned}&\left\|R_n^{*\prime}V_n^{-1/2}\left(\widehat{Q}_n^{-1} - Q_n^{-1}\right)Q_nV_n^{1/2}V_n^{-1/2}Q_n^{-1}\sqrt{n}\widehat{S}_n\right\| \\ \leq\ &\left\|R_n^{*\prime}V_n^{-1/2}Q_n^{-1/2}\left(Q_n^{1/2}\widehat{Q}_n^{-1}Q_n^{1/2} - I_k\right)Q_n^{1/2}V_n^{1/2}\right\|\left\|V_n^{-1/2}Q_n^{-1}\sqrt{n}\widehat{S}_n\right\| \\ \leq\ &\left\|R_n^{*\prime}V_n^{-1/2}Q_n^{-1/2}\left(Q_n^{1/2}\widehat{Q}_n^{-1}Q_n^{1/2} - I_k\right)Q_n^{1/2}V_nQ_n^{1/2}\left(Q_n^{1/2}\widehat{Q}_n^{-1}Q_n^{1/2} - I_k\right)Q_n^{-1/2}V_n^{-1/2}R_n^*\right\|^{1/2}O_p(1) \\ =\ &\left\|V_n^{-1/2}Q_n^{-1/2}\left(Q_n^{1/2}\widehat{Q}_n^{-1}Q_n^{1/2} - I_k\right)Q_n^{1/2}V_nQ_n^{1/2}\left(Q_n^{1/2}\widehat{Q}_n^{-1}Q_n^{1/2} - I_k\right)Q_n^{-1/2}V_n^{-1/2}\right\|^{1/2}O_p(1) \\ \leq\ &o_p(1).\end{aligned}$$

Together we have established (42).

For (43)

$$\begin{aligned}\left\|\left(R_n' V_n R_n\right)^{-1/2} R_n' \widehat{V}_n R_n \left(R_n' V_n R_n\right)^{-1/2} - I_q\right\| &= \left\|R_n^{*\prime}\left(V_n^{-1/2}\widehat{V}_n V_n^{-1/2} - I_k\right) R_n^*\right\| \\ &= \left\|V_n^{-1/2}\widehat{V}_n V_n^{-1/2} - I_k\right\|\end{aligned}$$

so it is equivalent to show

$$\left\|V_n^{-1/2}\widehat{V}_n V_n^{-1/2} - I_k\right\| \xrightarrow{p} 0.$$

Define

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

By the triangle and Cauchy-Schwarz inequalities

$$\begin{aligned}
& \left\|V_n^{-1/2}\widehat{V}_n V_n^{-1/2} - I_k\right\| \\
= & \left\|V_n^{-1/2}\widehat{Q}_n^{-1}\widehat{\Omega}_n\widehat{Q}_n^{-1}V_n^{-1/2} - I_k\right\| \\
\leq & \left\|V_n^{-1/2}Q_n^{-1}\widehat{\Omega}_n Q_n^{-1}V_n^{-1/2} - I_k\right\| + 2\left\|V_n^{-1/2}\left(\widehat{Q}_n^{-1} - Q_n^{-1}\right)\widehat{\Omega}_n Q_n^{-1}V_n^{-1/2}\right\| \\
& + \left\|V_n^{-1/2}\left(\widehat{Q}_n^{-1} - Q_n^{-1}\right)\widehat{\Omega}_n\left(\widehat{Q}_n^{-1} - Q_n^{-1}\right)V_n^{-1/2}\right\| \\
\leq & \left\|V_n^{-1/2}Q_n^{-1}\widehat{\Omega}_n Q_n^{-1}V_n^{-1/2} - I_k\right\| \\
& +2\left\|V_n^{-1/2}\left(\widehat{Q}_n^{-1} - Q_n^{-1}\right)\widehat{\Omega}_n\left(\widehat{Q}_n^{-1} - Q_n^{-1}\right)V_n^{-1/2}\right\|^{1/2}\left\|V_n^{-1/2}Q_n^{-1}\widehat{\Omega}_n Q_n^{-1}V_n^{-1/2}\right\|^{1/2} \\
& + \left\|V_n^{-1/2}\left(\widehat{Q}_n^{-1} - Q_n^{-1}\right)\widehat{\Omega}_n\left(\widehat{Q}_n^{-1} - Q_n^{-1}\right)V_n^{-1/2}\right\|.
\end{aligned}$$

Observe that

$$\begin{aligned}
& \left\|V_n^{-1/2}\left(\widehat{Q}_n^{-1} - Q_n^{-1}\right)\widehat{\Omega}_n\left(\widehat{Q}_n^{-1} - Q_n^{-1}\right)V_n^{-1/2}\right\| \\
= & \left\|\left(Q_n^{1/2}\widehat{Q}_n^{-1}Q_n^{1/2} - I_k\right)Q_n^{-1/2}\widehat{\Omega}_n Q_n^{-1/2}\left(Q_n^{1/2}\widehat{Q}_n^{-1}Q_n^{1/2} - I_k\right)Q_n^{-1/2}V_n^{-1}Q_n^{-1/2}\right\| \\
\leq & \left\|Q_n^{1/2}\widehat{Q}_n^{-1}Q_n^{1/2} - I_k\right\|^2\left\|\Omega_n^{-1/2}\widehat{\Omega}_n\Omega_n^{-1/2}\right\|
\end{aligned}$$

and

$$\begin{aligned}\left\|V_n^{-1/2}Q_n^{-1}\widehat{\Omega}_n Q_n^{-1}V_n^{-1/2} - I_k\right\| &= \left\|V_n^{-1/2}Q_n^{-1}\left(\widehat{\Omega}_n - \Omega_n\right)Q_n^{-1}V_n^{-1/2}\right\| \\ &= \left\|\Omega_n^{-1/2}\widehat{\Omega}_n\Omega_n^{-1/2} - I_k\right\|.\end{aligned}$$

Given (127), it is sufficient to show that

$$\left\|\Omega_n^{-1/2}\widehat{\Omega}_n\Omega_n^{-1/2} - I_k\right\| \xrightarrow{p} 0 \tag{132}$$

to establish (43).

Define

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

and

$$\begin{aligned}\widetilde{\Omega}_n^* &= \frac{1}{n}\sum_{g=1}^{G} \boldsymbol{X}_g' \left(\widehat{\boldsymbol{e}}_g - \boldsymbol{e}_g\right)\left(\widehat{\boldsymbol{e}}_g - \boldsymbol{e}_g\right)' \boldsymbol{X}_g \\ &= \frac{1}{n}\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{X}_g \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \boldsymbol{X}_g' \boldsymbol{X}_g.\end{aligned}$$

By the triangle and Cauchy-Schwarz inequalities

$$\begin{aligned}\left\|\Omega_n^{-1/2}\widehat{\Omega}_n\Omega_n^{-1/2} - I_p\right\| &\le \left\|\Omega_n^{-1/2}\widetilde{\Omega}_n\Omega_n^{-1/2} - I_k\right\| + \left\|\Omega_n^{-1/2}\frac{1}{n}\sum_{g=1}^{G} \boldsymbol{X}_g'\left(\widehat{\boldsymbol{e}}_g\widehat{\boldsymbol{e}}_g' - \boldsymbol{e}_g\boldsymbol{e}_g'\right)\boldsymbol{X}_g\Omega_n^{-1/2}\right\| \\ &\le \left\|\Omega_n^{-1/2}\widetilde{\Omega}_n\Omega_n^{-1/2} - I_k\right\| \\ &\quad +2\left\|\Omega_n^{-1/2}\frac{1}{n}\sum_{g=1}^{G} \boldsymbol{X}_g'\boldsymbol{e}_g\left(\widehat{\boldsymbol{e}}_g - \boldsymbol{e}_g\right)'\boldsymbol{X}_g\Omega_n^{-1/2}\right\| \\ &\quad +\left\|\Omega_n^{-1/2}\frac{1}{n}\sum_{g=1}^{G} \boldsymbol{X}_g'\left(\widehat{\boldsymbol{e}}_g - \boldsymbol{e}_g\right)\left(\widehat{\boldsymbol{e}}_g - \boldsymbol{e}_g\right)'\boldsymbol{X}_g\Omega_n^{-1/2}\right\| \\ &\le \left\|\Omega_n^{-1/2}\widetilde{\Omega}_n\Omega_n^{-1/2} - I_k\right\| + 2\left\|\Omega_n^{-1/2}\widetilde{\Omega}_n\Omega_n^{-1/2}\right\|^{1/2}\left\|\Omega_n^{-1/2}\widetilde{\Omega}_n^*\Omega_n^{-1/2}\right\|^{1/2} \\ &\quad +\left\|\Omega_n^{-1/2}\widetilde{\Omega}_n^*\Omega_n^{-1/2}\right\|. \qquad (133)\end{aligned}$$

