Asymptotic Properties of M-estimators with Finite Populations under Cluster Sampling and Cluster Assignment

Ruonan Xu*

Department of Economics, Rutgers University,

New Jersey Hall, New Brunswick, NJ 08901, USA (ruonan.xu@rutgers.edu)

Abstract

I establish asymptotic properties of M-estimators under finite populations with clustered data, allowing for unbalanced and unbounded cluster sizes in the limit. I distinguish between two situations that justify computing clustered standard errors: i) cluster sampling induced by random sampling of groups of units, and ii) cluster assignment caused by the correlated assignment of "treatment" within the same group. The finite population cluster-robust asymptotic variance (CRAV) is found to be less than or equal to the usual infinite population CRAV, in the matrix sense. I also show that one should only use clustered standard errors when there is cluster sampling or cluster assignment, or both, for a general class of linear and nonlinear estimators.

Keywords: *Finite population inference; M-estimation; Cluster-robust inference; Potential outcomes.*

JEL classification: *C18, C10*

*I am grateful to Jeffrey Wooldridge, Kyoo il Kim, Todd Elder, Timothy Vogelsang, and Anastasia Semykina for helpful comments and suggestions.

1 Introduction

The cluster-robust asymptotic variance (CRAV) has been studied extensively in the literature because of its wide and often inevitable application.[1] However, until recently, the statistical frameworks used to justify clustering largely assume an infinite population, at least implicitly. The infinite population approach yields proper inference in some cases. Intuitively, when the sampling fraction is small, such as the 1% U.S. Public Use Microdata Sample, it is harmless to assume the sample is drawn from an infinite population. In other cases, the paradigm of drawing a sample from an infinite population does not lead to usable inference. A leading case is when the sample and the population coincide, such as when data are available on all 3,142 counties in the U.S., or when we can collect data on exam performance for all fourth graders in a school district. Adopting the same finite population setting in Abadie, Athey, Imbens, and Wooldridge (2017) [hereafter, AAIW (2017)], this paper studies asymptotic properties of M-estimators where clusters are formed by either the sampling process or the assignment design.

There are three approaches to justifying clustering corrections of the standard errors. The conventional one is the model-based approach; see, for instance, Kloek (1981), Moulton (1986), and Moulton (1990). Empirical researchers often suspect that unobserved components in outcomes for individual units are correlated within groups. As a result, error components models are used to account for the potential within-group correlation, and the clustered standard errors follow after the model setup.

The problem with the model-based approach is the essentially arbitrary nature of the choice of clustering level. For instance, one researcher may claim that the unobservables are correlated at the zip code level; another may claim that correlation exists at the county

[1] See, for example, White (1984), Liang and Zeger (1986), Arellano (1987), Wooldridge (2003), Bertrand, Duflo, and Mullainathan (2004), Hansen (2007), Cameron and Miller (2015), and MacKinnon (2019).

or the state level. Some rules of thumb suggest clustering at the highest level possible until the number of clusters becomes too small to use standard asymptotics, and using the cluster-robust standard errors whenever there is an appreciable difference between the clustered standard errors and the Eicker-Huber-White (EHW) standard errors (Cameron and Miller, 2015, p. 333). This approach mainly addresses the question of when clustering makes a difference in the magnitude of the standard errors which, as shown by AAIW (2017), is not a justification for whether we should adjust standard errors for clustering.

Another approach is based purely on sampling considerations; see, for example, Kish and Frankel (1974), Scott and Holt (1982), Bell and McCaffrey (2002), and Bhattacharya (2005). Namely, one needs to adjust the standard errors for clustering when the primary sampling units (PSUs) are groups instead of individuals. There might be a second step in the sampling process, though, where individual units are sampled randomly within the selected groups. Note that when the entire population is used in the analysis, there is no cluster sampling.

A third approach to studying clustering is a design-based perspective. Related to randomized experiments literature, assignments are clustered when individual "treatments" are correlated within each group. In the leading case, individual assignments are perfectly correlated within clusters, such as the minimum wage law imposed on states. Because of cluster assignment, clustering adjustments are required even if the entire population is observed (i.e., no cluster sampling).

In the context of the difference-in-means estimator, AAIW (2017) show clustering is only necessary when there is either cluster sampling or cluster assignment, or both. For multiple regression and nonlinear estimators that are widely used in empirical studies, no such results are currently available. One contribution of the current paper is to fill in this gap in the literature; I find the same guidelines for clustering adjustments for the

difference-in-means estimator also hold for M-estimators. In addition, I provide a unified framework of deriving the finite population CRAV for M-estimators by accounting for and distinguishing between cluster sampling and cluster assignment, which also allows for unbalanced and unbounded cluster sizes in the limit. I find when the number of clusters in the sample is nonnegligible compared with the total number of clusters in the finite population, or when the sample coincides with the finite population, the usual CRAV is no less than the finite population CRAV, in the matrix sense. This means that in cases where the sampling proportion is reasonably large, the usually reported clustered standard errors would, in general, be too large.

The current paper contributes to two strands of literature. The first is on finite population inference methods. Abadie, Athey, Imbens, and Wooldridge (2020) [hereafter, AAIW (2020)] propose the finite population inference methods for ordinary least squares estimators incorporating both sampling-based and design-based uncertainties.[2] Xu (2020) extends AAIW (2020) to M-estimation while maintaining the assumption of independent sampling and independent assignment. AAIW (2017) examines the cluster-robust variance of the difference-in-means estimator caused by cluster sampling or cluster assignment under finite populations but do not provide a general form of CRAV for a broad class of linear and nonlinear estimators. As a result, the earlier research on finite population inference has limited applications but is contained as sub-cases of the unified framework derived in the current paper.

Second, this paper is related to the literature studying CRAV. The majority of this literature considers fixed cluster sizes or clusters of equal sizes in the setting of infinite populations. Recently, several articles contribute to the development of the asymptotic theory allowing for unbalanced and potentially unbounded cluster sizes; see Carter, Schne-

[2]See AAIW (2020) for a detailed explanation for sampling-based and design-based uncertainties.

pel, and Steigerwald (2017), Djogbenou, MacKinnon, and Nielsen (2019), and Hansen and Lee (2019). I extend the techniques developed by Hansen and Lee (2019) to the finite population asymptotics. In this way, the framework allows for heterogeneous and large cluster sizes in the samples.

The remaining of this paper is organized as follows. Section 2 derives the asymptotic distribution under finite populations for M-estimators with smooth objective functions. Even though the finite population CRAV is generally non-identifiable, Section 3 provides simple ways to find a less conservative upper bound of it. Section 4 derives the finite population CRAV of the functions containing M-estimators with the estimator of the average partial effect (APE) as a direct application. Simulation results are summarized in Section 5 and an empirical application is carried out in Section 6. Lastly, Section 7 concludes and points out directions for future research. Proofs and additional tables are collected in the appendix and online supplement.

2 Asymptotic Properties of M-estimators

2.1 Setup

Consider a sequence of finite populations indexed by population size M, where M diverges to infinity in deriving the asymptotic properties. Suppose there are G mutually exclusive clusters in population M defined as either the PSUs in the sampling scheme or the partition in the assignment design, where each cluster has M_g units, $g = 1, 2, \ldots, G$.[3] For each unit i within cluster g, we observe $(X_{igM}, z_{igM}, Y_{igM})$, where X_{igM} is the vector of assignment variables, z_{igM} is a set of attributes, and Y_{igM} is the realized outcome. The categorization of assignments and attributes is based on the posed empirical question.

[3] I assume here that sampling and assignments are clustered at the same level if there are both cluster sampling and cluster assignment, which can be relaxed by introducing more notation.

Typically, the key variables of interest in an empirical study could be viewed as assignment variables, and the rest covariates as attribute variables. There is no restriction in terms of the nature of the triple above: they can be discrete, continuous, or mixed. When the distinction across clusters is unnecessary, the triple is denoted by (X_{iM}, z_{iM}, Y_{iM}). For the most part, I denote $W_{igM} = (X_{igM}, Y_{igM})$ ($W_{iM} = (X_{iM}, Y_{iM})$) for brevity. Everything is denoted by a subscript M to emphasize their dependence on the population size.

Given the potential outcome framework, there exists a mapping, denoted by the potential outcome function $y_{igM}(x)$, from the assignment variables to the potential outcomes. For example, $y_{igM}(x) = x\theta_{01} + z_{igM}\theta_{02} + e_{igM}$ for continuous outcomes, and $y_{igM}(x) = \mathbb{1}[x\theta_{01} + z_{igM}\theta_{02} + e_{igM} > 0]$ for binary outcomes, where z_{igM} and e_{igM} are observed and unobserved attributes.[4] The potential outcome function, $y_{igM}(\cdot)$, along with the observed attributes z_{igM}, are non-stochastic.[5] By contrast, the assignment vector x is random, with X_{igM} denoting the assignment for unit i of cluster g in population M. As a result, the realized potential outcome, $Y_{igM} = y_{igM}(X_{igM})$, is random. Alternatively, the finite population setting can be understood as conditioned on the potential outcomes and attributes of the M units in the population.

As is the starting point in the infinite population paradigm, I study solutions to a population minimization problem, where the estimand of interest is a $k \times 1$ vector denoted by θ^*_M.

$$\begin{aligned}\theta^*_M &= \arg\min_\theta \frac{1}{M}\sum_{g=1}^{G}\sum_{i=1}^{M_g}\mathbb{E}_X\left[q_{igM}(W_{igM},\theta)\right] \\ &= \arg\min_\theta \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_X\left[q_{iM}(W_{iM},\theta)\right] \end{aligned} \tag{1}$$

[4] I emphasize x as the argument of the potential outcome function because it is the only stochastic variables in the function.

[5] This implies that the unobserved attributes are non-stochastic as well.

Notice that the expectation $\mathbb{E}_X$ in (1) is taken over the distribution of X since X is the source of randomness here.

Function $q_{igM}(\cdot,\cdot)$ is the objective function for a single unit. The subscripts of the objective function indicate its dependence on the non-stochastic attribute variables z_{igM}. When the objective function is a negative log-likelihood function, we are essentially performing quasi-maximum likelihood estimation (QMLE) given the discreteness of the unobservables in the finite population. For instance, for a binary response, the objective function could be $q_{iM}(W_{iM},\theta) = -\{Y_{iM}\log\left[\Phi(X_{iM}\theta_1 + z_{iM}\theta_2)\right] + (1-Y_{iM})\log\left[1-\Phi(X_{iM}\theta_1 + z_{iM}\theta_2)\right]\}$ as in the usual probit regression; even though this objective function is misspecified, it works well when the unobservables in the latent variable model approach a normal distribution as the population size gets large.

Let R_{igM} denote the binary sampling indicator, which is equal to one if unit i in cluster g is sampled. Hence, the sample size is $N = \sum_{g=1}^{G}\sum_{i=1}^{M_g} R_{igM} = \sum_{i=1}^{M} R_{iM}$.[6] The estimator of θ_M^* is denoted by $\hat{\theta}_N$, which solves the minimization problem in the sample.

$$\begin{aligned}\hat{\theta}_N &= \arg\min_{\theta} \frac{1}{N}\sum_{g=1}^{G}\sum_{i=1}^{M_g} R_{igM} q_{igM}(W_{igM},\theta) \\ &= \arg\min_{\theta} \frac{1}{N}\sum_{i=1}^{M} R_{iM} q_{iM}(W_{iM},\theta) \end{aligned} \tag{2}$$

I make the following assumptions throughout the paper. Additional regularity conditions are summarized in Appendix A.

Assumption 1. *(i) The sampling scheme consists of two steps. In the first step, a random group of clusters is drawn according to Bernoulli sampling with probability $\rho_{cM} > 0$; in the second step, units are independently sampled, according to a Bernoulli trial with probability $\rho_{uM} > 0$, from the subpopulation consisting of all the sampled clusters. (ii) The sequence*

[6] R_{igM} is suppressed as R_{iM} when the emphasis of clusters is unnecessary.

of sampling probabilities ρ_{cM} *and* ρ_{uM} *satisfies* $\rho_{cM} \to \rho_c \in (0,1]$ *and* $\rho_{uM} \to \rho_u \in (0,1]$, *as* $M \to \infty$.

Assumption 2. *The assignments are independent across clusters but allowed to be correlated within clusters.*

Assumption 3. *The vector of assignments is independent of the vector of sampling indicators.*

Assumption 4. $\max_{g \leq G} \frac{M_g}{M} \to 0$, *as* $M \to \infty$.

Assumption 5. $\frac{\sum_{g=1}^{G} M_g^2}{M} \leq C < \infty$ *and* $\max_{g \leq G} \frac{M_g^2}{M} \to 0$, *as* $M \to \infty$.

Based on Assumption 1, the sample size is random and cluster sampling occurs whenever $\rho_{cM} < 1$. If we assume that the cluster sizes are bounded, then zero can be included in the limiting value of both sampling probabilities given higher moments of the score functions are bounded. When the limiting sampling probabilities are zero, the infinite population inference can be nested in the framework as a special case.

Assumption 2 allows for clustered assignment, which is another source of within-cluster correlation in addition to cluster sampling. The assignment variables X_{igM} are not necessarily identically distributed, which allows the assignments to depend on fixed attributes z_{igM}. Assumption 3 implies that the sampling process and the assignment process are independent of each other. When $\rho_{cM} = 1$ and the assignment of X_{igM} is independent both across and within groups, Assumptions 1-3 contains independent sampling and independent assignment as a special case.

