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

Ruonan Xu*

November 7, 2019

[Click here for the latest version]

**Abstract**

When the sample is a relatively large proportion of the population, finite population inference serves as a more appealing alternative to the usual infinite population approach. Nevertheless, the finite population inference methods that are currently available only cover the difference-in-means estimator or independent observations. Consequently, these methods cannot be applied to the many branches of empirical research that use linear or nonlinear models where dependence due to clustering needs to be accounted for in computing the standard errors. In this paper, I establish asymptotic properties of M-estimators under finite populations with clustered data, allowing for unbalanced and unbounded cluster sizes. 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. I show that one should only adjust standard errors for clustering when there is cluster sampling or cluster assignment, or both, for a general class of linear and nonlinear estimators. I also find the finite population cluster-robust asymptotic variance (CRAV) is no larger than the usual infinite population CRAV, in the matrix sense. Consistent with the theoretical implication, the finite population clustered standard errors are smaller than the usual infinite population clustered standard errors by up to 30% in an empirical application.

**JEL Classification**: C18, C10

*Department of Economics, Michigan State University, Marshall-Adams Hall, East Lansing, MI 48824, USA. Email address: xuruonan@msu.edu. 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), this paper studies asymptotic properties of M-estimators where clusters are formed by either the sampling process or the assignment design.[2]

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 typically set up to account for the potential within-group correlation, and the clustered standard errors follow after the model setup.

As an example of the model-based approach, MacKinnon (2019) compares the standard errors based on different assumptions about how the disturbances are clustered. Using

[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).

[2] M-estimators include a broad class of extremum estimators that coincide with a vast majority of estimators used in empirical research [see Wooldridge (2010, Chapter 12)]. Besides linear regression, other leading cases of M-estimation include nonlinear least squares and (quasi-) maximum likelihood.

data from individuals across the 50 states and the District of Columbia from the U.S. Current Population Survey (CPS), MacKinnon (2019) evaluates the return of obtaining a postgraduate degree. The assumptions about clustering have an enormous impact on inference. Confidence intervals are constructed using standard errors clustered at various levels, including non-nested clusters such as states and years. These confidence intervals differ from each other and are almost all much wider than using the heteroskedasticity-robust standard errors that do not account for within-group correlation.

The problem with the model-based approach is the essentially arbitrary nature of the choice of clustering level. In the previous example, one researcher may claim that the unobservables are correlated at the zip code level. Another may claim that correlation exists at the county 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 Abadie et al. (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. For instance, cluster sampling occurs when a random group of hospitals is selected in the first step, followed by a random collection of individual patient data from

the selected hospitals in the second step for cost reasons. Note, 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, Abadie et al. (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 are also generally true 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. 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, clustered standard errors calculated based on finite population inference will be generally smaller than the usually reported clustered standard errors.

Samples that are large relative to the population or coincide with the population motivate the finite population setting in this paper. For instance, relatively large samples could

be drawn from aggregate levels, such as countries, states, or counties. I consider a sequence of finite populations while allowing the sampling probability, defined as the probability of each population unit being drawn into a sample, to depend on the population size and the sample size. As a result, the usual infinite population inference is nested in the framework by allowing the sampling probability to approach zero in the limit.

The current paper contributes to two strands of literature. The first is on finite population inference methods. Abadie, Athey, Imbens, and Wooldridge (2019) propose the finite population inference methods for ordinary least squares estimators incorporating both sampling-based and design-based uncertainties (definition for the two sources of uncertainty can be found in Section 2). Xu (2019) extends Abadie et al. (2019) to M-estimation with both smooth and nonsmooth objective functions. Both studies mentioned above assume independent sampling and independent assignment. Abadie et al. (2017) examines the cluster-robust variance of the difference-in-means estimator caused by cluster sampling or cluster assignment under finite populations but does 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, Schnepel, 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; for instance, in the patient data example, clusters could be

proportional to hospital sizes and hence tend to be unbalanced with the presence of both large and small hospitals. Short panel data are automatically contained in the framework as a special case, where the cluster sizes (the number of time periods) are bounded.