Theorem 3 implies that

$$\Omega_n^{-1/2}\widetilde{\Omega}_n\Omega_n^{-1/2} \xrightarrow{p} I_k. \qquad (134)$$

The proof is completed by showing that

$$\left\|\Omega_n^{-1/2}\widetilde{\Omega}_n^*\Omega_n^{-1/2}\right\| \xrightarrow{p} 0. \qquad (135)$$

Note that

$$\begin{aligned}\left\|\Omega_n^{-1/2}\widetilde{\Omega}_n^*\Omega_n^{-1/2}\right\| &= \left\|\Omega_n^{-1/2}\frac{1}{n}\sum_{g=1}^{G}\boldsymbol{X}_g'\boldsymbol{X}_g\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)'\boldsymbol{X}_g'\boldsymbol{X}_g\Omega_n^{-1/2}\right\| \\ &\leq \left\|V_n^{-1/2}\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right\|^2\left\|\frac{1}{n^2}\sum_{g=1}^{G}\boldsymbol{X}_g'\boldsymbol{X}_g\Omega_n^{-1}\boldsymbol{X}_g'\boldsymbol{X}_gV_n\right\| \\ &\leq O_p(1)\frac{1}{n^2}\sum_{g=1}^{G}\left\|\boldsymbol{X}_g'\boldsymbol{X}_g\right\|^2 \\ &= o_p(1)\end{aligned}$$

since by Minkowski's inequality and $E\left\|\boldsymbol{x}_{gj}\right\|^4\leq C$

$$E\left\|\boldsymbol{X}_g'\boldsymbol{X}_g\right\|^2=E\left\|\sum_{j=1}^{n_g}\boldsymbol{x}_{gj}\boldsymbol{x}_{gj}'\right\|^2\leq\left(\sum_{j=1}^{n_g}\left(E\left\|\boldsymbol{x}_{gj}\boldsymbol{x}_{gj}'\right\|^2\right)^{1/2}\right)^2\leq Cn_g^2$$

so

$$\frac{1}{n^2}\sum_{g=1}^{G}E\left\|\boldsymbol{X}_g'\boldsymbol{X}_g\right\|^2\leq\frac{C}{n^2}\sum_{g=1}^{G}n_g^2=o(1)$$

by (7). This shows (135). Equations (134) and (135) establish (132) and hence (43). (44) follows as in the proof of (13). ■

Proof of Theorem 10: Write

$$\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}=\left(\widehat{Q}_n'\widehat{W}_n^{-1}\widehat{Q}_n\right)^{-1}\widehat{Q}_n'\widehat{W}_n^{-1}\widehat{S}_n$$

where

$$\widehat{S}_n = \frac{1}{n}\sum_{g=1}^{G}\boldsymbol{Z}_g'\boldsymbol{e}_g=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_ie_i.$$

The random variables $(\boldsymbol{z}_i\boldsymbol{x}_i',\boldsymbol{z}_i\boldsymbol{z}_i',\boldsymbol{z}_ie_i)$ are uniformly integrable under the assumptions. By Theorem 1

$$\left\|\widehat{S}_n\right\| \xrightarrow{p} 0, \quad (136)$$

$$\left\|\widehat{Q}_n-Q_n\right\| \xrightarrow{p} 0, \quad (137)$$

$$\left\|\widehat{W}_n-W_n\right\| \xrightarrow{p} 0. \quad (138)$$

We first show

$$\left\|\widehat{W}_n^{-1}-W_n^{-1}\right\|\xrightarrow{p}0. \quad (139)$$

By $\lambda_{\min}(W_n) \geq C$ and (138),

$$\left\| W_n^{-1/2}\widehat{W}_n W_n^{-1/2} - I_l \right\| \leq \left\| W_n^{-1} \right\| \left\| \widehat{W}_n - W_n \right\| \leq C^{-1} \left\| \widehat{W}_n - W_n \right\| \xrightarrow{p} 0.$$

By the continuous mapping theorem,

$$\left(W_n^{-1/2}\widehat{W}_n W_n^{-1/2} \right)^{-1} \xrightarrow{p} I_l^{-1} = I_l.$$

Thus,

$$\begin{aligned} \left\| \widehat{W}_n^{-1} - W_n^{-1} \right\| &= \left\| W_n^{-1/2} \left(W_n^{1/2}\widehat{W}_n^{-1} W_n^{1/2} - I_l \right) W_n^{-1/2} \right\| \\ &\leq C^{-1} \left\| W_n^{1/2}\widehat{W}_n^{-1} W_n^{1/2} - I_l \right\| \xrightarrow{p} 0. \end{aligned}$$

Next we show

$$(Q_n' W_n^{-1} Q_n)^{1/2} (\widehat{Q}_n' \widehat{W}_n^{-1} \widehat{Q}_n)^{-1} (Q_n' W_n^{-1} Q_n)^{1/2} \xrightarrow{p} I_k. \tag{140}$$

By the continuous mapping theorem, (140) is equivalent to show

$$(Q_n' W_n^{-1} Q_n)^{-1/2} (\widehat{Q}_n' \widehat{W}_n^{-1} \widehat{Q}_n) (Q_n' W_n^{-1} Q_n)^{-1/2} \xrightarrow{p} I_k. \tag{141}$$

Under $\lambda_{\min}(W_n) \geq C > 0$ and the full column rank condition,

$$\lambda_{\min} \left(Q_n' W_n^{-1} Q_n \right) \geq C > 0. \tag{142}$$

Since

$$\begin{aligned} & \left\| (Q_n' W_n^{-1} Q_n)^{-1/2} (\widehat{Q}_n' \widehat{W}_n^{-1} \widehat{Q}_n) (Q_n' W_n^{-1} Q_n)^{-1/2} - I_k \right\| \\ \leq \;& C^{-1} \left\| \left(\widehat{Q}_n' \widehat{W}_n^{-1} \widehat{Q}_n - Q_n' W_n^{-1} Q_n \right) \right\| \end{aligned}$$

By (142) it is sufficient for (141) to show

$$\widehat{Q}_n' \widehat{W}_n^{-1} \widehat{Q}_n - Q_n' W_n^{-1} Q_n \xrightarrow{p} 0. \tag{143}$$

Observe that by Cauchy-Schwarz inequality,

$$\begin{aligned} \|Q_n\| &\leq \frac{1}{n} \sum_{g=1}^{G} \sum_{j=1}^{n_g} E \left\| \boldsymbol{z}_{gj} \boldsymbol{x}_{gj}' \right\| \\ &\leq \sup_i E \left\| \boldsymbol{z}_i \boldsymbol{x}_i' \right\| \\ &\leq \sup_i \left(E \left\| \boldsymbol{z}_i \right\|^2 \right)^{1/2} \left(E \left\| \boldsymbol{x}_i \right\|^2 \right)^{1/2} < \infty. \end{aligned} \tag{144}$$

By centering $\widehat{Q}_n$ and $\widehat{W}_n^{-1}$ around Q_n and W_n^{-1},

$$
\begin{aligned}
& \left\| \widehat{Q}_n' \widehat{W}_n^{-1} \widehat{Q}_n - Q_n' W_n^{-1} Q_n \right\| \\
\leq & \left\| \left(\widehat{Q}_n - Q_n\right)' \left(\widehat{W}_n^{-1} - W_n^{-1}\right) \left(\widehat{Q}_n - Q_n\right) \right\| \\
& +2 \left\| \left(\widehat{Q}_n - Q_n\right)' \left(\widehat{W}_n^{-1} - W_n^{-1}\right) Q_n \right\| + \left\| \left(\widehat{Q}_n - Q_n\right)' W_n^{-1} \left(\widehat{Q}_n - Q_n\right) \right\| \\
& + \left\| \left(\widehat{Q}_n - Q_n\right)' W_n^{-1} Q_n \right\| + \left\| Q_n' W_n^{-1} \left(\widehat{Q}_n - Q_n\right) \right\| + \left\| Q_n' \left(\widehat{W}_n^{-1} - W_n^{-1}\right) Q_n \right\| \\
\leq & \ o_p(1)
\end{aligned}
$$

by (137), (139), $\|W_n^{-1}\| \leq C^{-1}$, and (144).