Assumptions 4 and 5 are adapted from Hansen and Lee (2019) to restrict cluster heterogeneity and the growth rate of the cluster sizes relative to that of the population size. Since each cluster is asymptotically negligible, Assumption 4 rules out the case where a

particular group of clusters dominates the population and implies $G \to \infty$. The first part of Assumption 5 rules out clustered data with all clusters unbounded in the limit.[7] Once Assumption 5 is imposed, there is no need to impose Assumption 4 because the latter is implied by the former. However, to show consistency of M-estimators, only Assumption 4 is required.

2.2 Asymptotic Distribution

The following theorem shows that M-estimators are consistent estimators regardless of whether a finite- or infinite-population framework is adopted. I use large-G asymptotics here, where the number of clusters in the population $G \to \infty$ as $M \to \infty$.

Theorem 2.1. *Under Assumptions 1-4 and Assumption A.1 in the appendix,* $\hat{\theta}_N - \theta_M^* \xrightarrow{p} \mathbf{0}$.

Let $m_{iM}(W_{iM}, \theta)$ denote the score function of $q_{iM}(W_{iM}, \theta)$. The variance matrix of M-estimators is defined as

$$\begin{aligned} V_M =& H_M(\theta_M^*)^{-1}\big(\Delta_{ehw,M}(\theta_M^*) + \rho_{uM}\Delta_{cluster,M}(\theta_M^*) \\ & - \rho_{uM}\rho_{cM}\Delta_{E,M} - \rho_{uM}\rho_{cM}\Delta_{EC,M}\big)H_M(\theta_M^*)^{-1}, \end{aligned} \tag{3}$$

where

$$\Delta_{ehw,M}(\theta) = \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_X\big[m_{iM}(W_{iM},\theta)m_{iM}(W_{iM},\theta)'\big], \tag{4}$$

$$\Delta_{E,M} = \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_X\big[m_{iM}(W_{iM},\theta_M^*)\big]\mathbb{E}_X\big[m_{iM}(W_{iM},\theta_M^*)\big]', \tag{5}$$

$$\Delta_{cluster,M}(\theta) = \frac{1}{M}\sum_{g=1}^{G}\sum_{i=1}^{M_g}\sum_{j\neq i}^{M_g}\mathbb{E}_X\big[m_{igM}(W_{igM},\theta)m_{jgM}(W_{jgM},\theta)'\big], \tag{6}$$

[7] As a shorthand, clustered data is used in the remaining text to refer to the situation where there is cluster sampling, cluster assignment, or both.

$$\Delta_{EC,M} = \frac{1}{M}\sum_{g=1}^{G}\sum_{i=1}^{M_g}\sum_{j\neq i}^{M_g}\mathbb{E}_X\big[m_{igM}(W_{igM},\theta_M^*)]\mathbb{E}_X[m_{jgM}(W_{jgM},\theta_M^*)\big]', \tag{7}$$

and

$$H_M(\theta) = \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_X\big[\nabla_\theta m_{iM}(W_{iM},\theta)\big]. \tag{8}$$

The conventional infinite population variance matrix is denoted by

$$V_{SM} = H_M(\theta_M^*)^{-1}\big(\Delta_{ehw,M}(\theta_M^*) + \rho_{uM}\Delta_{cluster,M}(\theta_M^*)\big)H_M(\theta_M^*)^{-1}. \tag{9}$$

Notice that the middle part of the sandwich form of V_M is different from that of V_{SM} with two additional terms $\Delta_{E,M}$ and $\Delta_{EC,M}$ scaled by the composite sampling probability.

The usual cluster-robust variance estimator (CRVE) is given by

$$\hat{V}_{SN} = \hat{H}_N\big(\hat{\theta}_N\big)^{-1}(\hat{\Delta}_{ehw,N}(\hat{\theta}_N) + \hat{\Delta}_{cluster,N}(\hat{\theta}_N)\big)\hat{H}_N(\hat{\theta}_N)^{-1}, \tag{10}$$

where

$$\hat{H}_N(\theta) = \frac{1}{N}\sum_{i=1}^{M}R_{iM}\nabla_\theta m_{iM}(W_{iM},\theta), \tag{11}$$

$$\hat{\Delta}_{ehw,N}(\theta) = \frac{1}{N}\sum_{i=1}^{M}R_{iM}\cdot m_{iM}(W_{iM},\theta)m_{iM}(W_{iM},\theta)', \tag{12}$$

and

$$\hat{\Delta}_{cluster,N}(\theta) = \frac{1}{N}\sum_{g=1}^{G}\sum_{i=1}^{M_g}\sum_{j\neq i}^{M_g}R_{igM}R_{jgM}\cdot m_{igM}(W_{igM},\theta)m_{jgM}(W_{jgM},\theta)'. \tag{13}$$

Theorem 2.2. *Under Assumptions 1, 2, 3, 5, and Assumptions A.1 and A.2 in the appendix, (1)* $V_M^{-1/2}\sqrt{N}(\hat{\theta}_N - \theta_M^*) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, I_k)$*; (2)* $V_{SM}^{-1/2}\hat{V}_{SN}V_{SM}^{-1/2} \xrightarrow{p} I_k$.

Theorem 2.2 shows asymptotic normality with the alternate finite population CRAV.

Because I allow arbitrary within-cluster correlations, the composite $V_M^{-1/2}\sqrt{N}$ serves as the implicit convergence rate. In terms of the variance-covariance matrices, the term $\Delta_{cluster,M}(\theta_M^*)$ is scaled by the sampling probability ρ_{uM} because of the two-stage sampling scheme. Nevertheless, the usual CRVE, $\hat{V}_{SN}$, converges to V_{SM}, in which the estimation of ρ_{uM} has been accounted for.

Corollary 1. *Clustering is unnecessary unless there is cluster sampling ($\rho_{cM} < 1$) or cluster assignment ($\Delta_{cluster,M}(\theta_M^*) \neq \Delta_{EC,M}$), or both.*

The term related to clustering in the variance formula, $\Delta_{cluster,M}(\theta_M^*) - \rho_{cM}\Delta_{EC,M}(\theta_M^*)$, is zero if we have both independent sampling and independent assignment. Hence, ex-post clustering in the usual way makes the standard errors conservative. Corollary 1 implies that we should adjust standard errors of M-estimators for clustering at the level of cluster sampling or cluster assignment. It can be shown using similar arguments that when cluster sampling and cluster assignment occur at different but nested levels, one should cluster at the higher level. This conclusion may seem counterintuitive at first, since correlations among individual unobservables play no specific role in determining clustering adjustment. Instead of arbitrary clustering based on clustered errors, the guidelines in the corollary provide a more clear-cut of clustering adjustment: whenever the sampling schemes or the assignment rules are known, we have the idea of the appropriate level to cluster the standard errors.

Corollary 1 reproduces results of Corollary 1(i) in AAIW (2017) but in a much more generalized way. AAIW (2017) prove the case for the difference-in-means estimator, while the corollary above holds for all M-estimators with either continuous or discrete assignment variables.

Corollary 2. *The infinite population CRAV of M-estimators is no less than the finite population CRAV, in the matrix sense.*

I rewrite the two additional terms below to compare the asymptotic variance of M-estimators obtained in Theorem 2.2(1) to the usual infinite population CRAV. Because the sum of the two additional terms in (14) is positive semidefinite, we have the conclusion in Corollary 2, which is a generalization of the results in AAIW (2020) and Xu (2020) to clustered data.

$$\Delta_{E,M}+\Delta_{EC,M} = \frac{1}{M}\sum_{g=1}^{G}\left[\sum_{i=1}^{M_g}\mathbb{E}_X\left(m_{igM}(W_{igM},\theta_M^*)\right)\right]\left[\sum_{i=1}^{M_g}\mathbb{E}_X\left(m_{igM}(W_{igM},\theta_M^*)\right)\right]' \quad (14)$$

Corollary 2 together with Theorem 2.2(2) imply that the usual CRVE is often overly conservative. There are exceptional cases where using the usual CRVE for inference is approximately correct, though. The leading scenario is summarized in the corollary below.

Corollary 3. *If a relatively small number of clusters is sampled from a large population of clusters, i.e., ρ_{cM} is close to zero, or there is at most one unit sampled from each cluster, i.e., ρ_{uM} is close to zero, it is approximately correct to use the usual CRVE of M-estimators for inference.*

Corollary 3 reaches the same conclusion of Corollary 2(ii) and 2(iii) in AAIW (2017) but in a general framework of M-estimation. When ρ_{cM} is small, which is the case close to sampling from an infinite number of clusters, the finite population CRAV is close to its infinite population counterpart. In other words, the CRAV of M-estimators in the infinite population setting is in general conservative unless a relatively small number of clusters is sampled. When ρ_{uM} is close to zero so that there is at most one unit sampled from each cluster, the finite population CRAV is close to the usual EHW asymptotic variance. As a result, it is approximately correct to use the usual EHW variance estimator for inference, and so is the usual CRVE since clustering adjustment does not matter in this case.

Another special case for the usual CRVE to be correct for inference is when $\Delta_{E,M}$ +

$\Delta_{EC,M} = \mathbf{0}$, which is true if either $\mathbb{E}_X\big[m_{igM}(W_{igM}, \theta^*_M)\big] = \mathbf{0},\ \forall\ i = 1, \ldots, M_g,\ g = 1, \ldots, G$ or $\sum_{i=1}^{M_g} \mathbb{E}_X\big[m_{igM}(W_{igM}, \theta^*_M)\big] = \mathbf{0},\ \forall\ g = 1, \ldots, G$. The former is true for the variance of the coefficient estimator on the assignment variables under the sufficient conditions provided by AAIW (2020), including constant treatment effects and other linearity conditions. The latter holds if the finite population is composed of repetitions of the smallest cluster. With this kind of data structure, θ^*_M that solves $\mathbb{E}_X\Big[\sum_{g=1}^{G}\sum_{i=1}^{M_g} m_{igM}(W_{igM}, \theta^*_M)\Big] = \mathbf{0}$ is also the solution to $\mathbb{E}_X\Big[\sum_{i=1}^{M_g} m_{igM}(W_{igM}, \theta^*_M)\Big] = \mathbf{0}$ for each cluster g. However, these kinds of special cases rarely hold in practice.

3 Estimation of the Extra Terms in the Asymptotic Variance

The terms that show up in the usual CRAV can be estimated in the standard way. It is more challenging to estimate the two extra terms, $\Delta_{E,M}$ and $\Delta_{EC,M}$. The underlying reason is that $\mathbb{E}_X\big[m_{iM}(W_{iM}, \theta^*_M)\big]$ is generally non-identifiable due to the missing data problem of the potential outcome framework. For instance, with a binary assignment variable, $\mathbb{E}_X\big[m_{iM}(W_{iM}, \theta^*_M)\big] = P(X_{iM} = 1) \cdot m_{iM}\big((1, y_{iM}(1)), \theta^*_M\big) + P(X_{iM} = 0) \cdot m_{iM}\big((0, y_{iM}(0)), \theta^*_M\big)$. However, we do not observe both $y_{iM}(0)$ and $y_{iM}(1)$ at the same time. Here, I instead provide a menu of options valid under different circumstances to find a lower bound of the two extra terms, and essentially an upper bound of the finite population CRAV.

No matter whether there is second-step sampling within clusters or not, we can always remove part of $\Delta_{E,M}$ using the regression-based approach below. Consider the estimator,

$$\hat{\Delta}^Z_N = \frac{1}{N}\sum_{i=1}^{M} R_{iM}\hat{L}'_N z'_{iM} z_{iM}\hat{L}_N, \tag{15}$$

where $\hat{L}_N = \Big(\sum_{i=1}^{M} R_{iM} z_{iM}' z_{iM} \Big)^{-1} \Big[\sum_{i=1}^{M} R_{iM} z_{iM}' m_{iM}(W_{iM}, \hat{\theta}_N)' \Big]$.

Theorem 3.1. *In addition to Assumptions 1-4, Assumption A.1, and conditions (ii), (iii), and (viii) in Assumption A.2, assume that (i)* $\frac{1}{M}\sum_{i=1}^{M} z_{iM}' z_{iM}$ *is nonsingular; (ii)* $\sup_{i,M} \|z_{iM}\| < \infty$. *Then* $\boldsymbol{0} \leq \Delta_M^Z \leq \Delta_{E,M}$, *where* $\left\| \hat{\Delta}_N^Z - \Delta_M^Z \right\| \xrightarrow{p} 0$ *(all inequalities are in the matrix sense).*

With clustered data, we can include cluster dummies as regressors in the linear projection of $m_{iM}(W_{iM}, \hat{\theta}_N)$ onto the fixed attributes. Alternatively, when the variables in z_{iM} are discrete, we can partition the population into different strata based on the values of z_{iM}. Then $\mathbb{E}_X\big[m_{iM}(W_{iM}, \theta_M^*)\big]$ can be partially predicted by its within-stratum averages. However, the downside is that $\Delta_{EC,M}$, which contains $\sum_{g=1}^{G} M_g(M_g - 1)$ terms, still remains in the usual CRAV. Consequently, the adjusted finite population CRVE, using $\hat{\Delta}_N^Z$ to partially estimate $\Delta_{E,M}$, is still quite conservative.