The remaining of the paper is organized as follows. Section 2 illustrates the key concepts of finite population inference using a simple example of the difference-in-means estimator. Section 3 derives the asymptotic distribution under finite populations for M-estimators with smooth objective functions. Generally, the finite population CRAV is non-identifiable because of the missing data problem of the potential outcome framework. Nevertheless, Section 4 proposes two easy ways to bound the finite population CRAV by using control variables to partially predict the variance matrix. The resulting adjusted variance estimator is still conservative in large samples but smaller (in the matrix sense) than the usual cluster-robust variance estimator (CRVE). Section 5 derives the finite population CRAV of functions containing M-estimators with the estimator of the average partial effect (APE) as a direct application. Section 6 compares different standard errors of the APE estimator from pooled probit regressions with clustered data.[3] The simulation results are in line with the predictions from the large-sample theory. Section 7 summarizes an application to Antecol, Bedard, and Stearns (2018), who evaluate the effect of tenure clock stopping policies on tenure rates of female and male faculty members. In this example, the finite population clustered standard errors are smaller than the usual clustered standard errors by up to 30%. Lastly, Section 8 concludes and points out directions for future research.

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

## 2 A Simple Example of the Difference-in-Means Estimator

To introduce the concept of finite population inference, I summarize the example of the variance of the difference-in-means estimator given in Abadie et al. (2019). We start with a finite population of size $M$. There is a single binary treatment variable $X_{iM} \in \{0, 1\}$. Based on the framework of potential outcomes, $X_{iM}$ is a stochastic variable representing different states of the world. Correspondingly, there are two potential outcomes denoted by $\{Y_{iM}(0), Y_{iM}(1)\}$, which are fixed for unit $i$ irrespective of the realized value of $X_{iM}$. For each unit $i$, we could observe only one realization of the potential outcome $Y_{iM} = X_{iM}Y_{iM}(1) + (1 - X_{iM})Y_{iM}(0)$, which leads to the fundamental missing data problem of the potential outcome framework.

The randomness resulting from not observing all states of the world leads to design-based uncertainty.[4] For example, we are interested in examining the return to education by computing the difference in wage rates between all U.S. workers who have college degrees and those who do not. When all U.S. workers are treated as the population of interest, there is no sampling process in this example. The uncertainty of the estimator then stems from not observing the counterfactual wage rates, where workers have different years of schooling from what they actually received. Design-based uncertainty is often neglected in practice [see Freedman (2008a), Freedman (2008b) and Lin (2013) for a few exceptions].

From the population, a sample of fixed size $N$ is randomly drawn with $R_{iM}$ indicating whether unit $i$ is sampled ($R_{iM} = 1$) or not ($R_{iM} = 0$). The subsample sizes $N_1$ and $N_0$ denote the number of units in the sample with $X_{iM} = 1$ or $X_{iM} = 0$ respectively. The randomness arising from (possibly) not observing the entire population leads to sampling-based uncertainty. In the traditional inference methods, sampling-based uncertainty is the

[4]The naming, "design-based," can be traced back to randomized experiments literature, in which Neyman (1923) initially develops the idea of potential outcomes resulted from different assignment of the treatment.

only source of variation that induces the standard error of the coefficient estimators.

$$\tilde{\theta}_N = \frac{1}{N_1}\sum_{i=1}^{M} R_{iM}X_{iM}Y_{iM} - \frac{1}{N_0}\sum_{i=1}^{M} R_{iM}(1-X_{iM})Y_{iM} \tag{1}$$