Lastly, we show

$$\left\| \widehat{Q}_n' \widehat{W}_n^{-1} \widehat{S}_n \right\| \xrightarrow{p} 0. \tag{145}$$

Using (136), (137), (139), and $\|W_n^{-1}\| \leq C^{-1}$,

$$
\begin{aligned}
\left\| \widehat{Q}_n' \widehat{W}_n^{-1} \widehat{S}_n \right\| \leq & \left\| Q_n' W_n^{-1} \widehat{S}_n \right\| \\
& + \left\| Q_n' (\widehat{W}_n^{-1} - W_n^{-1}) \widehat{S}_n + (\widehat{Q}_n - Q_n)' W_n^{-1} \widehat{S}_n + (\widehat{Q}_n - Q_n)' \left(\widehat{W}_n^{-1} - W_n^{-1}\right) \widehat{S}_n \right\| \\
\leq & \left(O(1) C^{-1} + o_p(1)\right) \left\| \widehat{S}_n \right\| \xrightarrow{p} 0.
\end{aligned}
$$

Combining the results (140), (142), and (145),

$$
\begin{aligned}
\left\| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right\| \leq & \left(\left\| \left(Q_n' W_n^{-1} Q_n\right)^{-1} \right\| + \left\| \left(Q_n' W_n^{-1} Q_n\right)^{1/2} \left(\widehat{Q}_n' \widehat{W}_n^{-1} \widehat{Q}_n\right)^{-1} \left(Q_n' W_n^{-1} Q_n\right)^{1/2} - I_k \right\| \right) \\
& \cdot \left\| \widehat{Q}_n' \widehat{W}_n^{-1} \widehat{S}_n \right\| \\
\leq & \left(C^{-1} + o_p(1)\right) o_p(1).
\end{aligned}
$$

This completes the proof. ■

Proof of Theorem 11: We start by showing some useful results. Since $Q_n' W_n^{-1} Q_n = O(1)$, by (143),

$$
\begin{aligned}
\left(\widehat{Q}_n' \widehat{W}_n^{-1} \widehat{Q}_n\right)^{-1} &= \left(Q_n' W_n^{-1} Q_n\right)^{-1} \left(I_k + Q_n' W_n^{-1} Q_n \left((\widehat{Q}_n' \widehat{W}_n^{-1} \widehat{Q}_n)^{-1} - (Q_n' W_n^{-1} Q_n)^{-1} \right)\right) \\
&= \left(Q_n' W_n^{-1} Q_n\right)^{-1} \left(I_k + o_p(1)\right).
\end{aligned} \tag{146}
$$

In addition, by (137) and (139)

$$
\begin{aligned}
\widehat{Q}_n' \widehat{W}_n^{-1} &= Q_n' W_n^{-1} + Q_n' W_n^{-1} Q_n \left(Q_n' W_n^{-1} Q_n\right)^{-1} \left(\widehat{Q}_n' \widehat{W}_n^{-1} - Q_n' W_n^{-1}\right) \\
&= Q_n' W_n^{-1} \left(I_l + o_p(1)\right).
\end{aligned} \tag{147}
$$

By (146) and (147),

$$
\begin{aligned}
& \left(\widehat{Q}_n' \widehat{W}_n^{-1} \widehat{Q}_n\right)^{-1} \widehat{Q}_n' \widehat{W}_n^{-1} \sqrt{n} \widehat{S}_n \\
= \; & \left(Q_n' W_n^{-1} Q_n\right)^{-1} Q_n' W_n^{-1} \sqrt{n} \widehat{S}_n && (148) \\
& + \left(Q_n' W_n^{-1} Q_n\right)^{-1} o_p(1) Q_n' W_n^{-1} \sqrt{n} \widehat{S}_n + \left(Q_n' W_n^{-1} Q_n\right)^{-1} Q_n' W_n^{-1} o_p(1) \sqrt{n} \widehat{S}_n && (149) \\
& + \left(Q_n' W_n^{-1} Q_n\right)^{-1} o_p(1) Q_n' W_n^{-1} o_p(1) \sqrt{n} \widehat{S}_n && (150)
\end{aligned}
$$

Let

$$
R_n^* = V_n^{1/2} R_n \left(R_n' V_n R_n\right)^{-1/2}.
$$

First take (148). Since

$$
n\text{var}\left(V_n^{-1/2} \left(Q_n' W_n^{-1} Q_n\right)^{-1} Q_n' W_n^{-1} \widehat{S}_n\right) = I_k,
$$

we apply Theorem 2 to find

$$
R_n^{*\prime} V_n^{-1/2} \left(Q_n' W_n^{-1} Q_n\right)^{-1} Q_n' W_n^{-1} \sqrt{n} \widehat{S}_n \xrightarrow{d} N(0, I_q) \tag{151}
$$

In addition, (151) implies that (149) and (150) are bounded by $O_p(1)o_p(1) = o_p(1)$. Thus (48) follows under the assumptions.

For (49) we show that

$$
\left\| V_n^{-1/2} \widehat{V}_n V_n^{-1/2} - I_k \right\| \xrightarrow{p} 0. \tag{152}
$$

We first show that

$$
\left\| \Omega_n^{-1} \left(\widehat{\Omega}_n - \Omega_n\right) \right\| \xrightarrow{p} 0 \tag{153}
$$

which is equivalent of showing

$$
\left\| \Omega_n^{-1/2} \widehat{\Omega}_n \Omega_n^{-1/2} - I_l \right\| \xrightarrow{p} 0. \tag{154}
$$

Define

$$
\widetilde{\Omega}_n = \frac{1}{n} \sum_{g=1}^{G} \boldsymbol{Z}_g' \boldsymbol{e}_g \boldsymbol{e}_g' \boldsymbol{Z}_g
$$

and

$$
\begin{aligned}
\widetilde{\Omega}_n^* &= \frac{1}{n} \sum_{g=1}^{G} \boldsymbol{Z}_g' \left(\widehat{\boldsymbol{e}}_g - \boldsymbol{e}_g\right) \left(\widehat{\boldsymbol{e}}_g - \boldsymbol{e}_g\right)' \boldsymbol{Z}_g \\
&= \frac{1}{n} \sum_{g=1}^{G} \boldsymbol{Z}_g' \boldsymbol{X}_g \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \boldsymbol{X}_g' \boldsymbol{Z}_g.
\end{aligned}
$$

By the triangle and Cauchy-Schwarz inequalities

$$\begin{aligned}
\left\|\Omega_n^{-1/2}\widehat{\Omega}_n\Omega_n^{-1/2} - I_l\right\| &\le \left\|\Omega_n^{-1/2}\widetilde{\Omega}_n\Omega_n^{-1/2} - I_l\right\| + \left\|\Omega_n^{-1/2}\frac{1}{n}\sum_{g=1}^{G}\boldsymbol{Z}_g'\left(\widehat{\boldsymbol{e}}_g\widehat{\boldsymbol{e}}_g' - \boldsymbol{e}_g\boldsymbol{e}_g'\right)\boldsymbol{Z}_g\Omega_n^{-1/2}\right\| \\
&\le \left\|\Omega_n^{-1/2}\widetilde{\Omega}_n\Omega_n^{-1/2} - I_l\right\| + 2\left\|\Omega_n^{-1/2}\widetilde{\Omega}_n\Omega_n^{-1/2}\right\|^{1/2}\left\|\Omega_n^{-1/2}\widetilde{\Omega}_n^*\Omega_n^{-1/2}\right\|^{1/2} \\
&\quad + \left\|\Omega_n^{-1/2}\widetilde{\Omega}_n^*\Omega_n^{-1/2}\right\|.
\end{aligned}$$