We can do better if there is no second-step sampling within the selected clusters. In this case, the sampling indicator is denoted by R_{gM}, which is equal to one if cluster g is sampled. We could sum $m_{igM}(W_{igM}, \hat{\theta}_N)$ within each cluster, and linearly project $\sum_{i=1}^{M_g} m_{igM}(W_{igM}, \hat{\theta}_N)$ onto the fixed attributes. The number of observations in the linear projection would be the number of clusters in the sample. To reduce the dimensionality of the regressors, the fixed attributes can also be summed within clusters as one way of aggregation. As a result, $\sum_{i=1}^{M_g} \mathbb{E}_X\big[m_{igM}(W_{igM}, \theta_M^*)\big]$ can be partially estimated by its predicted value from the linear projection. Let

$$\tilde{z}_{gM} = \sum_{i=1}^{M_g} z_{igM}, \tag{16}$$

$$\tilde{m}_{gM}(\theta) = \sum_{i=1}^{M_g} m_{igM}(W_{igM}, \theta), \tag{17}$$

and

$$\hat{P}_N = \left(\sum_{g=1}^{G} R_{gM} \tilde{z}'_{gM} \tilde{z}_{gM} \right)^{-1} \left(\sum_{g=1}^{G} R_{gM} \tilde{z}'_{gM} \tilde{m}_{gM}(\hat{\theta}_N)' \right). \tag{18}$$

Estimate $\Delta_{E,M} + \Delta_{EC,M}$ by

$$\hat{\Delta}^{Z}_{CE,N} = \frac{1}{N} \sum_{g=1}^{G} R_{gM} \hat{P}'_N \tilde{z}'_{gM} \tilde{z}_{gM} \hat{P}_N. \tag{19}$$

Theorem 3.2. *In addition to Assumptions 1, 2, 3, 5, Assumption A.1, and conditions (ii), (iii), and (viii) in Assumption A.2, suppose that (i)* $\rho_{uM} = 1$*; (ii)* $\sum_{g=1}^{G} \tilde{z}'_{gM} \tilde{z}_{gM}$ *is nonsingular; (iii)* $\sup_{i,M} \|z_{iM}\| < \infty$*. Then* $0 \leq \Delta^{Z}_{CE,M} \leq \left(\Delta_{E,M} + \Delta_{EC,M}\right)$*, where* $\left\| \hat{\Delta}^{Z}_{CE,N} - \Delta^{Z}_{CE,M} \right\| \xrightarrow{p} 0$ *(all inequalities are in the matrix sense).*

Theorem 3.2 proposes an easy way to partially remove $\Delta_{E,M} + \Delta_{EC,M}$ all at once. The composite sampling probability $\rho_{uM}\rho_{cM}$ can be estimated by N/M, where the population size M is assumed to be known. If the entire population is observed, $\rho_{uM}\rho_{cM}$ is simply one. Since $\hat{\Delta}^{Z}_{CE,N}$ is positive semidefinite,

$$\hat{\Delta}_{ehw,N}(\hat{\theta}_N) + \hat{\Delta}_{cluster,N}(\hat{\theta}_N) - \frac{N}{M} \hat{\Delta}^{Z}_{CE,N} \leq \hat{\Delta}_{ehw,N}(\hat{\theta}_N) + \hat{\Delta}_{cluster,N}(\hat{\theta}_N) \tag{20}$$

(in the matrix sense) is an algebraic fact with finite samples. With large samples, even though the limit of the adjusted finite population CRVE is still conservative, it improves over the limit of the usual CRVE.

4 Asymptotic Distribution of the Functions of M-estimators

Sometimes, we are interested in the functions of M-estimators rather than M-estimators themselves. Let $f_{iM}(W_{iM}, \theta^*_M)$ be a $q \times 1$ function of W_{iM} and θ^*_M. Suppose we wish

to estimate $\gamma_M^* = \frac{1}{M}\sum_{i=1}^{M} \mathbb{E}_X\big[f_{iM}(W_{iM}, \theta_M^*)\big]$. As an example, γ_M^* could be the APE from nonlinear models, where $f(\cdot,\cdot)$ is some partial derivative for continuous variables or some difference function for discrete variables. Let $\hat{\gamma}_N = \frac{1}{N}\sum_{i=1}^{M} R_{iM} f_{iM}(W_{iM}, \hat{\theta}_N)$ be the estimator of γ_M^*. Its asymptotic distribution is given in Theorem 4.1 (see below).

Denote the finite population variance matrix by

$$V_{f,M} = \Delta^f_{ehw,M} + \rho_{uM}\Delta^f_{cluster,M} - \rho_{uM}\rho_{cM}\Delta^f_{E,M} - \rho_{uM}\rho_{cM}\Delta^f_{EC,M}. \tag{21}$$

The infinite population variance matrix is then $V_{f,SM} = \Delta^f_{ehw,M} + \rho_{uM}\Delta^f_{cluster,M}$. And the usual CRVE is denoted by $\hat{V}_{f,SN} = \hat{\Delta}^f_{ehw,N} + \hat{\Delta}^f_{cluster,N}$. The detailed definition of each term can be found in Appendix A.

Theorem 4.1. *Under Assumptions 1, 2, 3, 5, and Assumptions A.1-A.3 in the appendix, (1)* $V_{f,M}^{-1/2}\sqrt{N}(\hat{\gamma}_N - \gamma_M^*) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, I_q)$*; (2)* $V_{f,SM}^{-1/2} \cdot \hat{V}_{f,SN} \cdot V_{f,SM}^{-1/2} \xrightarrow{p} I_q$.

Theorem 4.1 shows that the conservative property of the usual CRVE of M-estimators carries over to the usual CRVE of any functions of M-estimators. We can also apply the same techniques in Section 3 to estimate the two extra terms, $\Delta^f_{E,M}$ and $\Delta^f_{EC,M}$. The only difference is that the dependent variables in the regression-based approach would be

$$f_{igM}(W_{igM}, \hat{\theta}_N) - \hat{\gamma}_N - \hat{F}_N(\hat{\theta}_N)\hat{H}_N(\hat{\theta}_N)^{-1} m_{igM}(W_{igM}, \hat{\theta}_N) \tag{22}$$

or the cluster sum of it rather than $m_{igM}(W_{igM}, \hat{\theta}_N)$ alone. (See the details of the notation in Appendix A.)

5 Simulation

In this section, I compare the Monte Carlo standard deviation of the APE estimator of the assignment variable in a binary response model with a set of different standard errors. In the population generating process, there is a single assignment variable $X_{igM} \in \{0, 1\}$ and a single attribute variable $z_{igM} \in \{-1, 1\}$, each equal to one with a probability of $1/2$. The potential outcome of a binary response is generated as

$$y_{igM}(x) = \mathbb{1}[x + 2z_{igM} \cdot x + c_{gM} + e_{igM} > 0]. \tag{23}$$

Because of the cluster setup, there is an unobserved group heterogeneity for each cluster, c_{gM}, which is generated as residuals from regressing random realization of a standard normal distribution on z_{igM}. The idiosyncratic unobservable e_{igM} is the residual from regressing random realization of a standard normal distribution on z_{igM} and c_{gM}. The data of z_{igM}, c_{gM}, and e_{igM} are generated once and kept fixed in the population M.

The random assignment of X_{igM} involves two stages. In the first stage, an assignment probability $p_{gM} \in [0, 1]$ for cluster g is drawn randomly from a distribution $h(\cdot)$ with mean $1/2$ and variance σ^2. In the second stage, X_{igM} in cluster g is assigned to one independently, with cluster specific probability p_{gM}. If $\sigma^2 > 0$, we have correlated assignments within each cluster but independent assignments across clusters. In the simulation, p_{gM} is either drawn from the standard uniform distribution or kept fixed at $1/2$. Hence, $\sigma^2 \in \{0, 1/12\}$.

There are 10,000 replications for each design. For each replication, X_{igM} is assigned according to the assignment rules above and then clusters are sampled with probability $\rho_c \in \{0.1, 0.5, 1\}$ from the finite population. Each resembles the case of sampling from an infinite number of clusters, drawing a nonnegligible chunk of clusters in the population, and observing all clusters in the population, respectively. There is no second-step sampling,

i.e., $\rho_u = 1$.

The expected sample size is kept the same across different designs with varying population sizes. Results with two different expected cluster numbers in the sample, 50 and 100, are reported. Within each population M, half of the clusters have four units and another half have eight units. Hence, the expected sample size is 300 and 600 respectively.

Estimates from the pooled probit regression of Y_{igM} on 1, X_{igM}, and z_{igM} are displayed in Table 1 below. To report the analytical standard errors, θ_M^* is computed by minimizing the finite population objective function as in (1) with each population size, where $q_{iM}(W_{iM}, \theta)$ is the Bernoulli log-likelihood function. In the left panel (columns (1)-(3)), the assignment variable X_{igM} is independently assigned for each unit in the population given that the assignment probability p_{gM} is fixed at 0.5 for all clusters. While in the right panel (columns (4)-(6)), assignments within clusters are correlated as each cluster has its specific assignment probability. Cluster sampling occurs when $\rho_c < 1$. As a result, for columns (1) and (2), there is cluster sampling but no cluster assignment; for column (3), there is neither cluster sampling nor cluster assignment; for columns (4) and (5), there are both cluster sampling and cluster assignment; while for column (6), there is cluster assignment but no cluster sampling.

Within each sample size, the first row in Table 1 reports the APE of X in the population obtained from the potential outcome function and the second row reports the average of the APE estimates across the replications. The population APEs vary across columns because each column has a different population. Even though there is some gap between the population APEs and the estimated APEs due to misspecification of the model, the estimated ones are not too off from the truth. With QMLE, the hope is to get the best approximation to the population APE given the model specified.

The third row reports the Monte Carlo standard deviation of the APE estimator. The

Table 1: Standard Errors and Coverage Rates for Pooled Probit: APE

		No Cluster Assignment			With Cluster Assignment		
		$\rho_c = 0.1$ (1)	$\rho_c = 0.5$ (2)	$\rho_c = 1$ (3)	$\rho_c = 0.1$ (4)	$\rho_c = 0.5$ (5)	$\rho_c = 1$ (6)
$G\rho_c = 50$	APE_M^*	0.1007	0.1283	0.1033	0.1230	0.0983	0.1267
	$\overline{\widehat{APE}}$	0.1090	0.1371	0.1094	0.1317	0.1084	0.1365
	$sd(\widehat{APE})$	0.0727	0.0604	0.0391	0.0819	0.0730	0.0521
	$\overline{se}_{cluster}$	0.0761	0.0755	0.0751	0.0830	0.0881	0.0875
	$cov_{cluster}$	(0.958)	(0.983)	(1.000)	(0.948)	(0.979)	(0.999)
	$\overline{se}_{adj}$	0.0741	0.0650	0.0528	0.0811	0.0766	0.0610
	cov_{adj}	(0.953)	(0.965)	(0.992)	(0.944)	(0.958)	(0.975)
	$\overline{se}_{ehw,adj}$	0.0513	0.0497	0.0469	0.0516	0.0496	0.0468
$G\rho_c = 100$	APE_M^*	0.1030	0.0908	0.0950	0.1133	0.1167	0.1100
	$\overline{\widehat{APE}}$	0.1117	0.1006	0.1044	0.1226	0.1254	0.1173
	$sd(\widehat{APE})$	0.0526	0.0421	0.0282	0.0581	0.0500	0.0368
	$\overline{se}_{cluster}$	0.0555	0.0535	0.0530	0.0602	0.0599	0.0589
	$cov_{cluster}$	(0.959)	(0.987)	(1.000)	(0.956)	(0.981)	(0.998)
	$\overline{se}_{adj}$	0.0540	0.0457	0.0370	0.0587	0.0530	0.0430
	cov_{adj}	(0.955)	(0.967)	(0.990)	(0.952)	(0.961)	(0.975)
	$\overline{se}_{ehw,adj}$	0.0363	0.0352	0.0337	0.0365	0.0351	0.0332

[1] G is the number of clusters in the population; ρ_c is the sampling probability of clusters; thus, $G\rho_c$ is the expected number of clusters in the sample.

[2] For cluster assignment, the variance of the assignment probability across clusters is 1/12.

[3] APE_M^* stands for the population APE; $\overline{\widehat{APE}}$ stands for the average of the APE estimates across replications; $sd(\widehat{APE})$ stands for the Monte Carlo standard deviation of the APE estimator; $\overline{se}_{cluster}$ stands for the average of the usual infinite population cluster-robust standard error; $cov_{cluster}$ stands for the coverage rate of the 95% confidence interval based on the usual cluster-robust standard error; $\overline{se}_{adj}$ stands for the average of the adjusted finite population cluster-robust standard error; cov_{adj} stands for the coverage rate of the 95% confidence interval based on the adjusted finite population cluster-robust standard error; $\overline{se}_{ehw,adj}$ stands for the average of the adjusted finite population EHW standard error.

[4] In the construction of the confidence intervals, 97.5^{th} percentile of $t(G\rho_c - 1)$ is used as the critical value.

next two rows report the average of the usual cluster-robust standard error and the corresponding coverage rate of the 95% confidence interval.[8] Consistent with the theory, the usual cluster-robust standard errors are always larger than the Monte Carlo standard deviation of the APE estimator. Consequently, the coverage rates of the confidence intervals are always above its nominal level. The discrepancy between the usual cluster-robust standard error and the standard deviation is the smallest when the sampling probability is 0.1, as this is the case closest to sampling from an infinite number of clusters.

The sixth and seventh rows report the average of the adjusted finite population cluster-robust standard error and the coverage rate of the corresponding 95% confidence interval. Since the fixed attribute z_{igM} is correlated with the score function, the adjusted finite population cluster-robust standard error is quite a bit smaller than the usual infinite population cluster-robust standard error, making the coverage rate of the confidence interval closer to its nominal level.