I focus on the difference-in-means estimator $\tilde{\theta}_N$ in (1) in this section, which is also the coefficient estimator on $X_{iM}$ from the simple regression of $Y_{iM}$ on 1 and $X_{iM}$. Since $\tilde{\theta}_N$ is a function of both $R_{iM}$ and $X_{iM}$, its variance incorporates both sampling-based and design-based uncertainties, which has the following form [see (2.3) in Abadie et al. (2019)]:

$$\mathbb{V}(\tilde{\theta}_N|N_0, N_1) = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\theta^2}{M}, \tag{2}$$

where

$$S_x^2 = \frac{1}{M-1}\sum_{i=1}^{M}\big(Y_{iM}(x) - \frac{1}{M}\sum_{j=1}^{M} Y_{jM}(x)\big)^2,\ x = 1, 0 \tag{3}$$

and

$$S_\theta^2 = \frac{1}{M-1}\sum_{i=1}^{M}\Big(Y_{iM}(1) - Y_{iM}(0) - \frac{1}{M}\sum_{j=1}^{M}\big(Y_{jM}(1) - Y_{jM}(0)\big)\Big)^2. \tag{4}$$

The variance formula in (2) has two implications. First, the last term, $\frac{S_\theta^2}{M}$, is zero either when the population size $M$ tends to infinity or when the treatment effect is constant across units, i.e., $Y_{iM}(1) - Y_{iM}(0) = \tau,\ \forall\ i$. Let us focus on the general case of heterogeneous treatment effects. The difference between finite population inference and the usual infinite population inference results from the last term. Since $\frac{S_\theta^2}{M}$ is positive when $M$ is finite, the finite population variance of the difference-in-means estimator is smaller than the usual infinite population variance. In Section 3, I show that the conservative property of the usual variance of the difference-in-means estimator can be generalized to M-estimators with clustered data.

Second, the extra term, $\frac{S_\theta^2}{M}$, is non-identified in general because we cannot observe both potential outcomes $\{Y_{iM}(0), Y_{iM}(1)\}$ at the same time. The common practice in randomized experiments literature is to ignore the additional term and use the usual overly conservative variance estimator. As shown in Section 4, we can use regression-based approach to estimate the adjusted finite population variance, which is still conservative but smaller than the usual infinite population variance, if there are fixed control variables available.

# 3 Asymptotic Properties of M-estimators

## 3.1 Setup

In this section, I derive the general theory of M-estimators. I use the same setup of M-estimation with smooth objective functions in Xu (2019) but relax the assumptions of independent sampling and independent assignment. Consider a sequence of finite populations indexed by population size $M$. 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 $M_g$ observations, $g = 1, 2, \ldots, G$. I assume for now that sampling and assignments are clustered at the same level if there are both cluster sampling and cluster assignment. 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. 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.

Given the potential outcome framework, there exists a mapping, denoted by the po-

tential 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 unrestricted outcomes, and $y_{igM}(x) = \mathbb{1}[x\theta_{01} + z_{igM}\theta_{02} + e_{igM} > 0]$ for binary outcomes. The potential outcome function, $y_{igM}(x)$, along with the observed attributes $z_{igM}$ and the unobserved attributes $e_{igM}$, are non-stochastic. 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. The subscripts attached to the functions emphasize its dependence on both the fixed attributes and the fixed unobservables.

The distinction between the stochastic assignment variables and the fixed attribute variables is justified by having a thought experiment of certain "policy" or "treatment" assignments. The variables in $X_{iM}$ are the "intervention" variables of interest, and $z_{iM}$ effectively contains control variables. In the example given by Imbens and Rubin (2015) in evaluating the effect of job training on future earnings, the assignment variable is the treatment status, and the "attributes may include age, previous educational achievement, family, and socio-economic status, or pre-training earnings" (p. 15). They elaborate, "the key characteristics of these covariates is that they are a priori known to be unaffected by the treatment assignment. This knowledge often comes from the fact that they are permanent characteristics of units, or that they took on their values prior to the treatment being assigned" (Imbens and Rubin, 2015, p. 16). In observational studies, the assignment variables can be any variables of interest, such as years of education in the study of return to schooling, but are not restricted to treatment variables only. The categorization of assignments and attributes is subjective in practice and is related to the empirical question under research.