Under the assumption, Theorem 3 implies that

$$\Omega_n^{-1/2}\widetilde{\Omega}_n\Omega_n^{-1/2} \xrightarrow{p} I_l. \tag{155}$$

The proof of (153) is completed by showing that

$$\left\|\Omega_n^{-1/2}\widetilde{\Omega}_n^*\Omega_n^{-1/2}\right\| \xrightarrow{p} 0. \tag{156}$$

Since $\|(Q_n' W_n^{-1} Q_n)^{-1}\| \le C^{-1}$ and $\|W_n^{-1}\| \le C^{-1}$,

$$\begin{aligned}
\left\|\Omega_n^{-1/2}\widetilde{\Omega}_n^*\Omega_n^{-1/2}\right\| &= \left\|\Omega_n^{-1/2}\frac{1}{n}\sum_{g=1}^{G}\boldsymbol{Z}_g'\boldsymbol{X}_g\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\boldsymbol{X}_g'\boldsymbol{Z}_g\Omega_n^{-1/2}\right\| \\
&\le \left\|V_n^{-1/2}\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right\|^2\left\|\frac{1}{n^2}\sum_{g=1}^{G}\boldsymbol{X}_g'\boldsymbol{Z}_g\Omega_n^{-1}\boldsymbol{Z}_g'\boldsymbol{X}_g V_n\right\| \\
&\le O_p(1)C^{-4}\frac{1}{n^2}\sum_{g=1}^{G}\left\|\boldsymbol{Z}_g'\boldsymbol{X}_g\right\|^2 \\
&= o_p(1)
\end{aligned}$$

since by Minkowski's and Cauchy-Schwarz inequalities, and $E\|\boldsymbol{x}_{gi}\|^4 \le C$, and $E\|\boldsymbol{z}_{gi}\|^4 \le C$,

$$\begin{aligned}
E\left\|\boldsymbol{Z}_g'\boldsymbol{X}_g\right\|^2 = E\left\|\sum_{i=1}^{n_g}\boldsymbol{z}_{gi}\boldsymbol{x}_{gi}'\right\|^2 &\le \left(\sum_{i=1}^{n_g}\left(E\left\|\boldsymbol{z}_{gi}\boldsymbol{x}_{gi}'\right\|^2\right)^{1/2}\right)^2 \\
&\le \left(\sum_{i=1}^{n_g}\left(E\left\|\boldsymbol{z}_{gi}\right\|^2\left\|\boldsymbol{x}_{gi}\right\|^2\right)^{1/2}\right)^2 \\
&\le \left(\sum_{i=1}^{n_g}\left(E\left\|\boldsymbol{z}_{gi}\right\|^4\right)^{1/4}\left(E\left\|\boldsymbol{x}_{gi}\right\|^4\right)^{1/4}\right)^2 \\
&\le Cn_g^2
\end{aligned}$$

so

$$\frac{1}{n^2}\sum_{g=1}^{G}E\left\|\boldsymbol{Z}_g'\boldsymbol{X}_g\right\|^2 \le \frac{C}{n^2}\sum_{g=1}^{G}n_g^2 = o(1)$$

by (7).

By (153),

$$\begin{aligned}\widehat{\Omega}_n &= \Omega_n\left(I_l+\Omega_n^{-1}\left(\widehat{\Omega}_n-\Omega_n\right)\right)\\ &= \Omega_n(I_l+o_p(1)).\end{aligned} \tag{157}$$

By (146), (147), (157), and the triangle inequality

$$\begin{aligned}&\left\|V_n^{-1/2}\widehat{V}_nV_n^{-1/2}-I_k\right\|\\ =&\left\|V_n^{-1/2}\left(\widehat{Q}_n'\widehat{W}_n^{-1}\widehat{Q}_n\right)^{-1}\widehat{Q}_n'\widehat{W}_n^{-1}\widehat{\Omega}_n\widehat{W}_n^{-1}\widehat{Q}_n\left(\widehat{Q}_n'\widehat{W}_n^{-1}\widehat{Q}_n\right)^{-1}V_n^{-1/2}-I_k\right\|\\ =&\left\|V_n^{-1/2}\left(Q_n'W_n^{-1}Q_n\right)^{-1}(I_k+o_p(1))Q_n'W_n^{-1}(I_l+o_p(1))\Omega_n(I_l+o_p(1))\right.\\ &\left.\cdot(I_l+o_p(1))W_n^{-1}Q_n\left(Q_n'W_n^{-1}Q_n\right)^{-1}(I_k+o_p(1))V_n^{-1/2}-I_k\right\|\\ \leq&\left\|V_n^{-1/2}V_nV_n^{-1/2}-I_k\right\|+\left\|V_n^{-1/2}V_nV_n^{-1/2}\right\|o_p(1)\\ \leq&\ o_p(1).\end{aligned}$$

Therefore, (49) is proved.

Finally, (50) follows as in the proof of Theorem 3 (13). ■

Proof of Theorem 12: We proceed by verifying the conditions of Theorem 2.1 of Newey and McFadden (1994) where $E[\log f(X_i,\boldsymbol{\theta})]$ and $L_n(\boldsymbol{\theta})$ correspond to their $Q_0(\boldsymbol{\theta})$ and $\widehat{Q}_n(\boldsymbol{\theta})$.

Their condition (i) holds by Lemma 2.2 of Newey and McFadden (1994) under our conditions 2 and 3.

Their condition (ii) is our condition 1.

Their conditions (iii) and (iv) hold by our Theorem 5 under our conditions 1, 3, 4, and (2). ■

Proof of Theorem 13: We start by showing that

$$E\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log f(X_i,\boldsymbol{\theta}_0)\right]=0, \tag{158}$$

which holds by Lemma 3.6 of Newey and McFadden (1994) under our conditions 2(a) and 2(b). By Theorem 12 and condition 1, $\widehat{\boldsymbol{\theta}}$ is in the interior of $\boldsymbol{\Theta}$ with probability approaching one and the first-order condition (FOC) holds:

$$\frac{1}{n}\sum_{i=1}^n\frac{\partial}{\partial\boldsymbol{\theta}}\log f(X_i,\widehat{\boldsymbol{\theta}})=0. \tag{159}$$

Define

$$R_n^*=V_n^{1/2}R_n\left(R_n'V_nR_n\right)^{-1/2}$$

and

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

By the mean value theorem,

$$\begin{aligned}\left(R_n' V_n R_n\right)^{-1/2} R_n' \sqrt{n}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right) &= R_n^{*\prime} V_n^{-1/2}\sqrt{n}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)\\ &= -R_n^{*\prime} V_n^{-1/2}\widehat{H}_n\left(\overline{\boldsymbol{\theta}}\right)^{-1}\sqrt{n}\widehat{S}_n\left(\boldsymbol{\theta}_0\right) \qquad (160)\end{aligned}$$

where $\overline{\boldsymbol{\theta}}$ is a mean value lies on a line segment joining $\boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\theta}}$. Let $\mathcal{N}$ be a neighborhood of $\boldsymbol{\theta}_0$. Take n large enough so that $\widehat{\boldsymbol{\theta}} \in \mathcal{N}$.