The averages of the adjusted finite population EHW standard error are reported in the last row, which are almost always smaller than the Monte Carlo standard deviations except in column (3).[9] For the design in column (3), the adjusted finite population EHW standard error is smaller than the cluster-robust standard error but larger than the standard deviation. Therefore, when there is neither cluster assignment nor cluster sampling, the usual cluster-robust standard error is overly conservative because the population is incorrectly treated as infinite and the clustering is unnecessary even though there are common unobserved components within clusters.

[8] The 97.5^{th} percentile of $t(G\rho_c - 1)$ is used as the critical value in constructing the confidence intervals. The coverage rates here, especially those in the top panel, should be interpreted with caution due to the cluster heterogeneity and the relatively small cluster number in the sample. Nevertheless, since I use the same set of critical values across confidence intervals, it is still fare to compare their coverage rates resulted from different standard errors.

[9] The adjusted EHW standard errors are obtained using the regression-based approach in Theorem 3.1 without the clustering terms.

All in all, we can conclude from the simulation results that the usual cluster-robust standard error is overly conservative unless the sample is a small proportion of the number of clusters in the population. When there are fixed attributes available, they can be used to estimate an uppper bound of the finite population CRAV. Although the adjusted finite population cluster-robust standard error is still conservative, it often improves over the usual cluster-robust standard error.

6 Application

The adjusted finite population CRVE proposed in Theorem 3.2 is applied to Antecol, Bedard, and Stearns (2018), who study the effects of tenure clock stopping policies on tenure rates among assistant professors. The unique dataset collected by them contains all assistant professor hires at the top-50 Economics departments from 1980-2005 as pooled cross sections, resulting in 1,392 observations in total. Furthermore, the tenure clock stopping policies are assigned at the university level while the data are collected at the individual level, implying that we have a setting of observing the entire population with cluster assignment.[10] The standard errors in Antecol et al. (2018) are clustered at the policy university level, which is the correct level to cluster the standard errors as implied by Corollary 1. As a result, there are 49 clusters in total with cluster sizes ranging from 11 to 57.

Since the dependent variable is a binary response, I analyze the linear probability model (LPM) given in Antecol et al. (2018) along with an additional probit model given in (24)

[10]This group of assistant professors is treated as the population.

below, which adopts the same notation from their paper.

$$
\begin{aligned}
&P(Y_{ugit} = 1|GN_{ut}, F_{ugit}, E_{ut}, FO_{ut}, X_{ugit}, Z_{ut}, \rho_{gt}, \psi_{ug}) = \\
&\Phi(\beta_0 + \beta_1 GN_{ut} + \beta_2 GN_{ut} \times F_{ugit} + \beta_3 GN_{ut} \times E_{ut} + \beta_4 GN_{ut} \times E_{ut} \times F_{ugit} \\
&\quad + \beta_5 FO_{ut} + \beta_6 FO_{ut} \times F_{ugit} + \beta_7 FO_{ut} \times E_{ut} + \beta_8 FO_{ut} \times E_{ut} \times F_{ugit} \\
&\quad + X_{ugit}\xi + Z_{ut}\eta + \rho_{gt} + \psi_{ug})
\end{aligned}
\tag{24}
$$

The dependent variable Y is an indicator of obtaining tenure at the university of initial placement. Binary variables GN and FO are indicators of gender-neutral and female-only tenure clock stopping policies respectively. The dummy variable F is the indicator for females. The variable E is an indicator of starting jobs in years zero through three after policy adoption. The vector X contains individual characteristics and the vector Z includes university level controls.[11] The parameter ρ captures gender-specific time trend and ψ represents gender-specific university heterogeneity. The subscripts, u, g, i, t, are indicators for university, gender, individual, and the year the job started, respectively.

Antecol et al. (2018) include gender-specific university dummies to capture different unobserved university heterogeneity for males and females. Adding group dummies in the linear model is equivalent to performing fixed effects with clustered data. However, adding group dummies in the nonlinear model may cause the incidental parameter problem. Since the cluster sizes are unbalanced, I use pooled probit with correlated random effects as suggested by Wooldridge (2010) to allow correlation between the gender-specific university heterogeneity and the covariates. Using Chamberlain-Mundlak device, the cluster size, the gender-specific university averages of individual and university characteristics and policies, and their interactions with cluster sizes are included as additional controls.

Given the probit model above is a nonlinear "difference-in-differences" model, the com-

[11] I refer to Antecol et al. (2018) for the details of the variables included as controls.

Table 2: Effects of Clock Stopping Policies on the Probability of Tenure at the University of Initial Placement

	LPM			Probit		
	APE	Standard Error		APE	Standard Error	
		inf pop	finite pop		inf pop	finite pop
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Policy Effects Years 0-3						
Men FOCS	-0.0085	0.0670	0.0616	-0.0068	0.0621	0.0557
Women FOCS	0.1723	0.1405	0.1191	0.1454	0.1771	0.1503
Men GNCS	0.0511	0.0787	0.0757	0.0446	0.0725	0.0699
Women GNCS	-0.0166	0.1071	0.0959	0.0220	0.1033	0.0957
Panel B. Policy Effects Years 4+						
Men FOCS	0.0023	0.0747	0.0701	-0.0055	0.0650	0.0606
Women FOCS	0.0493	0.1015	0.0797	0.0415	0.0891	0.0680
Men GNCS	0.1757	0.0826	0.0794	0.1537	0.0764	0.0732
Women GNCS	-0.1945	0.1057	0.0899	-0.1856	0.0989	0.0865

[1] Standard errors are clustered at the university level.
[2] Columns (1) and (4) report the APEs under the linear probability model and the correlated random effects probit model, respectively; columns (2) and (5) report the usual infinite population cluster-robust standard errors of the APE estimators (linear functions of the coefficient estimators in the case of the LPM); columns (3) and (6) report the adjusted finite population cluster-robust standard errors of the APE estimators.
[3] I refer to Antecol et al. (2018) for detailed control variables.

mon trend assumption is imposed on the latent outcome variable following Lechner (2011) and Puhani (2012). The treatment effects are defined as the differences in the probit probabilities when the treatment variables equal one or zero. I report the average of the treatment effect for those actually treated by the specific policy.[12] Assume that ψ conditional on the sufficient statistics (the additional controls included) follows a normal distribution, APEs can be obtained via pooled probit.

In Table 2, panel A presents the total effects for men and women hired in years zero through three after policy adoption, and panel B shows the effects for those employed in years four or later. The left panel (columns (1)-(3)) summarizes the results under the

[12] This may or may not be the true average treatment effect on the treated.

LPM. Columns (1) and (2) report the total effects and the standard errors, as shown in column (1) of Table 2 in Antecol et al. (2018), while column (3) reports the adjusted finite population clustered standard errors. The coefficients (APEs) are interpreted as the policy effect on the tenure attainment of the assistant professors compared with those of the same genders at the same university but without any clock stopping policies. For example, the coefficient in the third row of panel B shows that "men whose first job was at a top-50 university with a gender-neutral tenure clock stopping policy in place for more than three years have a 17.6 percentage point tenure rate advantage over men at the same university prior to the implementation of any policy" (Antecol et al., 2018, p. 2429-2430).

To estimate the adjusted finite population CRAV, I sum all the estimated score functions and control variables within clusters and apply the variance estimator in the left-hand side of (20) together with the usual estimator of the Hessian matrix. Since the number of control variables exceeds the number of clusters in the data, I only include university characteristics as the fixed attributes in the linear projection, resulting in a linear regression with 49 observations and eight independent variables. Compared with the usual cluster-robust standard errors, the finite population cluster-robust standard errors shrink by about 4% to 21% across the eight treatment groups. In terms of the statistical significance, the effect of gender-neutral policy for women hired three or more years after the policy adoption is significant at the 5% rather than the 10% level based on the adjusted finite population cluster-robust standard error. The same result holds when the critical values from $t(48)$ distribution are used.

In the right panel (columns (4)-(6)), we can see that the APEs from the probit regression are close in magnitudes to those from the linear model. The adjusted finite population CRAV is estimated applying Theorem 3.2 and the delta method.[13] Using the same set of

[13]When the entire population is observed, applying the delta method, in this case, is equivalent to the approach in Section 4 because the individual partial effect does not contain stochastic assignment variables.

university characteristics as the attribute variables, the reduction from the usual clustered standard errors to the finite population clustered standard errors ranges from 4% to 24%. Based on the critical values from $t(48)$, the effect of gender-neutral policy for women hired in later years is significant at the 5% level rather than the 10% level when the finite population clustered standard error is adopted.

Table B.1 in the online supplement provides empirical results under an alternative specification of the correlated random effects probit model, where the cluster size in the set of sufficient statistics is replaced by the dummy variables indicating different bins of cluster sizes. The APE estimates from the more flexible functional form are quite similar to those in the right panel of Table 2. The only exception is that the positive effect of female-only clock stopping policy on female assistant professors, in the early years of policy adoption, is significant at the 10% level when the finite population clustered standard error is used. Under this specification, the finite population clustered standard errors are smaller than the infinite population clustered standard errors by up to 23%.

To sum up, control variables can help shrink the standard errors when the population is treated as finite in both linear and nonlinear models. The empirical evidence suggests that gender-neutral tenure clock stopping policy is beneficial to men in obtaining tenured positions but detrimental to women. Furthermore, there is evidence that female-only policy helps women without hurting men, which is not found previously using the LPM and the infinite population clustered standard errors.

7 Conclusion

This paper develops finite population inference methods for M-estimators with clustered data. The takeaway for empirical practice is summarized as follows. One should only adjust standard errors for clustering if there is cluster sampling or cluster assignment. If

the number of clusters in the sample is minimal compared with the number of clusters in the population, one can use the usual cluster-robust standard error. However, if the sample contains a moderate fraction of clusters in the population or the entire population is used in the analysis, one can obtain M-estimators with smaller cluster-robust standard errors if the population is treated as finite rather than infinite. When there are control variables available, such as the baseline characteristics, they can be used to provide a better estimate of the finite population CRAV.

The current paper focuses on the asymptotics as the number of clusters tends to infinity in the limit. For a small number of clusters or wildly unbalanced clusters, the wild cluster bootstrap[14] has been proposed as a better-performing inference method for linear models in the setting of infinite populations. The finite population inference method for few heterogeneous clusters remains an interesting future research topic.

References

Abadie, A., Athey, S., Imbens, G.W., and Wooldridge, J.M. (2017), When should you adjust standard errors for clustering? Tech. rep., NBER Working Paper No. 24003.

Abadie, A., Athey, S., Imbens, G.W., and Wooldridge, J.M. (2020), Sampling-based versus design-based uncertainty in regression analysis. *Econometrica* 88, 265–296.

Antecol, H., Bedard, K., and Stearns, J. (2018), Equal but inequitable: Who benefits from gender-neutral tenure clock stopping policies? *American Economic Review* 108, 2420–2441.

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

[14] See, for example, Cameron, Gelbach, and Miller (2008) and MacKinnon and Webb (2017).

Bell, R.M. and McCaffrey, D.F. (2002), Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology* 28, 169–181.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004), How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics* 119, 249–275.

Bhattacharya, D. (2005), Asymptotic inference from multi-stage samples. *Journal of Econometrics* 126, 145–171.

Cameron, A.C., Gelbach, J.B., and Miller, D.L. (2008), Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics* 90, 414–427.

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

Carter, A.V., Schnepel, K.T., and Steigerwald, D.G. (2017), Asymptotic behavior of a t-test robust to cluster heterogeneity. *Review of Economics and Statistics* 99, 698–709.

Djogbenou, A.A., MacKinnon, J.G., and Nielsen, M.Ø. (2019), Asymptotic theory and wild bootstrap inference with clustered errors. *Journal of Econometrics* 212, 393–412.

Hansen, B.E. and Lee, S. (2019), Asymptotic theory for clustered samples. *Journal of Econometrics* 210, 268–290.

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

Kish, L. and Frankel, M.R. (1974), Inference from complex samples. *Journal of the Royal Statistical Society: Series B (Methodological)* 36, 1–37.

Kloek, T. (1981), OLS estimation in a model where a microvariable is explained by aggregates and contemporaneous disturbances are equicorrelated. *Econometrica* 49, 205–207.

Lechner, M. (2011), The estimation of causal effects by difference-in-difference methods. *Foundations and Trends in Econometrics* 4, 165–224.

Liang, K. and Zeger, S.L. (1986), Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13–22.

MacKinnon, J.G. (2019), How cluster-robust inference is changing applied econometrics. *Canadian Journal of Economics* 52, 851–881.

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

Moulton, B.R. (1986), Random group effects and the precision of regression estimates. *Journal of Econometrics* 32, 385–397.

Moulton, B.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.

Newey, W.K. (1991), Uniform convergence in probability and stochastic equicontinuity. *Econometrica* 59, 1161–1167.

Newey, W.K. and McFadden, D. (1994), Large sample estimation and hypothesis testing. In R.F. Engle and D.L. McFadden (eds.), *Handbook of Econometrics*, vol. 4, pp. 2111–2245, Elsevier.

Puhani, P.A. (2012), The treatment effect, the cross difference, and the interaction term in nonlinear "difference-in-differences" models. *Economics Letters* 115, 85–87.

Scott, A.J. and Holt, D. (1982), The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association* 77, 848–854.

White, H. (1984), *Asymptotic theory for econometricians*. Academic press.

Wooldridge, J.M. (2003), Cluster-sample methods in applied econometrics. *American Economic Review* 93, 133–138.

Wooldridge, J.M. (2010), *Econometric analysis of cross section and panel data (2nd ed.)*. MIT press.