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 $\theta_M^*$.

$$\begin{aligned}\theta_M^* &= \arg\min_{\theta} \frac{1}{M} \sum_{g=1}^{G} \sum_{i=1}^{M_g} \mathbb{E}_X \big[ q_{igM}(W_{igM}, \theta) \big] \\ &= \arg\min_{\theta} \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_X \big[ q_{iM}(W_{iM}, \theta) \big]\end{aligned} \tag{5}$$

Note that the expectation in (5) is taken over the distribution of $X$ since $X$ is the source of randomness here. Function $q_{iM}(\cdot, \cdot)$ is the objective function for a single unit. For example, $q_{iM}(W_{iM}, \theta) = (Y_{iM} - X_{iM}\theta_1 - z_{iM}\theta_2)^2$ for linear regression, $q_{iM}(W_{iM}, \theta) = -log\big[f_{iM}(W_{iM}, \theta)\big]$ for (quasi-) maximum likelihood estimation (MLE), in which $f_{iM}(\cdot, \cdot)$ is some density function. I focus on pooled estimation. For MLE, pooled MLE is adopted since it is challenging to specify the joint density in practice. Besides, pooled MLE can still estimate interesting quantities such as the APE.

For each finite population $M$, the sampling process involves two steps.[5] In the first step, a random group of clusters is drawn according to Bernoulli sampling with sampling probability $\rho_{cM}$. Therefore, we have clustered samples whenever $\rho_{cM} < 1$. In the second step, each unit within the selected clusters is again sampled independently according to a Bernoulli trial with probability $\rho_{uM}$. As a result, there is a binary sampling indicator $R_{gM}$, which is equal to one if cluster $g$ is sampled, and another sampling indicator $\tilde{R}_{igM}$, which is equal to one if unit $i$ would be sampled in a one-step sampling process with probability $\rho_{uM}$. The appearance of unit $i$ in the sample is then denoted by a composite sampling indicator $R_{igM} = R_{gM} \cdot \tilde{R}_{igM}$. Occasionally, $R_{igM}$ is suppressed as $R_{iM}$ when the emphasis of clusters is unnecessary. Consequently, we have a random sample size $N = \sum_{g=1}^{G} \sum_{i=1}^{M_g} R_{igM} = \sum_{i=1}^{M} R_{iM}$, where $\mathbb{E}(R_{iM}) = \mathbb{E}(R_{igM}) = \rho_{uM}\rho_{cM}$. Besides cluster sampling, the assignment can also be clustered in the sense that within-cluster covariance of the assignment variables is nonzero.

[5] The sampling process can be generalized to multi-level cluster sampling, though formal results are not provided in the paper.

The estimator of $\theta_M^*$ is denoted by $\hat{\theta}_N$, which solves the minimization problem in the sample.

$$\begin{aligned}\hat{\theta}_N &= \arg\min_{\theta} \frac{1}{N} \sum_{g=1}^{G} \sum_{i=1}^{M_g} R_{igM} q_{igM}(W_{igM}, \theta) \\ &= \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{M} R_{iM} q_{iM}(W_{iM}, \theta)\end{aligned} \tag{6}$$

Panel data could be thought of as a special case of the cluster framework described above without second-layer sampling, i.e., $\rho_{uM} = 1$. In a balanced panel, $M_g$ has equal values across clusters (cross-sectional units). Even with unbalanced panels, the results can be applied as long as the observation of cross-sectional units at certain time periods is random.

Also, note that the objective function is abstract of any correlation between the unobserved cluster heterogeneity and the covariates. In the linear model, the objective function allows for cluster fixed effects, where either $z_{igM}$ includes a set of cluster dummies or the variables are cluster-demeaned. In nonlinear models, Wooldridge (2019) proposes a generalization of correlated random effects model to unbalanced panels, assuming both strict "exogeneity" of the selection indicators and the covariates. The same Chamberlain-Mundlak device can be applied to clustered data where selection indicators go away by nature. The only difference from the standard case would be that now the objective function contains additional sufficient statistics composed of functions of the covariates and cluster sizes.[6]

I make the following assumptions throughout the paper.