We first show

$$\widehat{H}_n(\overline{\boldsymbol{\theta}}) = H_n(\boldsymbol{\theta}_0)(I_k + o_p(1)). \qquad (161)$$

Since we can write

$$\widehat{H}_n(\overline{\boldsymbol{\theta}}) = H_n(\boldsymbol{\theta}_0)\left(I_k + H_n(\boldsymbol{\theta}_0)^{-1}\left(\widehat{H}_n(\overline{\boldsymbol{\theta}}) - H_n(\boldsymbol{\theta}_0)\right)\right),$$

it suffices to show

$$\left\|H_n(\boldsymbol{\theta}_0)^{-1}\left(\widehat{H}_n(\overline{\boldsymbol{\theta}}) - H_n(\boldsymbol{\theta}_0)\right)\right\| \xrightarrow{p} 0. \qquad (162)$$

But by the triangle inequality and Theorem 5,

$$\begin{aligned}&\left\|H_n(\boldsymbol{\theta}_0)^{-1}\left(\widehat{H}_n(\overline{\boldsymbol{\theta}}) - H_n(\boldsymbol{\theta}_0)\right)\right\|\\ \leq\ & C^{-1}\left(\sup_{\boldsymbol{\theta}\in\mathcal{N}}\left\|\widehat{H}_n(\boldsymbol{\theta}) - H_n(\boldsymbol{\theta})\right\| + \left\|H_n(\overline{\boldsymbol{\theta}}) - H_n(\boldsymbol{\theta}_0)\right\|\right) \xrightarrow{p} 0.\end{aligned}$$

By Woodbury matrix identity, (161) implies

$$\widehat{H}_n(\overline{\boldsymbol{\theta}})^{-1} = H_n(\boldsymbol{\theta}_0)^{-1}(I_k + o_p(1)). \qquad (163)$$

Using (163), (160) can be written as

$$\begin{aligned}-R_n^{*\prime} V_n^{-1/2}\widehat{H}_n\left(\overline{\boldsymbol{\theta}}\right)^{-1}\sqrt{n}\widehat{S}_n\left(\boldsymbol{\theta}_0\right) &= -R_n^{*\prime} V_n^{-1/2} H_n\left(\boldsymbol{\theta}_0\right)^{-1}\sqrt{n}\widehat{S}_n\left(\boldsymbol{\theta}_0\right) \qquad (164)\\ &\quad -R_n^{*\prime} V_n^{-1/2} H_n\left(\boldsymbol{\theta}_0\right)^{-1} o_p(1)\sqrt{n}\widehat{S}_n\left(\boldsymbol{\theta}_0\right). \qquad (165)\end{aligned}$$

First take the RHS of (164). Since $\operatorname{var}\left(V_n^{-1/2} H_n\left(\boldsymbol{\theta}_0\right)^{-1}\sqrt{n}\widehat{S}_n\left(\boldsymbol{\theta}_0\right)\right) = I_k$, by Theorem 2 under the conditions,

$$-R_n^{*\prime} V_n^{-1/2} H_n\left(\boldsymbol{\theta}_0\right)^{-1}\sqrt{n}\widehat{S}_n\left(\boldsymbol{\theta}_0\right) \xrightarrow{d} N(0, I_q). \qquad (166)$$

Next, given (166), (165) can be bounded by $O_p(1)o_p(1)$. Thus, (54) is proved.

To show (55) it is equivalent to show

$$\left\| V_n^{-1/2} \widehat{V}_n V_n^{-1/2} - I_k \right\| \xrightarrow{p} 0.$$

Since (163) holds by replacing $\overline{\boldsymbol{\theta}}$ with $\widehat{\boldsymbol{\theta}}$,

$$\widehat{H}_n(\widehat{\boldsymbol{\theta}})^{-1} = H_n(\boldsymbol{\theta}_0)^{-1}(I_k + o_p(1)). \tag{167}$$

Since $\lambda_{\min}(\Omega_n(\boldsymbol{\theta})) \geq \lambda > 0$, with probability approaching one,

$$\begin{aligned} \widehat{\Omega}_n(\widehat{\boldsymbol{\theta}}) &= \Omega_n(\boldsymbol{\theta}_0)\left(I_k + \Omega_n(\boldsymbol{\theta}_0)^{-1}\Omega_n(\widehat{\boldsymbol{\theta}})\Omega_n(\widehat{\boldsymbol{\theta}})^{-1}\left(\widehat{\Omega}_n(\widehat{\boldsymbol{\theta}}) - \Omega_n(\boldsymbol{\theta}_0)\right)\right) \\ &= \Omega_n(\boldsymbol{\theta}_0)(I_k + o_p(1)), \end{aligned} \tag{168}$$

because

$$\begin{aligned} & \left\| \Omega_n(\boldsymbol{\theta}_0)^{-1}\Omega_n(\widehat{\boldsymbol{\theta}})\Omega_n(\widehat{\boldsymbol{\theta}})^{-1}\left(\widehat{\Omega}_n(\widehat{\boldsymbol{\theta}}) - \Omega_n(\boldsymbol{\theta}_0)\right)\right\| \\ \leq\ & \left\| \Omega_n(\boldsymbol{\theta}_0)^{-1/2}\Omega_n(\widehat{\boldsymbol{\theta}})\Omega_n(\boldsymbol{\theta}_0)^{-1/2} \right\| \\ & \cdot \left(\sup_{\boldsymbol{\theta}\in\mathcal{N}} \left\| \Omega_n(\boldsymbol{\theta})^{-1/2}\widehat{\Omega}_n(\boldsymbol{\theta})\Omega_n(\boldsymbol{\theta})^{-1/2} - I_l \right\| + \left\| \Omega_n(\widehat{\boldsymbol{\theta}})^{-1/2}\Omega_n(\boldsymbol{\theta}_0)\Omega_n(\widehat{\boldsymbol{\theta}})^{-1/2} - I_l \right\| \right) \\ \leq\ & o_p(1). \end{aligned}$$

The first inequality holds by the triangle and Schwarz Matrix inequalities. The second inequality holds by Theorem 6 (21), $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$, and continuity of $\Omega_n(\boldsymbol{\theta})$ in $\boldsymbol{\theta}$ under condition 2(a).

By using (167) and (168),

$$\begin{aligned} & \left\| V_n^{-1/2}\widehat{V}_n V_n^{-1/2} - I_k \right\| \\ =\ & \left\| V_n^{-1/2}\widehat{H}_n(\widehat{\boldsymbol{\theta}})^{-1}\widehat{\Omega}_n(\widehat{\boldsymbol{\theta}})\widehat{H}_n(\widehat{\boldsymbol{\theta}})^{-1}V_n^{-1/2} - I_k \right\| \\ =\ & \left\| V_n^{-1/2}H_n(\boldsymbol{\theta}_0)^{-1}(I_k + o_p(1))\,\Omega_n(\boldsymbol{\theta}_0)(I_k + o_p(1))\,H_n(\boldsymbol{\theta}_0)^{-1}(I_k + o_p(1))\,V_n^{-1/2} - I_k \right\| \\ \leq\ & \left\| V_n^{-1/2}V_n V_n^{-1/2} - I_k \right\| + \left\| V_n^{-1/2}V_n V_n^{-1/2} \right\| o_p(1) \\ \leq\ & o_p(1). \end{aligned}$$

Thus, (55) is proved.

Finally, (56) follows as in the proof of (13). ■

Proof of Theorem 14: Write $m_n(\boldsymbol{\theta}) = E\overline{m}_n(\boldsymbol{\theta})$. Define the population GMM criterion function as

$$J_n(\boldsymbol{\theta}) = n \cdot m_n(\boldsymbol{\theta})' W_n^{-1} m_n(\boldsymbol{\theta}). \tag{169}$$

We proceed by verifying the conditions of Theorem 2.1 of Newey and McFadden (1994). Let $-n^{-1}J_n(\boldsymbol{\theta})$ be their $Q_0(\boldsymbol{\theta})$.

Their condition (i) holds by our conditions 2 and 5.

Their condition (ii) is our condition 1.

By Theorem 5 under (2) and conditions 3 and 4, $m_n(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ uniformly over $\boldsymbol{\theta} \in \Theta$ and

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\overline{m}_n(\boldsymbol{\theta}) - m_n(\boldsymbol{\theta})\| \xrightarrow{p} 0. \tag{170}$$

Since $-n^{-1}J_n(\boldsymbol{\theta})$ is continuous and their condition (iii) holds.