Xu, R. (2020), Potential outcomes and finite-population inference for M-estimators. *Econometrics Journal* forthcoming.

A Notation and Regularity Conditions

The following notation provides details of the asymptotic variances and the variance estimator in Section 4:

$$\Delta^f_{ehw,M} = \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_X\Big\{\big[f_{iM}(W_{iM},\theta^*_M)-\gamma^*_M-F_M(\theta^*_M)H_M(\theta^*_M)^{-1}m_{iM}(W_{iM},\theta^*_M)\big]\cdot \big[f_{iM}(W_{iM},\theta^*_M)-\gamma^*_M-F_M(\theta^*_M)H_M(\theta^*_M)^{-1}m_{iM}(W_{iM},\theta^*_M)\big]'\Big\}, \tag{A.1}$$

$$\Delta^f_{E,M} = \frac{1}{M}\sum_{i=1}^{M}\Big\{\mathbb{E}_X\big[f_{iM}(W_{iM},\theta^*_M)-\gamma^*_M-F_M(\theta^*_M)H_M(\theta^*_M)^{-1}m_{iM}(W_{iM},\theta^*_M)\big]\cdot \mathbb{E}_X\big[f_{iM}(W_{iM},\theta^*_M)-\gamma^*_M-F_M(\theta^*_M)H_M(\theta^*_M)^{-1}m_{iM}(W_{iM},\theta^*_M)\big]'\Big\}, \tag{A.2}$$

$$\Delta^f_{cluster,M} = \frac{1}{M}\sum_{g=1}^{G}\sum_{i=1}^{M_g}\sum_{j\neq i}^{M_g}\mathbb{E}_X\Big\{\big[f_{igM}(W_{igM},\theta^*_M)-\gamma^*_M-F_M(\theta^*_M)H_M(\theta^*_M)^{-1}m_{igM}(W_{igM},\theta^*_M)\big]\cdot \big[f_{jgM}(W_{jgM},\theta^*_M)-\gamma^*_M-F_M(\theta^*_M)H_M(\theta^*_M)^{-1}m_{jgM}(W_{jgM},\theta^*_M)\big]'\Big\}, \tag{A.3}$$

$$\Delta^f_{EC,M} = \frac{1}{M}\sum_{g=1}^{G}\sum_{i=1}^{M_g}\sum_{j\neq i}^{M_g}\Big\{\mathbb{E}_X\big[f_{igM}(W_{igM},\theta^*_M) - \gamma^*_M - F_M(\theta^*_M)H_M(\theta^*_M)^{-1}m_{igM}(W_{igM},\theta^*_M)\big]\cdot$$
$$\mathbb{E}_X\Big[f_{jgM}(W_{jgM},\theta^*_M) - \gamma^*_M - F_M(\theta^*_M)H_M(\theta^*_M)^{-1}m_{jgM}(W_{jgM},\theta^*_M)\Big]'\Big\}, \tag{A.4}$$

$$\hat{\Delta}^f_{ehw,N} = \frac{1}{N}\sum_{i=1}^{M} R_{iM}\big[f_{iM}(W_{iM},\hat{\theta}_N) - \hat{\gamma}_N - \hat{F}_N(\hat{\theta}_N)\hat{H}_N(\hat{\theta}_N)^{-1}m_{iM}(W_{iM},\hat{\theta}_N)\big]\cdot$$
$$\big[f_{iM}(W_{iM},\hat{\theta}_N) - \hat{\gamma}_N - \hat{F}_N(\hat{\theta}_N)\hat{H}_N(\hat{\theta}_N)^{-1}m_{iM}(W_{iM},\hat{\theta}_N)\big]', \tag{A.5}$$

and

$$\hat{\Delta}^f_{cluster,N} = \frac{1}{N}\sum_{g=1}^{G}\sum_{i=1}^{M_g}\sum_{j\neq i}^{M_g} R_{igM}R_{jgM}\big[f_{igM}(W_{igM},\hat{\theta}_N) - \hat{\gamma}_N - \hat{F}_N(\hat{\theta}_N)\hat{H}_N(\hat{\theta}_N)^{-1}m_{igM}(W_{igM},\hat{\theta}_N)\big]\cdot$$
$$\big[f_{jgM}(W_{jgM},\hat{\theta}_N) - \hat{\gamma}_N - \hat{F}_N(\hat{\theta}_N)\hat{H}_N(\hat{\theta}_N)^{-1}m_{jgM}(W_{jgM},\hat{\theta}_N)\big]', \tag{A.6}$$

where

$$F_M(\theta) = \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_X\big[\nabla_\theta f_{iM}(W_{iM},\theta)\big] \tag{A.7}$$

and

$$\hat{F}_N(\theta) = \frac{1}{N}\sum_{i=1}^{M} R_{iM}\nabla_\theta f_{iM}(W_{iM},\theta). \tag{A.8}$$

I impose the following regularity conditions for the theorems in the paper.

Assumption A.1. *(i) Suppose* $Q(\theta) \equiv \lim_{M\to\infty}\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_X\big[q_{iM}(W_{iM},\theta)\big]$ *exists and is uniquely minimized at* θ^*;[15] *(ii)* Θ *is compact; (iii)* $q_{iM}(w,\theta)$ *is continuous in* θ *for all* w *in the support of* W_{iM}, $\forall\ i,M$; *(iv)* $\sup_{i,M}\mathbb{E}_X\Big[\sup_{\theta\in\Theta}|q_{iM}(W_{iM},\theta)|^r\Big] < \infty$ *for some* $r > 1$; *(v) there is* $h(u)\downarrow 0$ *as* $u\downarrow 0$ *and* $b_1(\cdot):\mathcal{W}\to R$ *such that* $\sup_{i,M}\mathbb{E}_X\big[b_{1,iM}(W_{iM})\big] < \infty$, *and for all* $\tilde{\theta},\theta\in\Theta$, $\big|q_{iM}(W_{iM},\tilde{\theta}) - q_{iM}(W_{iM},\theta)\big| \le b_{1,iM}(W_{iM})h(\|\tilde{\theta}-\theta\|)$.

Assumption A.2. *suppose that* $\frac{1}{N}\sum_{i=1}^{M} R_{iM}\cdot m_{iM}(W_{iM},\hat{\theta}_N) = o_p(N^{-1/2})$ *and (i)* $\theta^* \in$

[15] The introduction of θ^* and the assumption of the existence of $Q(\theta)$ is not needed for what follows, but it entails little loss of generality and simplifies regularity conditions.

int(Θ)*; (ii)* $q_{iM}(w,\theta)$ *is twice continuously differentiable on int*(Θ) *for all* w *in the support of* W_{iM}*,* $\forall\ i, M$*; (iii)* $\sup_{i,M}\mathbb{E}_X\Big[\sup_{\theta\in\Theta}\|m_{iM}(W_{iM},\theta)\|^r\Big] < \infty$ *for some* $r > 2$*; (iv)* $\Delta_{ehw,M}(\theta^*_M) - \rho_{uM}\rho_{cM}\Delta_{E,M} + \rho_{uM}\Delta_{cluster,M}(\theta^*_M) - \rho_{uM}\rho_{cM}\Delta_{EC,M}$ *is nonsingular; (v)* $\sup_{i,M}\mathbb{E}_X\Big[\sup_{\theta\in\Theta}\|\nabla_\theta m_{iM}(W_{iM},\theta)\|^r\Big] < \infty$ *for some* $r > 1$*; (vi) there is* $h(u)\downarrow 0$ *as* $u\downarrow 0$ *and* $b_2(\cdot):\mathcal{W}\to R$ *such that* $\sup_{i,M}\mathbb{E}_X\big[b_{2,iM}(W_{iM})\big] < \infty$*, and for all* $\tilde\theta,\theta\in\Theta$*,* $\Big\|\nabla_\theta m_{iM}(W_{iM},\tilde\theta) - \nabla_\theta m_{iM}(W_{iM},\theta)\Big\| \le b_{2,iM}(W_{iM})h(\|\tilde\theta-\theta\|)$*; (vii)* $H_M(\theta^*_M)$ *is nonsingular; (viii) there is* $h(u)\downarrow 0$ *as* $u\downarrow 0$ *and* $b_3(\cdot):\mathcal{W}\to R$ *such that* $\sup_{i,M}\mathbb{E}_X\big[b_{3,iM}(W_{iM})^2\big] < \infty$*, and for all* $\tilde\theta,\theta\in\Theta$*,* $\Big\|m_{iM}(W_{iM},\tilde\theta) - m_{iM}(W_{iM},\theta)\Big\| \le b_{3,iM}(W_{iM})h(\|\tilde\theta-\theta\|)$.

Assumption A.3. *suppose that (i)* $f_{iM}(w,\theta)$ *is continuously differentiable on int*(Θ) *for all* w *in the support of* W_{iM}*,* $\forall\ i, M$*; (ii)* $\sup_{i,M}\mathbb{E}_X\Big[\sup_{\theta\in\Theta}\|f_{iM}(W_{iM},\theta)\|^r\Big] < \infty$ *for some* $r > 2$*; (iii)* $V_{f,M}$ *is nonsingular; (iv)* $\sup_{i,M}\mathbb{E}_X\Big[\sup_{\theta\in\Theta}\|\nabla_\theta f_{iM}(W_{iM},\theta)\|^r\Big] < \infty$ *for some* $r > 1$*; (v) there is* $h(u)\downarrow 0$ *as* $u\downarrow 0$ *and* $b_4(\cdot):\mathcal{W}\to R$ *such that* $\sup_{i,M}\mathbb{E}_X\big[b_{4,iM}(W_{iM})\big] < \infty$*, and for all* $\tilde\theta,\theta\in\Theta$*,* $\Big\|\nabla_\theta f_{iM}(W_{iM},\tilde\theta) - \nabla_\theta f_{iM}(W_{iM},\theta)\Big\| \le b_{4,iM}(W_{iM})h(\|\tilde\theta-\theta\|)$*; (vi) there is* $h(u)\downarrow 0$ *as* $u\downarrow 0$ *and* $b_5(\cdot):\mathcal{W}\to R$ *such that* $\sup_{i,M}\mathbb{E}_X\big[b_{5,iM}(W_{iM})^2\big] < \infty$*, and for all* $\tilde\theta,\theta\in\Theta$*,* $\Big\|f_{iM}(W_{iM},\tilde\theta) - f_{iM}(W_{iM},\theta)\Big\| \le b_{5,iM}(W_{iM})h(\|\tilde\theta-\theta\|)$.

B Proof

In the following proofs, C denotes a generic positive constant that may be different in different circumstances.

Proof of Theorem 2.1:

To prove Theorem 2.1, I proceed by verifying the conditions of Theorem 2.1 in Newey and McFadden (1994).

Their first two conditions are the same as conditions (i) and (ii) in Assumption A.1. Their condition (iii) holds under conditions (iii) and (iv) in Assumption A.1 by the dom-

inated convergence theorem (DCT) and Jensen's inequality. To show their condition (iv) holds under the conditions in Theorem 2.1, first note that

$$\frac{1}{N}\sum_{i=1}^{M} R_{iM} q_{iM}(W_{iM},\theta) = \frac{M\rho_{uM}\rho_{cM}}{N}\frac{1}{M}\sum_{i=1}^{M}\frac{R_{iM}}{\rho_{uM}\rho_{cM}} q_{iM}(W_{iM},\theta). \tag{B.1}$$

By Lemma A.1 in the supplement and the continuous mapping theorem, $\frac{M\rho_{uM}\rho_{cM}}{N} \xrightarrow{p} 1$. Hence, it is sufficient to show that for each $\theta \in \Theta$

$$\left\| \frac{1}{M}\sum_{i=1}^{M}\frac{R_{iM}}{\rho_{uM}\rho_{cM}} q_{iM}(W_{iM},\theta) - \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_X\big[q_{iM}(W_{iM},\theta)\big] \right\| \xrightarrow{p} 0. \tag{B.2}$$

Condition (iv) in Assumption A.1 implies $\forall\ \theta \in \Theta$

$$\sup_{i,M}\mathbb{E}_X\left[\left|\frac{R_{iM}}{\rho_{uM}\rho_{cM}} q_{iM}(W_{iM},\theta)\right|^r\right] \le \frac{1}{(\rho_{uM}\rho_{cM})^{r-1}}\sup_{i,M}\mathbb{E}_X\Big[\sup_{\theta\in\Theta}|q_{iM}(W_{iM},\theta)|^r\Big] < \infty \tag{B.3}$$

for some $r > 1$, which further implies

$$\lim_{C\to\infty}\sup_{i,M}\left\{\mathbb{E}\left[\left|\frac{R_{iM}}{\rho_{uM}\rho_{cM}} q_{iM}(W_{iM},\theta)\right|\cdot \mathbb{1}\left(\left|\frac{R_{iM}}{\rho_{uM}\rho_{cM}} q_{iM}(W_{iM},\theta)\right| > C\right)\right]\right\} = 0. \tag{B.4}$$

(B.2) thus follows by Theorem 1 in Hansen and Lee (2019) under Assumption 4. As a result, condition (iv) in Newey and McFadden (1994) holds by Lemma A.2 in the supplement and Corollary 2.2 in Newey (1991) under condition (v) in Assumption A.1. Because $\theta_M^* - \theta^* \to \mathbf{0}$ by definition, $\hat{\theta}_N - \theta_M^* \xrightarrow{p} \mathbf{0}$.

Proof of Theorem 2.2:

The proof is modification of the proof of Theorem 11 in Hansen and Lee (2019).