Finally, by $\boldsymbol{\Theta}$ compact, $m_n(\boldsymbol{\theta})$ is bounded on $\boldsymbol{\Theta}$. By the triangle and Schwarz Matrix inequalities,

$$\begin{aligned}
& \left\|\overline{m}_n(\boldsymbol{\theta})'\widehat{W}_n^{-1}\overline{m}_n(\boldsymbol{\theta}) - m_n(\boldsymbol{\theta})'W_n^{-1}m_n(\boldsymbol{\theta})\right\| \\
\leq\ & \left\|(\overline{m}_n(\boldsymbol{\theta}) - m_n(\boldsymbol{\theta}))'\left(\widehat{W}_n^{-1} - W_n^{-1}\right)(\overline{m}_n(\boldsymbol{\theta}) - m_n(\boldsymbol{\theta}))\right\| \\
& + \left\|(\overline{m}_n(\boldsymbol{\theta}) - m_n(\boldsymbol{\theta}))'W_n^{-1}(\overline{m}_n(\boldsymbol{\theta}) - m_n(\boldsymbol{\theta}))\right\| + 2\left\|(\overline{m}_n(\boldsymbol{\theta}) - m_n(\boldsymbol{\theta}))'\left(\widehat{W}_n^{-1} - W_n^{-1}\right)m_n(\boldsymbol{\theta})\right\| \\
& + \left\|m_n(\boldsymbol{\theta})'\left(\widehat{W}_n^{-1} - W_n^{-1}\right)m_n(\boldsymbol{\theta})\right\| + 2\left\|(\overline{m}_n(\boldsymbol{\theta}) - m_n(\boldsymbol{\theta}))'W_n^{-1}m_n(\boldsymbol{\theta})\right\| \\
\leq\ & \|\overline{m}_n(\boldsymbol{\theta}) - m_n(\boldsymbol{\theta})\|\left(\|\overline{m}_n(\boldsymbol{\theta}) - m_n(\boldsymbol{\theta})\| + 2\|m_n(\boldsymbol{\theta})\|\right)\left(\left\|\widehat{W}_n^{-1} - W_n^{-1}\right\| + C^{-1}\right) \\
& + \|m_n(\boldsymbol{\theta})\|^2\left\|\widehat{W}_n^{-1} - W_n^{-1}\right\|.
\end{aligned}$$

By taking the supremum over $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ on both sides, their condition (iv) holds by (170) and our condition 6. ■

Proof of Theorem 15: By conditions 1, 2(a), 3, and Theorem 14, the sample FOC

$$2\widehat{Q}_n'\widehat{W}_n^{-1}\overline{m}_n(\widehat{\boldsymbol{\theta}}) = 0 \tag{171}$$

is satisfied with probability approaching one. Define

$$R_n^* = V_n^{1/2}R_n\left(R_n'V_nR_n\right)^{-1/2}.$$

By the mean value theorem, we can write

$$\begin{aligned}
\left(R_n'V_nR_n\right)^{-1/2}R_n'\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) &= R_n^{*\prime}V_n^{-1/2}\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\
&= -R_n^{*\prime}V_n^{-1/2}\left(\widehat{Q}_n'\widehat{W}_n^{-1}\widehat{Q}_n(\overline{\boldsymbol{\theta}})\right)^{-1}\widehat{Q}_n'\widehat{W}_n^{-1}\sqrt{n}\overline{m}_n(\boldsymbol{\theta}_0)
\end{aligned} \tag{172}$$

where $\overline{\boldsymbol{\theta}}$ is a mean value lies on a line segment joining $\boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\theta}}$.

First we show

$$\widehat{Q}_n' = Q_n'\left(I_l + o_p(1)\right). \tag{173}$$

Since Q_n is full rank and $\lambda_{\min}(W_n) \geq C > 0$,

$$\lambda_{\min}\left(Q_n'W_nQ_n\right) \geq C > 0. \tag{174}$$

We can write

$$\widehat{Q}_n' = Q_n' \left\{ I_l + W_n^{-1} Q_n \left(Q_n' W_n^{-1} Q_n \right)^{-1} \left(\widehat{Q}_n - Q_n(\widehat{\boldsymbol{\theta}}) + Q_n(\widehat{\boldsymbol{\theta}}) - Q_n \right) \right\}. \tag{175}$$

Let $\mathcal{N}$ be a neighborhood of $\boldsymbol{\theta}_0$. Take n large enough so that $\widehat{\boldsymbol{\theta}} \in \mathcal{N}$. By Theorem 5 under the assumptions,

$$\left\| \widehat{Q}_n - Q_n(\widehat{\boldsymbol{\theta}}) \right\| \leq \sup_{\boldsymbol{\theta} \in \mathcal{N}} \left\| \widehat{Q}_n(\boldsymbol{\theta}) - Q_n(\boldsymbol{\theta}) \right\| \xrightarrow{p} 0, \tag{176}$$

and

$$\left\| Q_n(\widetilde{\boldsymbol{\theta}}) - Q_n \right\| \xrightarrow{p} 0. \tag{177}$$

Since $Q_n = O(1)$,

$$\left\| W_n^{-1} Q_n \left(Q_n' W_n^{-1} Q_n \right)^{-1} \left(\widehat{Q}_n - Q_n \right) \right\| \leq C^{-2} O(1) o_p(1).$$

Thus, (173) is shown. Using the same argument we also have

$$\widehat{Q}_n'(\overline{\boldsymbol{\theta}}) = Q_n' \left(I_l + o_p(1) \right). \tag{178}$$

Next, we show

$$\widehat{W}_n = W_n \left(I_l + o_p(1) \right). \tag{179}$$

We can write

$$\widehat{W}_n = W_n \left(I_l + W_n^{-1} W_n(\widetilde{\boldsymbol{\theta}}) W_n(\widetilde{\boldsymbol{\theta}})^{-1} (\widehat{W}_n - W_n) \right)$$

but by the triangle and Schwarz matrix inequalities

$$\begin{aligned}
& \left\| W_n^{-1} W_n(\widetilde{\boldsymbol{\theta}}) W_n(\widetilde{\boldsymbol{\theta}})^{-1} (\widehat{W}_n - W_n) \right\| \\
\leq \; & \left\| W_n^{-1/2} W_n(\widetilde{\boldsymbol{\theta}}) W_n^{-1/2} \right\| \\
& \cdot \left(\sup_{\boldsymbol{\theta} \in \mathcal{N}} \left\| W_n(\boldsymbol{\theta})^{-1/2} \widehat{W}_n(\boldsymbol{\theta}) W_n(\boldsymbol{\theta})^{-1/2} - I_l \right\| + \left\| W_n(\widetilde{\boldsymbol{\theta}})^{-1/2} W_n W_n(\widetilde{\boldsymbol{\theta}})^{-1/2} - I_l \right\| \right) \\
\leq \; & o_p(1)
\end{aligned} \tag{180}$$

if $W_n(\boldsymbol{\theta})$ and $W_n(\boldsymbol{\theta})^{-1}$ are continuous in $\boldsymbol{\theta} \in \mathcal{N}$ and

$$\sup_{\boldsymbol{\theta} \in \mathcal{N}} \left\| W_n(\boldsymbol{\theta})^{-1/2} \widehat{W}_n(\boldsymbol{\theta}) W_n(\boldsymbol{\theta})^{-1/2} - I_l \right\| \xrightarrow{p} 0. \tag{181}$$

Since $m(x, \boldsymbol{\theta})$ is continuously differentiable and $W_n(\boldsymbol{\theta})$ is nonsingular for all $\boldsymbol{\theta} \in \mathcal{N}$, $W_n(\boldsymbol{\theta})$ and $W_n(\boldsymbol{\theta})^{-1}$ are continuous. Thus, it suffices to show (181).

First consider $\widehat{W}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n m(X_i,\boldsymbol{\theta})m(X_i,\boldsymbol{\theta})'$. Then by Theorem 5,

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\mathcal{N}}\left\|W_n(\boldsymbol{\theta})^{-1/2}\widehat{W}_n(\boldsymbol{\theta})W_n(\boldsymbol{\theta})^{-1/2} - I_l\right\| &= \sup_{\boldsymbol{\theta}\in\mathcal{N}}\left\|W_n(\boldsymbol{\theta})^{-1/2}\left(\widehat{W}_n(\boldsymbol{\theta}) - W_n(\boldsymbol{\theta})\right)W_n(\boldsymbol{\theta})^{-1/2}\right\| \\ &\leq C^{-1}\sup_{\boldsymbol{\theta}\in\mathcal{N}}\left\|\widehat{W}_n(\boldsymbol{\theta}) - W_n(\boldsymbol{\theta})\right\| \xrightarrow{p} 0. \qquad (182)\end{aligned}$$

The case for $\widehat{W}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n m(X_i,\boldsymbol{\theta})m(X_i,\boldsymbol{\theta})' - \overline{m}_n(\boldsymbol{\theta})\overline{m}_n(\boldsymbol{\theta})'$ can be shown similarly by Theorems 4 and 5. Next consider $\widehat{W}_n(\boldsymbol{\theta}) = \widehat{\Omega}_n(\boldsymbol{\theta})$. Then (181) holds by Theorem 6. Thus, (179) is shown.

By Woodbury matrix identity,

$$\begin{aligned}\widehat{W}_n^{-1} &= W_n^{-1}\left(I_l + o_p(1)\right), \qquad (183)\\ \left(\widehat{Q}_n'\widehat{W}_n^{-1}\widehat{Q}_n(\overline{\boldsymbol{\theta}})\right)^{-1} &= \left(Q_n'W_n^{-1}Q_n\right)^{-1}(I_k + o_p(1)). \qquad (184)\end{aligned}$$

By (173), (183), and (184), (172) can be written as

$$\begin{aligned}&-R_n^{*\prime}V_n^{-1/2}\left(\widehat{Q}_n'\widehat{W}_n^{-1}\widehat{Q}_n(\overline{\boldsymbol{\theta}})\right)^{-1}\widehat{Q}_n(\widetilde{\boldsymbol{\theta}})'\sqrt{n}\overline{m}_n(\boldsymbol{\theta}_0)\\ =\ &-R_n^{*\prime}V_n^{-1/2}\left(Q_n'W_n^{-1}Q_n\right)^{-1}(I_k + o_p(1))Q_n'(I_l + o_p(1))W_n^{-1}(I_l + o_p(1))\sqrt{n}\overline{m}_n(\boldsymbol{\theta}_0)\\ =\ &-R_n^{*\prime}V_n^{-1/2}\left(Q_n'W_n^{-1}Q_n\right)^{-1}Q_n'W_n^{-1}\sqrt{n}\overline{m}_n(\boldsymbol{\theta}_0) + U_n \qquad (185)\end{aligned}$$

where $\|U_n\| = o_p(1)$. Since

$$n\text{var}\left(-R_n^{*\prime}V_n^{-1/2}\left(Q_n'W_n^{-1}Q_n\right)^{-1}Q_n'W_n^{-1}\overline{m}_n(\boldsymbol{\theta}_0)\right) = I_k,$$

we apply Theorem 2 to find

$$-R_n^{*\prime}V_n^{-1/2}\left(Q_n'W_n^{-1}Q_n\right)^{-1}Q_n'W_n^{-1}\sqrt{n}\overline{m}_n(\boldsymbol{\theta}_0) \xrightarrow{d} N(0, I_q). \qquad (186)$$

Thus, (65) is shown.

To show (66) it is equivalent to show

$$\left\|V_n^{-1/2}\widehat{V}_nV_n^{-1/2} - I_k\right\| \xrightarrow{p} 0.$$

Since (179) holds with $\widehat{W}_n = \widehat{\Omega}_n(\widetilde{\boldsymbol{\theta}})$ and both $\widetilde{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}$ are consistent, using the same argument with (179) we have

$$\widehat{\Omega}_n = \Omega_n(I_l + o_p(1)). \qquad (187)$$

By using (173), (183), (184), and (187),

$$\begin{aligned}
& \left\| V_n^{-1/2}\widehat{V}_n V_n^{-1/2} - I_k \right\| \\
= & \left\| V_n^{-1/2}(\widehat{Q}_n'\widehat{W}_n^{-1}\widehat{Q}_n)^{-1}\widehat{Q}_n'\widehat{W}_n^{-1}\Omega_n\widehat{W}_n^{-1}\widehat{Q}_n(\widehat{Q}_n'\widehat{W}_n^{-1}\widehat{Q}_n)^{-1}V_n^{-1/2} - I_k \right\| \\
\leq & \left\| V_n^{-1/2}V_n V_n^{-1/2} - I_k \right\| + \left\| V_n^{-1/2}V_n V_n^{-1/2} \right\| o_p(1) \\
\leq & \; o_p(1).
\end{aligned}$$

For the efficient weight matrix case,

$$\begin{aligned}
\left\| V_n^{-1/2}\widehat{V}_n V_n^{-1/2} - I_k \right\| &= \left\| V_n^{-1/2}(\widehat{Q}_n'\widehat{\Omega}_n^{-1}\widehat{Q}_n)^{-1}V_n^{-1/2} - I_k \right\| \\
&\leq \left\| V_n^{-1/2}V_n V_n^{-1/2} - I_k \right\| + \left\| V_n^{-1/2}V_n V_n^{-1/2} \right\| o_p(1) \\
&\leq o_p(1).
\end{aligned}$$

Thus, (66) is proved.

Next, (67) follows as in the proof of (13).

Finally, we show (68). By the mean value theorem, the triangle inequality, (173), and Theorems 2 and 15 (65),

$$\begin{aligned}
& \left\| \Omega_n^{-1/2}\sqrt{n}\overline{m}_n(\widehat{\boldsymbol{\theta}}) \right\| \\
\leq & \left\| \Omega_n^{-1/2}\sqrt{n}\overline{m}_n(\boldsymbol{\theta}_0) \right\| + \left\| \Omega_n^{-1/2}Q_n\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \right\| (1 + o_p(1)) \\
\leq & \; O_p(1) + \left\| \sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' V_n^{-1/2} V_n^{1/2} Q_n'\Omega_n^{-1}Q_n V_n^{1/2} V_n^{-1/2}\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \right\|^{1/2} (1 + o_p(1)) \\
\leq & \; O_p(1) + \left\| V_n^{-1/2}\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \right\| (1 + o_p(1)) \\
\leq & \; O_p(1).
\end{aligned}$$

Since by (187) and Woodbury matrix identity,

$$\begin{aligned}
\left\| n \cdot \overline{m}_n(\widehat{\boldsymbol{\theta}})'\widehat{\Omega}_n^{-1}\overline{m}_n(\widehat{\boldsymbol{\theta}}) - n \cdot \overline{m}_n(\widehat{\boldsymbol{\theta}})'\Omega_n^{-1}\overline{m}_n(\widehat{\boldsymbol{\theta}}) \right\| &= \left\| n \cdot \overline{m}_n(\widehat{\boldsymbol{\theta}})' \left(\widehat{\Omega}_n^{-1} - \Omega_n^{-1} \right) \overline{m}_n(\widehat{\boldsymbol{\theta}}) \right\| \\
&\leq \left\| \Omega_n^{-1/2}\sqrt{n}\overline{m}_n(\widehat{\boldsymbol{\theta}}) \right\|^2 o_p(1) \\
&\leq O_p(1)o_p(1),
\end{aligned}$$

it suffices to show

$$n \cdot \overline{m}_n(\widehat{\boldsymbol{\theta}})'\Omega_n^{-1}\overline{m}_n(\widehat{\boldsymbol{\theta}}) \xrightarrow{d} \chi^2_{l-k}. \tag{188}$$

Using (172), (173), (183), and (184), we can write

$$\begin{aligned}
\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) &= -\left(\widehat{Q}_n'\widehat{\Omega}_n^{-1}\widehat{Q}_n(\overline{\boldsymbol{\theta}}) \right)^{-1} \widehat{Q}_n'\widehat{\Omega}_n^{-1}\sqrt{n}\overline{m}_n(\boldsymbol{\theta}_0) \\
&= -\left(Q_n'\Omega_n^{-1}Q_n \right)^{-1} Q_n'\Omega_n^{-1}\sqrt{n}\overline{m}_n(\boldsymbol{\theta}_0) + o_p(1).
\end{aligned} \tag{189}$$