I start by verifying that

$$\sum_{i=1}^{M} \mathbb{E}_X\big[m_{iM}(W_{iM}, \theta_M^*)\big] = \mathbf{0}, \tag{B.5}$$

which holds by Lemma 3.6 in Newey and McFadden (1994) and Jensen's inequality under conditions (ii) and (iii) in Assumption A.2.

By the element-by-element mean value expansion around θ_M^*,

$$\begin{aligned} o_p(N^{-1/2}) &= V_M^{-1/2} \frac{1}{N} \sum_{i=1}^{M} R_{iM} \cdot m_{iM}(W_{iM}, \hat{\theta}_N) \\ =& V_M^{-1/2} \frac{1}{N} \sum_{i=1}^{M} R_{iM} \cdot m_{iM}(W_{iM}, \theta_M^*) + V_M^{-1/2} \frac{1}{N} \sum_{i=1}^{M} R_{iM} \nabla_\theta m_{iM}(W_{iM}, \check{\theta})(\hat{\theta}_N - \theta_M^*), \end{aligned} \tag{B.6}$$

where $\check{\theta}$ lies on the line segment connecting θ_M^* and $\hat{\theta}_N$.

I first show

$$\hat{H}_N(\check{\theta}) = H_M(\theta_M^*)\big(I_k + o_p(1)\big). \tag{B.7}$$

Since we can write

$$\hat{H}_N(\check{\theta}) = H_M(\theta_M^*)\Big[I_k + H_M(\theta_M^*)^{-1}\big(\hat{H}_N(\check{\theta}) - H_M(\theta_M^*)\big)\Big], \tag{B.8}$$

it suffices to show

$$\Big\| H_M(\theta_M^*)^{-1}\big(\hat{H}_N(\check{\theta}) - H_M(\theta_M^*)\big)\Big\| \xrightarrow{p} 0. \tag{B.9}$$

We can write

$$\begin{aligned} \hat{H}_N(\theta) =& \frac{M\rho_{uM}\rho_{cM}}{N} \frac{1}{M} \sum_{i=1}^{M} \frac{R_{iM}}{\rho_{uM}\rho_{cM}} \nabla_\theta m_{iM}(W_{iM}, \theta) \\ =& \big(1 + o_p(1)\big) \frac{1}{M} \sum_{i=1}^{M} \frac{R_{iM}}{\rho_{uM}\rho_{cM}} \nabla_\theta m_{iM}(W_{iM}, \theta). \end{aligned} \tag{B.10}$$

Since $\forall\ \theta \in \Theta$

$$\begin{aligned}
&\sup_{i,M} \mathbb{E}_X \left[\left\| \frac{R_{iM}}{\rho_{uM}\rho_{cM}} \nabla_\theta m_{iM}(W_{iM}, \theta) \right\|^r \right] \\
\leq &\frac{1}{(\rho_{uM}\rho_{cM})^{r-1}} \sup_{i,M} \mathbb{E}_X \left[\sup_{\theta \in \Theta} \|\nabla_\theta m_{iM}(W_{iM}, \theta)\|^r \right] < \infty
\end{aligned} \tag{B.11}$$

for some $r > 1$,

$$\left\| \frac{1}{M} \sum_{i=1}^{M} \frac{R_{iM}}{\rho_{uM}\rho_{cM}} \nabla_\theta m_{iM}(W_{iM}, \theta) - H_M(\theta) \right\| \xrightarrow{p} 0 \tag{B.12}$$

by Theorem 1 in Hansen and Lee (2019) under Assumption 4 (implied by Assumption 5) and condition (v) in Assumption A.2. Note that $H_M(\theta)$ is continuous in θ for all M by the DCT and Jensen's inequality under conditions (ii) and (v) in Assumption A.2. By Corollary 2.2 in Newey (1991) and Lemma A.2 in the supplement,

$$\begin{aligned}
&\left\| H_M(\theta_M^*)^{-1} \big(\hat{H}_N(\check{\theta}) - H_M(\theta_M^*) \big) \right\| \\
\leq &C \left(\sup_{\theta \in \Theta} \left\| \hat{H}_N(\theta) - H_M(\theta) \right\| + \left\| H_M(\check{\theta}) - H_M(\theta_M^*) \right\| \right) \xrightarrow{p} 0
\end{aligned} \tag{B.13}$$

under conditions (vi) and (vii) in Assumption A.2.

(B.7) implies

$$\hat{H}_N(\check{\theta})^{-1} = H_M(\theta_M^*)^{-1}(I_k + o_p(1)). \tag{B.14}$$

Using (B.14), (B.6) can be written as

$$\begin{aligned}
V_M^{-1/2}\sqrt{N}(\hat{\theta}_N - \theta_M^*) = &- V_M^{-1/2} H_M(\theta_M^*)^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{M} R_{iM} \cdot m_{iM}(W_{iM}, \theta_M^*) \\
&- V_M^{-1/2} H_M(\theta_M^*)^{-1} o_p(1) \frac{1}{\sqrt{N}} \sum_{i=1}^{M} R_{iM} \cdot m_{iM}(W_{iM}, \theta_M^*) + o_p(1).
\end{aligned} \tag{B.15}$$

We can write

$$\begin{aligned}\frac{1}{\sqrt{N}}\sum_{i=1}^{M} R_{iM}\cdot m_{iM}(W_{iM},\theta_M^*) &= \sqrt{\frac{M\rho_{uM}\rho_{cM}}{N}}\frac{1}{\sqrt{M}}\sum_{i=1}^{M}\frac{R_{iM}}{\sqrt{\rho_{uM}\rho_{cM}}}m_{iM}(W_{iM},\theta_M^*)\\ &= \big(1+o_p(1)\big)\frac{1}{\sqrt{M}}\sum_{i=1}^{M}\frac{R_{iM}}{\sqrt{\rho_{uM}\rho_{cM}}}m_{iM}(W_{iM},\theta_M^*).\end{aligned} \tag{B.16}$$

Plug (B.16) into (B.15), we have

$$\begin{aligned}&V_M^{-1/2}\sqrt{N}(\hat{\theta}_N-\theta_M^*)\\ =&-V_M^{-1/2}H_M(\theta_M^*)^{-1}\frac{1}{\sqrt{M}}\sum_{i=1}^{M}\frac{R_{iM}}{\sqrt{\rho_{uM}\rho_{cM}}}m_{iM}(W_{iM},\theta_M^*)\\ &-V_M^{-1/2}H_M(\theta_M^*)^{-1}\frac{1}{\sqrt{M}}\sum_{i=1}^{M}\frac{R_{iM}}{\sqrt{\rho_{uM}\rho_{cM}}}m_{iM}(W_{iM},\theta_M^*)\cdot o_p(1)+o_p(1).\end{aligned} \tag{B.17}$$

Since

$$\begin{aligned}&\mathbb{V}_X\left(\frac{1}{\sqrt{M}}\sum_{i=1}^{M}\frac{R_{iM}}{\sqrt{\rho_{uM}\rho_{cM}}}m_{iM}(W_{iM},\theta_M^*)\right)\\ =&\frac{1}{M\rho_{uM}\rho_{cM}}\Bigg\{\sum_{i=1}^{M}\mathbb{V}_X\big[R_{iM}\cdot m_{iM}(W_{iM},\theta_M^*)\big]\\ &+\sum_{g=1}^{G}\sum_{i=1}^{M_g}\sum_{j\neq i}^{M_g}\mathbb{COV}_X\big[R_{igM}\cdot m_{igM}(W_{igM},\theta_M^*),R_{jgM}\cdot m_{jgM}(W_{jgM},\theta_M^*)\big]\Bigg\}\\ =&\frac{1}{M\rho_{uM}\rho_{cM}}\Bigg\{\sum_{i=1}^{M}\Big[\mathbb{E}_X\big(R_{iM}\cdot m_{iM}(W_{iM},\theta_M^*)m_{iM}(W_{iM},\theta_M^*)'\big)\\ &-\mathbb{E}_X\big(R_{iM}\cdot m_{iM}(W_{iM},\theta_M^*)\big)\mathbb{E}_X\big(R_{iM}\cdot m_{iM}(W_{iM},\theta_M^*)\big)'\Big]\\ &+\sum_{g=1}^{G}\sum_{i=1}^{M_g}\sum_{j\neq i}^{M_g}\Big[\mathbb{E}_X\big(R_{igM}R_{jgM}\cdot m_{igM}(W_{igM},\theta_M^*)m_{jgM}(W_{jgM},\theta_M^*)'\big)\\ &-\mathbb{E}_X\big(R_{igM}\cdot m_{igM}(W_{igM},\theta_M^*)\big)\mathbb{E}_X\big(R_{jgM}\cdot m_{jgM}(W_{jgM},\theta_M^*)\big)'\Big]\Bigg\}\end{aligned}$$

$$
\begin{aligned}
=&\frac{1}{M}\Bigg\{\sum_{i=1}^{M}\Big[\mathbb{E}_X\big(m_{iM}(W_{iM},\theta_M^*)m_{iM}(W_{iM},\theta_M^*)'\big)\\
&-\rho_{uM}\rho_{cM}\mathbb{E}_X\big(m_{iM}(W_{iM},\theta_M^*)\big)\mathbb{E}_X\big(m_{iM}(W_{iM},\theta_M^*)\big)'\Big]\\
&+\sum_{g=1}^{G}\sum_{i=1}^{M_g}\sum_{j\neq i}^{M_g}\Big[\rho_{uM}\mathbb{E}_X\big(m_{igM}(W_{igM},\theta_M^*)m_{jgM}(W_{jgM},\theta_M^*)'\big)\\
&-\rho_{uM}\rho_{cM}\mathbb{E}_X\big(m_{igM}(W_{igM},\theta_M^*)\big)\mathbb{E}_X\big(m_{jgM}(W_{jgM},\theta_M^*)\big)'\Big]\Bigg\}\\
&=\Delta_{ehw,M}(\theta_M^*)-\rho_{uM}\rho_{cM}\Delta_{E,M}+\rho_{uM}\Delta_{cluster,M}(\theta_M^*)-\rho_{uM}\rho_{cM}\Delta_{EC,M},
\end{aligned}
\tag{B.18}
$$

we have

$$
\mathbb{V}_X\left(V_M^{-1/2}H_M(\theta_M^*)^{-1}\frac{1}{\sqrt{M}}\sum_{i=1}^{M}\frac{R_{iM}}{\sqrt{\rho_{uM}\rho_{cM}}}m_{iM}(W_{iM},\theta_M^*)\right)=I_k. \tag{B.19}
$$

Given $\forall\ \theta\in\Theta$

$$
\sup_{i,M}\mathbb{E}_X\left[\left\|\frac{R_{iM}}{\sqrt{\rho_{uM}\rho_{cM}}}m_{iM}(W_{iM},\theta)\right\|^r\right]\leq\frac{1}{(\rho_{uM}\rho_{cM})^{r/2-1}}\sup_{i,M}\mathbb{E}_X\Big[\sup_{\theta\in\Theta}\|m_{iM}(W_{iM},\theta)\|^r\Big]<\infty \tag{B.20}
$$

for some $r>2$ under condition (iii) in Assumption A.2,

$$
V_M^{-1/2}H_M(\theta_M^*)^{-1}\frac{1}{\sqrt{M}}\sum_{i=1}^{M}\frac{R_{iM}}{\sqrt{\rho_{uM}\rho_{cM}}}m_{iM}(W_{iM},\theta_M^*)\xrightarrow{d}\mathcal{N}(\mathbf{0},I_k) \tag{B.21}
$$

by Theorem 2 in Hansen and Lee (2019) under Assumption 5 and condition (iv) in Assumption A.2.

Because of (B.21),

$$
\begin{aligned}
V_M^{-1/2}\sqrt{N}(\hat{\theta}_N-\theta_M^*)=&-V_M^{-1/2}H_M(\theta_M^*)^{-1}\frac{1}{\sqrt{M}}\sum_{i=1}^{M}\frac{R_{iM}}{\sqrt{\rho_{uM}\rho_{cM}}}m_{iM}(W_{iM},\theta_M^*)\\
&+o_p(1)O_p(1)+o_p(1)\xrightarrow{d}\mathcal{N}(\mathbf{0},I_k).
\end{aligned}
\tag{B.22}
$$

As for Theorem 2.2(2), it is equivalent to show $\left\|V_{SM}^{-1/2}\hat{V}_{SN}V_{SM}^{-1/2}-I_k\right\| \xrightarrow{p} 0$.

Since (B.14) holds by replacing $\check{\theta}$ with $\hat{\theta}_N$,

$$\hat{H}_N(\hat{\theta}_N)^{-1} = H_M(\theta_M^*)^{-1}\big(I_k + o_p(1)\big). \tag{B.23}$$

We can write

$$\begin{aligned}
&\hat{\Delta}_{ehw,N}(\theta) + \hat{\Delta}_{cluster,N}(\theta)\\
=&\frac{1}{N}\sum_{g=1}^{G}\left[\sum_{i=1}^{M_g} R_{igM}\cdot m_{igM}(W_{igM},\theta)\right]\left[\sum_{i=1}^{M_g} R_{igM}\cdot m_{igM}(W_{igM},\theta)\right]'\\
=&\frac{M\rho_{uM}\rho_{cM}}{N}\frac{1}{M}\sum_{g=1}^{G}\left[\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}} m_{igM}(W_{igM},\theta)\right]\left[\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}} m_{igM}(W_{igM},\theta)\right]'\\
=&\big(1+o_p(1)\big)\frac{1}{M}\sum_{g=1}^{G}\left[\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}} m_{igM}(W_{igM},\theta)\right]\left[\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}} m_{igM}(W_{igM},\theta)\right]'.
\end{aligned} \tag{B.24}$$

Note that

$$\begin{aligned}
&\mathbb{E}_X\left\{\frac{1}{M}\sum_{g=1}^{G}\left[\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}} m_{igM}(W_{igM},\theta)\right]\left[\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}} m_{igM}(W_{igM},\theta)\right]'\right\}\\
=&\mathbb{E}_X\left[\frac{1}{M}\sum_{i=1}^{M}\frac{R_{iM}}{\rho_{uM}\rho_{cM}} m_{iM}(W_{iM},\theta)m_{iM}(W_{iM},\theta)'\right]\\
&+\mathbb{E}_X\left[\frac{1}{M}\sum_{g=1}^{G}\sum_{i=1}^{M_g}\sum_{j\neq i}^{M_g}\frac{R_{igM}R_{jgM}}{\rho_{uM}\rho_{cM}} m_{igM}(W_{igM},\theta)m_{jgM}(W_{jgM},\theta)'\right]\\
=&\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_X\big[m_{iM}(W_{iM},\theta)m_{iM}(W_{iM},\theta)'\big]\\
&+\frac{1}{M}\sum_{g=1}^{G}\sum_{i=1}^{M_g}\sum_{j\neq i}^{M_g}\rho_{uM}\mathbb{E}_X\big[m_{igM}(W_{igM},\theta)m_{jgM}(W_{jgM},\theta)'\big]
\end{aligned}$$

$$=\Delta_{ehw,M}(\theta)+\rho_{uM}\Delta_{cluster,M}(\theta). \tag{B.25}$$

Hence, $\forall\ \theta\in\Theta$

$$\begin{aligned}&\left\|\frac{1}{M}\sum_{g=1}^{G}\left[\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}}m_{igM}(W_{igM},\theta)\right]\left[\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}}m_{igM}(W_{igM},\theta)\right]'\right.\\&\left.-\left(\Delta_{ehw,M}(\theta)+\rho_{uM}\Delta_{cluster,M}(\theta)\right)\right\|\xrightarrow{p}0\end{aligned} \tag{B.26}$$

follows by (B.20) and the same proof of (62) in Hansen and Lee (2019) under Assumption 5. Also, $\Delta_{ehw,M}(\theta)+\rho_{uM}\Delta_{cluster,M}(\theta)$ is continuous in θ for all M by the DCT, Jensen's inequality, and Cauchy-Schwarz Inequality under conditions (ii) and (iii) in Assumption A.2. In addition,

$$\begin{aligned}&\left\|\hat{\Delta}_{ehw,N}(\tilde{\theta})+\hat{\Delta}_{cluster,N}(\tilde{\theta})-\left(\hat{\Delta}_{ehw,N}(\theta)+\hat{\Delta}_{cluster,N}(\theta)\right)\right\|\\\leq&\frac{1}{N}\sum_{g=1}^{G}\left\|\left[\sum_{i=1}^{M_g}R_{igM}\cdot m_{igM}(W_{igM},\tilde{\theta})\right]\left[\sum_{i=1}^{M_g}R_{igM}\cdot m_{igM}(W_{igM},\tilde{\theta})\right]'\right.\\&\left.-\left[\sum_{i=1}^{M_g}R_{igM}\cdot m_{igM}(W_{igM},\theta)\right]\left[\sum_{i=1}^{M_g}R_{igM}\cdot m_{igM}(W_{igM},\theta)\right]'\right\|\\\leq&\frac{1}{N}\sum_{g=1}^{G}2\sup_{\theta\in\Theta}\left\|\sum_{i=1}^{M_g}R_{igM}\cdot m_{igM}(W_{igM},\theta)\right\|\cdot\\&\left\|\sum_{i=1}^{M_g}R_{igM}\cdot m_{igM}(W_{igM},\tilde{\theta})-\sum_{i=1}^{M_g}R_{igM}\cdot m_{igM}(W_{igM},\theta)\right\|\\\leq&\frac{2}{N}\sum_{g=1}^{G}\sup_{\theta\in\Theta}\left\|\sum_{i=1}^{M_g}R_{igM}\cdot m_{igM}(W_{igM},\theta)\right\|\sum_{i=1}^{M_g}R_{igM}b_{3,igM}(W_{igM})h(\|\tilde{\theta}-\theta\|).\end{aligned} \tag{B.27}$$

under condition (viii) in Assumption A.2. Let

$$
\begin{aligned}
B_N^1 \equiv & \frac{2}{N}\sum_{g=1}^{G}\sup_{\theta\in\Theta}\left\|\sum_{i=1}^{M_g} R_{igM}\cdot m_{igM}(W_{igM},\theta)\right\|\left\|\sum_{i=1}^{M_g} R_{igM} b_{3,igM}(W_{igM})\right. \\
= & 2\frac{M\rho_{uM}\rho_{cM}}{N}\frac{1}{M}\sum_{g=1}^{G}\sup_{\theta\in\Theta}\left\|\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}} m_{igM}(W_{igM},\theta)\right\|\left\|\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}} b_{3,igM}(W_{igM})\right. \\
= & \big(1+o_p(1)\big)\frac{2}{M}\sum_{g=1}^{G}\sup_{\theta\in\Theta}\left\|\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}} m_{igM}(W_{igM},\theta)\right\|\left\|\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}} b_{3,igM}(W_{igM}).\right.
\end{aligned}
\tag{B.28}
$$

Since

$$
\mathbb{E}_X\left[\sup_{\theta\in\Theta}\left\|\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}} m_{igM}(W_{igM},\theta)\right\|^2\right] < CM_g^2 \tag{B.29}
$$

by Cr inequality and Jensen's inequality under condition (iii) in Assumption A.2,

$$
\begin{aligned}
& \mathbb{E}_X\left[\frac{2}{M}\sum_{g=1}^{G}\sup_{\theta\in\Theta}\left\|\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}} m_{igM}(W_{igM},\theta)\right\|\left\|\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}} b_{3,igM}(W_{igM})\right.\right] \\
\leq & \frac{2}{M}\sum_{g=1}^{G}\sum_{i=1}^{M_g}\left\{\mathbb{E}_X\left[\sup_{\theta\in\Theta}\left\|\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}} m_{igM}(W_{igM},\theta)\right\|^2\right]\right\}^{1/2}\cdot \\
& \left\{\mathbb{E}_X\left[\frac{R_{igM}}{\rho_{uM}\rho_{cM}} b_{3,igM}(W_{igM})^2\right]\right\}^{1/2} \\
\leq & C\frac{1}{M}\sum_{g=1}^{G} M_g^2 < \infty
\end{aligned}
\tag{B.30}
$$

by Cauchy-Schwarz inequality under Assumption 5 and condition (viii) in Assumption A.2. As a result, $B_N^1 = O_p(1)$ by Markov's inequality. Therefore, given condition (iv) in

Assumption A.2,

$$
\begin{aligned}
&\Big\| \big[\Delta_{ehw,M}(\theta_M^*) + \rho_{uM}\Delta_{cluster,M}(\theta_M^*)\big]^{-1} \\
&\big[\hat{\Delta}_{ehw,N}(\hat{\theta}_N) + \hat{\Delta}_{cluster,N}(\hat{\theta}_N) - \Delta_{ehw,M}(\theta_M^*) - \rho_{uM}\Delta_{cluster,M}(\theta_M^*)\big]\Big\| \\
\leq & C\Bigg(\sup_{\theta\in\Theta} \|\hat{\Delta}_{ehw,N}(\theta) + \hat{\Delta}_{cluster,N}(\theta) - \Delta_{ehw,M}(\theta) - \rho_{uM}\Delta_{cluster,M}(\theta)\| \\
&+ \Big\|\Delta_{ehw,M}(\hat{\theta}_N) + \rho_{uM}\Delta_{cluster,M}(\hat{\theta}_N) - \Delta_{ehw,M}(\theta_M^*) - \rho_{uM}\Delta_{cluster,M}(\theta_M^*)\Big\| \Bigg) = o_p(1)
\end{aligned} \tag{B.31}
$$

by Corollary 2.2 in Newey (1991) under $\hat{\theta}_N - \theta_M^* \xrightarrow{p} \mathbf{0}$ (implied by Theorem 2.1). Hence,

$$
\begin{aligned}
&\hat{\Delta}_{ehw,N}(\hat{\theta}_N) + \hat{\Delta}_{cluster,N}(\hat{\theta}_N) \\
=&\big(\Delta_{ehw,M}(\theta_M^*) + \rho_{uM}\Delta_{cluster,M}(\theta_M^*)\big)\Big[I_k + \big(\Delta_{ehw,M}(\theta_M^*) + \rho_{uM}\Delta_{cluster,M}(\theta_M^*)\big)^{-1} \\
&\big(\hat{\Delta}_{ehw,N}(\hat{\theta}_N) + \hat{\Delta}_{cluster,N}(\hat{\theta}_N) - \Delta_{ehw,M}(\theta_M^*) - \rho_{uM}\Delta_{cluster,M}(\theta_M^*)\big)\Big] \\
=&\big(\Delta_{ehw,M}(\theta_M^*) + \rho_{uM}\Delta_{cluster,M}(\theta_M^*)\big)\big(I_k + o_p(1)\big).
\end{aligned} \tag{B.32}
$$

Using (B.23) and (B.32),

$$
\begin{aligned}
&\Big\|V_{SM}^{-1/2}\hat{V}_{SN}V_{SM}^{-1/2} - I_k\Big\| \\
=&\Big\|V_{SM}^{-1/2}\hat{H}_N(\hat{\theta}_N)^{-1}\big(\hat{\Delta}_{ehw,N}(\hat{\theta}_N) + \hat{\Delta}_{cluster,N}(\hat{\theta}_N)\big)\hat{H}_N(\hat{\theta}_N)^{-1}V_{SM}^{-1/2} - I_k\Big\| \\
=&\Big\|V_{SM}^{-1/2}H_M(\theta_M^*)^{-1}\big(I_k + o_p(1)\big)\big(\Delta_{ehw,M}(\theta_M^*) + \rho_{uM}\Delta_{cluster,M}(\theta_M^*)\big)\big(I_k + o_p(1)\big)\cdot \\
&H_M(\theta_M^*)^{-1}\big(I_k + o_p(1)\big)V_{SM}^{-1/2} - I_k\Big\| \\
\leq&\Big\|V_{SM}^{-1/2}V_{SM}V_{SM}^{-1/2} - I_k\Big\| + \Big\|V_{SM}^{-1/2}V_{SM}V_{SM}^{-1/2}\Big\|\, o_p(1) \\
=&o_p(1).
\end{aligned} \tag{B.33}
$$

Hence the result.

Proof of Theorem 3.2:

Let

$$P_M = \left[\sum_{g=1}^{G} \tilde{z}'_{gM}\tilde{z}_{gM}\right]^{-1} \sum_{g=1}^{G} \tilde{z}'_{gM}\mathbb{E}_X\left[\tilde{m}_{gM}(\theta^*_M)\right]'. \tag{B.34}$$

To show $\left\|\hat{P}_N - P_M\right\| \xrightarrow{p} 0$, I first show

$$\left\|\frac{1}{N}\sum_{g=1}^{G} R_{gM}\tilde{z}'_{gM}\tilde{z}_{gM} - \frac{1}{M}\sum_{g=1}^{G} \tilde{z}'_{gM}\tilde{z}_{gM}\right\| \xrightarrow{p} 0. \tag{B.35}$$

We can write

$$\frac{1}{N}\sum_{g=1}^{G} R_{gM}\tilde{z}'_{gM}\tilde{z}_{gM} = \frac{M\rho_{cM}}{N}\frac{1}{M}\sum_{g=1}^{G}\frac{R_{gM}}{\rho_{cM}}\tilde{z}'_{gM}\tilde{z}_{gM}. \tag{B.36}$$

Since $\rho_{uM} = 1$, $\frac{M\rho_{cM}}{N} \xrightarrow{p} 1$. Hence, it suffices to show

$$\left\|\frac{1}{M}\sum_{g=1}^{G}\frac{R_{gM}}{\rho_{cM}}\tilde{z}'_{gM}\tilde{z}_{gM} - \frac{1}{M}\sum_{g=1}^{G} \tilde{z}'_{gM}\tilde{z}_{gM}\right\| \xrightarrow{p} 0, \tag{B.37}$$

Because for some $r > 2$

$$\sup_{i,g,M} \mathbb{E}_X\left(\left\|\frac{R_{gM}}{\sqrt{\rho_{cM}}} z_{igM}\right\|^r\right) < \infty \tag{B.38}$$

under condition (iii) in Theorem 3.2, (B.37) follows by the proof of (62) in Hansen and Lee (2019) under Assumption 5.