By the mean value theorem and (189),

$$
\begin{aligned}
\Omega_n^{-1/2}\sqrt{n}\overline{m}_n(\widehat{\boldsymbol{\theta}}) &= \Omega_n^{-1/2}\sqrt{n}\overline{m}_n(\boldsymbol{\theta}_0) \\
&\quad -\Omega_n^{-1/2}Q_n(I_k+o_p(1))\left(\left(Q_n'\Omega_n^{-1}Q_n\right)^{-1}Q_n'\Omega_n^{-1}\sqrt{n}\overline{m}_n(\boldsymbol{\theta}_0)+o_p(1)\right) \\
&= \left(I_l-\Omega_n^{-1/2}Q_n\left(Q_n'\Omega_n^{-1}Q_n\right)^{-1}Q_n'\Omega_n^{-1/2}\right)\Omega_n^{-1/2}\sqrt{n}\overline{m}_n(\boldsymbol{\theta}_0) \qquad (190)\\
&\quad -\Omega_n^{-1/2}Q_n o_p(1)\left(Q_n'\Omega_n^{-1}Q_n\right)^{-1}Q_n'\Omega_n^{-1}\sqrt{n}\overline{m}_n(\boldsymbol{\theta}_0) \qquad (191)\\
&\quad -\Omega_n^{-1/2}Q_n(I_k+o_p(1))o_p(1). \qquad (192)
\end{aligned}
$$

Take (190). Since $I_l-\Omega_n^{-1/2}Q_n\left(Q_n'\Omega_n^{-1}Q_n\right)^{-1}Q_n'\Omega_n^{-1/2}$ is idempotent with rank $l-k$, (190) has the χ^2_{l-k} distribution asymptotically. For (191),

$$
\begin{aligned}
&\left\|\Omega_n^{-1/2}Q_n o_p(1)\left(Q_n'\Omega_n^{-1}Q_n\right)^{-1}Q_n'\Omega_n^{-1}\sqrt{n}\overline{m}_n(\boldsymbol{\theta}_0)\right\| \\
\le\ &\left\|\Omega_n^{-1/2}Q_n\left(Q_n'\Omega_n^{-1}Q_n\right)^{-1}Q_n'\Omega_n^{-1/2}\right\|O_p(1)o_p(1) \\
\le\ &o_p(1).
\end{aligned}
$$

For (192),

$$
\left\|\Omega_n^{-1/2}Q_n(I_k+o_p(1))o_p(1)\right\| \le \lambda^{-1/2}O(1)o_p(1).
$$

Thus, (68) is shown and the proof is completed. ■

References

[1] Andrews, Donald W.K. (1992). Generic uniform convergence. *Econometric Theory*, 8 (2), 241-257.

[2] Angrist, Joshua D. and J. S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

[3] Arellano, Manuel (1987). Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics*, 49, 431-434.

[4] Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004). How much should we trust differences-in-differences estimates?. *The Quarterly Journal of Economics*, 119(1), 249-275.

[5] Bester, C. Alan, Timothy G. Conley, and Christian B. Hansen (2011). Inference with dependent data using cluster covariance estimators. *Journal of Econometrics*, 165(2), 137-151.

[6] Cameron, A. Colin, and Douglas L. Miller (2011). Robust inference with clustered data. *Handbook of Empirical Economics and Finance*, ed. A. Ullah and D.E. Giles, New York: CRC Press, 1-28.

[7] Cameron, A. Colin, and Douglas L. Miller (2015). A practitioner's guide to cluster robust inference. *Journal of Human Resources*, 50, 317-372.

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

[9] Canay, Ivan A., Joseph P. Romano, and Azeem M. Shaikh (2017). Randomization tests under an approximate symmetry assumption. *Econometrica*, 85(3), 1013-1030.

[10] Carter, Andrew V., Kevin T. Schnepel, and Douglas G. Steigerwald (2017). Asymptotic behavior of a t test robust to cluster heterogeneity. *Review of Economics and Statistics*, forthcoming.

[11] Conley, Timothy G., and Christopher R. Taber (2011). Inference with 'difference in differences' with a small number of policy changes. *Review of Economics and Statistics*, 93, 113-125.

[12] Djogbenou, Antoine A., Morten O. Nielsen and James G. MacKinnon (2017). Validity of wild bootstrap with clustered errors. Working paper.

[13] Donald, Stephen G., and Kevin Lang (2007). Inference with difference in differences and other panel data. *Review of Economics and Statistics*, 89, 221-223.

[14] Etemadi, Nasrollah (2006). Convergence of weighted averages of random variables revisited. *Proceedings of the American Mathematical Society*, 134(9), 2739-2744.

[15] Hall, Alastair R. and Atsushi Inoue (2003). The large sample behaviour of the generalized method of moments estimator in misspecified models. *Journal of Econometrics*, 114(2), 361-394.

[16] Hansen, Bruce E. (2018). *Econometrics*, http://www.ssc.wisc.edu/˜bhansen/econometrics/.

[17] Hansen, Bruce E. and Seojeong Lee (2017). Inference for iterated GMM under misspecification and clustering. Working paper.

[18] Hansen, Christian B. (2007). Asymptotic properties of a robust variance matrix estimator for panel data when T is large. *Journal of Econometrics*, 141, 597-620.

[19] Hwang, Jungbin. (2016). Simple and trustworthy cluster-robust GMM inference. Working paper, University of Connecticut.

[20] Ibragimov, Rustam, and Ulrich K. Müller (2010). t-statistic based correlation and heterogeneity robust inference. *Journal of Business and Economic Statistics*, 28, 453-468.

[21] Ibragimov, Rustam, and Ulrich K. Müller (2016). Inference with few heterogeneous clusters. *Review of Economics and Statistics*, 98, 83-96.

[22] Imbens, Guido W., and Michal Kolesár (2016). Robust standard errors in small samples: Some practical advice. *Review of Economics and Statistics*, 98, 701-712.

[23] Lee, Seojeong. (2014). Asymptotic refinements of a misspecification-robust bootstrap for generalized method of moments estimators. *Journal of Econometrics*, 178, 398-413.

[24] Liang, Kung-Yee, and Scott L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

[25] MacKinnon, James G. (2012). Thirty years of heteroskedasticity-robust inference. *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, ed. Xiaohong Chen and Norman R. Swanson. New York, Springer, 437-461.

[26] MacKinnon, James G. (2016). Inference with large clustered datasets. Working paper, Queen's University.

[27] MacKinnon, James G., and Matthew D. Webb (2017a). Wild bootstrap inference for wildly different cluster sizes. *Journal of Applied Econometrics*, 32, 233-254.

[28] MacKinnon, James G., and Matthew D. Webb (2017b) The wild bootstrap for few (treated) clusters. *The Econometrics Journal*, forthcoming.

[29] Moulton, Brent R. (1986). Random group effects and the precision of regression estimates. *Journal of Econometrics*, 32, 385-397.

[30] Moulton, Brent R. (1990). An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *Review of Economics and Statistics*, 72, 334-338.

[31] Newey, Whitney K., and Daniel McFadden (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4, 2111-2245.

[32] Rogers, William (1994). Regression standard errors in clustered samples. *STATA Technical Bulletin*, 13, 19-23.

[33] van der Vaart, Aad W.(1998). *Asymptotic Statistics*, New York: Cambridge University Press.

[34] White, Halbert (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48(4), 817-838.

[35] White, Halbert (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1), 1-25.

[36] White, Halbert (1984). *Asymptotic Theory for Econometricians.*

[37] Wooldridge, Jeffrey M. (2003). Cluster-sample methods in applied econometrics. *American Economic Review*, 93, 133-138.

[38] Wooldridge, Jeffrey M. (2010) *Econometric analysis of cross section and panel data, 2nd edition*. Cambridge: MIT Press.

[39] Young, Alwyn (2016). Improved, nearly exact, statistical inference with robust and clustered covariance matrices using effective degrees of freedom corrections. Working paper, London School of Economics.