Next, I show

$$\left\|\frac{1}{N}\sum_{g=1}^{G} R_{gM}\tilde{m}_{gM}(\hat{\theta}_N)\tilde{z}_{gM} - \frac{1}{M}\sum_{g=1}^{G}\mathbb{E}_X\left[\tilde{m}_{gM}(\theta^*_M)\right]\tilde{z}_{gM}\right\| \xrightarrow{p} 0. \tag{B.39}$$

Again, we can write

$$\begin{aligned}\frac{1}{N}\sum_{g=1}^{G} R_{gM}\tilde{m}_{gM}(\hat{\theta}_N)\tilde{z}_{gM} =& \frac{M\rho_{cM}}{N}\frac{1}{M}\sum_{g=1}^{G}\frac{R_{gM}}{\rho_{cM}}\tilde{m}_{gM}(\hat{\theta}_N)\tilde{z}_{gM} \\ =& \big(1+o_p(1)\big)\frac{1}{M}\sum_{g=1}^{G}\frac{R_{gM}}{\rho_{cM}}\tilde{m}_{gM}(\hat{\theta}_N)\tilde{z}_{gM}\end{aligned} \tag{B.40}$$

As a first step, I show $\forall\ \theta \in \Theta$

$$\left\| \frac{1}{M}\sum_{g=1}^{G}\frac{R_{gM}}{\rho_{cM}}\tilde{m}_{gM}(\theta)\tilde{z}_{gM} - \frac{1}{M}\sum_{g=1}^{G}\mathbb{E}_X\big[\tilde{m}_{gM}(\theta)\big]\tilde{z}_{gM} \right\| \xrightarrow{p} 0. \tag{B.41}$$

Fix $\delta > 0$. Set $\epsilon = (\delta/C)^2$. Let

$$\tilde{l}_{gM} = \frac{R_{gM}}{\rho_{cM}}\tilde{m}_{gM}(\theta)\tilde{z}_{gM}\mathbb{1}\left(\frac{R_{gM}}{\rho_{cM}}\left\|\tilde{m}_{gM}(\theta)\tilde{z}_{gM}\right\| \le M\epsilon\right). \tag{B.42}$$

Then

$$\begin{aligned}&\mathbb{E}_X\left[\left\| \frac{1}{M}\sum_{g=1}^{G}\frac{R_{gM}}{\rho_{cM}}\tilde{m}_{gM}(\theta)\tilde{z}_{gM} - \frac{1}{M}\sum_{g=1}^{G}\mathbb{E}_X\big[\tilde{m}_{gM}(\theta)\big]\tilde{z}_{gM}\right\|\right] \\ \le& \frac{1}{M}\mathbb{E}_X\left\{\left\|\sum_{g=1}^{G}\Big[\tilde{l}_{gM} - \mathbb{E}_X\big(\tilde{l}_{gM}\big)\Big]\right\|\right\} \\ &+ \frac{2}{M}\sum_{g=1}^{G}\mathbb{E}_X\left[\left\|\tilde{m}_{gM}(\theta)\tilde{z}_{gM}\right\|\mathbb{1}\left(\frac{R_{gM}}{\rho_{cM}}\left\|\tilde{m}_{gM}(\theta)\tilde{z}_{gM}\right\| > M\epsilon\right)\right].\end{aligned} \tag{B.43}$$

Observe that

$$\begin{aligned}&\frac{1}{M}\mathbb{E}_X\left[\left\|\sum_{g=1}^{G}\Big(\tilde{l}_{gM} - \mathbb{E}_X\big(\tilde{l}_{gM}\big)\Big)\right\|\right] \\ \le& \frac{1}{M}\left\{\mathbb{E}_X\left[\left\|\sum_{g=1}^{G}\Big(\tilde{l}_{gM} - \mathbb{E}_X\big(\tilde{l}_{gM}\big)\Big)\right\|^2\right]\right\}^{1/2}\end{aligned}$$

$$\leq \frac{1}{M}\left\{\sum_{g=1}^{G}\mathbb{E}_X\left[\left\|\tilde{l}_{gM}\right\|^2\right]\right\}^{1/2} \leq (\epsilon C)^{1/2}\left(\frac{1}{M}\sum_{g=1}^{G}M_g^2\right)^{1/2} \leq \delta \tag{B.44}$$

by Jensen's inequality and Cr inequality under Assumption 5, condition (iii) in Assumption A.2, and condition (iii) in Theorem 3.2. Also for some $r > 1$

$$\begin{aligned}&\sup_{g,M}\mathbb{E}_X\left[\left\|\frac{R_{gM}}{\rho_{cM}}\tilde{m}_{gM}(\theta)\tilde{z}_{gM}/M_g^2\right\|^r\right]\\ \leq &\frac{1}{\rho_{cM}^{r-1}}\left\{\sup_{g,M}\mathbb{E}_X\left[\sup_{\theta\in\Theta}\|\tilde{m}_{gM}(\theta)/M_g\|^{2r}\right]\right\}^{1/2}\sup_{g,M}\|\tilde{z}_{gM}/M_g\|^r < \infty\end{aligned} \tag{B.45}$$

by Jensen's inequality under condition (iii) in Assumption A.2 and condition (iii) in Theorem 3.2. Hence, we can pick B sufficiently large so that

$$\sup_{g,M}\mathbb{E}_X\left[\left\|\frac{R_{gM}}{\rho_{cM}}\tilde{m}_{gM}(\theta)\tilde{z}_{gM}/M_g^2\right\|\mathbb{1}\left(\left\|\frac{R_{gM}}{\rho_{cM}}\tilde{m}_{gM}(\theta)\tilde{z}_{gM}/M_g^2\right\| > B\right)\right] \leq \frac{\delta}{C}. \tag{B.46}$$

Pick M large enough so that

$$\max_{g\leq G}\frac{M_g^2}{M} \leq \frac{\epsilon}{B}, \tag{B.47}$$

which is feasible under Assumption 5. Then,

$$\frac{2}{M}\sum_{g=1}^{G}\mathbb{E}_X\left[\|\tilde{m}_{gM}(\theta)\tilde{z}_{gM}\|\,\mathbb{1}\left(\frac{R_{gM}}{\rho_{cM}}\|\tilde{m}_{gM}(\theta)\tilde{z}_{gM}\| > M\epsilon\right)\right] \leq \frac{2}{M}\sum_{g=1}^{G}M_g^2\frac{\delta}{C} \leq 2\delta. \tag{B.48}$$

Combining (B.44) and (B.48), (B.41) holds by Markov's inequality.

Next,

$$\begin{aligned}&\left\|\frac{1}{N}\sum_{g=1}^{G}R_{gM}\left[\tilde{m}_{gM}(\tilde{\theta})\tilde{z}_{gM} - \tilde{m}_{gM}(\theta)\tilde{z}_{gM}\right]\right\|\\ \leq &\frac{1}{N}\sum_{g=1}^{G}R_{gM}\left\|\tilde{m}_{gM}(\tilde{\theta})\tilde{z}_{gM} - \tilde{m}_{gM}(\theta)\tilde{z}_{gM}\right\|\end{aligned}$$

$$
\begin{aligned}
=&\frac{1}{N}\sum_{g=1}^{G} R_{gM}\left\|\sum_{i=1}^{M_g}\sum_{j=1}^{M_g} m_{igM}(W_{igM},\tilde{\theta})z_{jgM}-\sum_{i=1}^{M_g}\sum_{j=1}^{M_g} m_{igM}(W_{igM},\theta)z_{jgM}\right\| \\
\leq&\frac{1}{N}\sum_{g=1}^{G} R_{gM}\sum_{i=1}^{M_g}\sum_{j=1}^{M_g}\left\|m_{igM}(W_{igM},\tilde{\theta})-m_{igM}(W_{igM},\theta)\right\|\cdot\|z_{jgM}\| \\
\leq&\frac{1}{N}\sum_{g=1}^{G} R_{gM}\sum_{i=1}^{M_g}\sum_{j=1}^{M_g} b_{3,igM}(W_{igM})\cdot\|z_{jgM}\|\cdot h(\|\tilde{\theta}-\theta\|)
\end{aligned}
\tag{B.49}
$$

Let

$$
\begin{aligned}
B_N^2 \equiv&\frac{1}{N}\sum_{g=1}^{G} R_{gM}\sum_{i=1}^{M_g}\sum_{j=1}^{M_g} b_{3,igM}(W_{igM})\cdot\|z_{jgM}\| \\
=&\frac{M\rho_{cM}}{N}\frac{1}{M}\sum_{g=1}^{G}\frac{R_{gM}}{\rho_{cM}}\sum_{i=1}^{M_g}\sum_{j=1}^{M_g} b_{3,igM}(W_{igM})\cdot\|z_{jgM}\| \\
=&\big(1+o_p(1)\big)\frac{1}{M}\sum_{g=1}^{G}\frac{R_{gM}}{\rho_{cM}}\sum_{i=1}^{M_g}\sum_{j=1}^{M_g} b_{3,igM}(W_{igM})\cdot\|z_{jgM}\|
\end{aligned}
\tag{B.50}
$$

Since

$$
\begin{aligned}
&\mathbb{E}_X\left[\frac{1}{M}\sum_{g=1}^{G}\frac{R_{gM}}{\rho_{cM}}\sum_{i=1}^{M_g}\sum_{j=1}^{M_g} b_{3,igM}(W_{igM})\cdot\|z_{igM}\|\right] \\
=&\frac{1}{M}\sum_{g=1}^{G}\mathbb{E}\Big(\frac{R_{gM}}{\rho_{cM}}\Big)\sum_{i=1}^{M_g}\sum_{j=1}^{M_g}\mathbb{E}_X\big[b_{3,igM}(W_{igM})\big]\,\|z_{igM}\| \\
\leq&\frac{1}{M}\sum_{g=1}^{G} M_g^2\sup_{i,g,M}\Big\{\mathbb{E}_X\big[b_{3,igM}(W_{igM})^2\big]\Big\}^{1/2}\sup_{i,g,M}\|z_{igM}\|<\infty
\end{aligned}
\tag{B.51}
$$

by Jensen's inequality under condition (viii) in Assumption A.2, condition (iii) in Theorem 3.2, and Assumption 5, $B_N^2 = O_p(1)$ by Markov's inequality. Also, $\frac{1}{M}\sum_{g=1}^{G}\mathbb{E}_X\big[\tilde{m}_{gM}(\theta)\big]\tilde{z}_{gM}$ is continuous in θ for all M by the DCT and Jensen's inequality under Assumption 5 and conditions (ii) and (iii) in Assumption A.2. As a result,

$$
\left\|\frac{1}{N}\sum_{g=1}^{G} R_{gM}\tilde{m}_{gM}(\hat{\theta}_N)\tilde{z}_{gM}-\frac{1}{M}\sum_{g=1}^{G}\mathbb{E}_X\big[\tilde{m}_{gM}(\theta_M^*)\big]\tilde{z}_{gM}\right\|
$$

$$
\begin{aligned}
&\leq \sup_{\theta\in\Theta} \left\| \frac{1}{N}\sum_{g=1}^{G} R_{gM}\tilde{m}_{gM}(\theta)\tilde{z}_{gM} - \frac{1}{M}\sum_{g=1}^{G}\mathbb{E}_X\big[\tilde{m}_{gM}(\theta)\big]\tilde{z}_{gM} \right\| \\
&\quad + \left\| \frac{1}{M}\sum_{g=1}^{G}\mathbb{E}_X\big[\tilde{m}_{gM}(\hat{\theta}_N)\big]\tilde{z}_{gM} - \frac{1}{M}\sum_{g=1}^{G}\mathbb{E}_X\big[\tilde{m}_{gM}(\theta^*_M)\big]\tilde{z}_{gM} \right\| \xrightarrow{p} 0. \qquad \text{(B.52)}
\end{aligned}
$$

follows by Corollary 2.2 in Newey (1991) under $\hat{\theta}_N - \theta^*_M \xrightarrow{p} \mathbf{0}$.

The result $\left\|\hat{P}_N - P_M\right\| \xrightarrow{p} 0$ is immediately implied by (B.35) and (B.39) under the continuity of inversion and multiplication.

Denote $\Delta^Z_{CE,M} \equiv \frac{1}{M}\sum_{g=1}^{G} P'_M\tilde{z}'_{gM}\tilde{z}_{gM}P_M$. We can write

$$
\begin{aligned}
\hat{\Delta}^Z_{CE,N} =& \frac{M\rho_{cM}}{N}\frac{1}{M}\sum_{g=1}^{G}\frac{R_{gM}}{\rho_{cM}}\hat{P}'_N\tilde{z}'_{gM}\tilde{z}_{gM}\hat{P}_N \\
=& \big(1+o_p(1)\big)\frac{1}{M}\sum_{g=1}^{G}\frac{R_{gM}}{\rho_{cM}}\big(P'_M+o_p(1)\big)\tilde{z}'_{gM}\tilde{z}_{gM}\big(P_M+o_p(1)\big).
\end{aligned} \qquad \text{(B.53)}
$$

$$
\left\|\hat{\Delta}^Z_{CE,N} - \Delta^Z_{CE,M}\right\| = o_p(1) \qquad \text{(B.54)}
$$

given (B.37).

To show the ordering of the variance-covariance matrices in Theorem 3.2, notice that

$$
\begin{aligned}
&\Delta_{E,M} + \Delta_{EC,M} - \Delta^Z_{CE,M} \\
=& \frac{1}{M}\sum_{g=1}^{G}\mathbb{E}_X\big[\tilde{m}_{gM}(\theta^*_M)\big]\mathbb{E}_X\big[\tilde{m}_{gM}(\theta^*_M)\big]' \\
&- \frac{1}{M}\sum_{g=1}^{G}\mathbb{E}_X\big[\tilde{m}_{gM}(\theta^*_M)\big]\tilde{z}_{gM}\left[\frac{1}{M}\sum_{g=1}^{G}\tilde{z}'_{gM}\tilde{z}_{gM}\right]^{-1}\frac{1}{M}\sum_{g=1}^{G}\tilde{z}'_{gM}\mathbb{E}_X\big[\tilde{m}_{gM}(\theta^*_M)\big]', \qquad \text{(B.55)}
\end{aligned}
$$

which is positive semidefinite.

Hence, the result.