**Assumption 1.** *(i) The sampling scheme consists of two stages. In the first stage, clusters*

[6] Alternative approaches to dealing with the correlation between the unobserved group heterogeneity and covariates in the nonlinear model, such as bias-corrected fixed effects estimators (e.g., Fernández-Val and Weidner (2016)), cannot be easily incorporated in the framework of M-estimation; thus they are not considered here.

*are randomly sampled with probability $\rho_{cM}$; in the second stage, units are randomly sampled from the subpopulation consisting of all the sampled clusters with probability $\rho_{uM}$. Hence, the sample size is $N = \sum_{i=1}^{M} R_{iM}$. (ii) The sequence of sampling probabilities $\rho_{cM}$ and $\rho_{uM}$ satisfies $M\rho_{uM}\rho_{cM} \to \infty$, $G\rho_{cM} \to \infty$, and $\rho_{cM} \to \rho_c \in [0,1]$, $\rho_{uM} \to \rho_u \in [0,1]$ as $M \to \infty$.*

**Assumption 2.** *(i) The assignments $\{X_{igM}$, i=1, 2,..., $M_g$, g=1, 2,...,$G\}$ are not (necessarily) identically distributed; (ii) 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* $M \to \infty$.

Assumption 1(i) formalizes the sampling process. Similar to the independent sampling case, the sample size is random, which does not affect the asymptotic distribution of M-estimators as long as the sampling fraction $\frac{N}{M}$ converges to the composite sampling probability $\rho_u\rho_c$ [see the lemma of asymptotic equivalence in Rao (1973, p. 122)]. Assumption 1(ii) assumes both the expected sample size, $\mathbb{E}(N) = M\rho_{uM}\rho_{cM}$, and the expected number of clusters in the sample, $\mathbb{E}(G_N) = G\rho_{cM}$, tend to infinity along with the population size since I adopt the large-$G$ asymptotics throughout the paper. Note that the limiting sampling probabilities $\rho_c$ and $\rho_u$ are allowed to take value zero, which not only nests infinite populations in the framework but also allows for unbounded cluster sizes in the limit, at least in some clusters.

Assumption 2(i) allows for either identically distributed or nonidentically distributed

assignment variables $X_{igM}$. The latter allows the assignments to depend on fixed attributes $z_{igM}$. Assumption 2(ii) allows for clustered assignment, which is another source of within-cluster correlation in addition to within-cluster correlation of the composite sampling indicators. 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 clusters, Assumptions 1-3 contains independent sampling and independent assignment as a special case. Hence, this paper generalizes results in Xu (2019).

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. The cluster sizes in the sample and the overall sample size in Hansen and Lee (2019) are replaced by their population counterparts. Assumption 4 rules out the case where a particular group of clusters dominates the population since each cluster is asymptotically negligible. Suppose all clusters are of the same size; then $M_g = \frac{M}{G}$. As a result, $\max_{g \leq G} \frac{M_g}{M} = \frac{1}{G}$. An implication of Assumption 4 is thus $G \to \infty$. Assumption 5 strengthens Assumption 4 since it is used to show asymptotic normality, which involves higher moments. We can see that

$$\max_{g \leq G} \frac{M_g}{M} \leq \max_{g \leq G} \frac{M_g^2}{M} \to 0. \tag{7}$$

Therefore, once Assumption 5 is imposed, there is no need to impose Assumption 4 since the latter is implied by the former. However, to show consistency, only Assumption 4 is required.

The first part of Assumption 5 deserves the most detailed discussion. The finite popula-

tion counterpart of the original assumption in Hansen and Lee (2019) is for some $2 \leq r < \infty$

$$\frac{\Big(\sum_{g=1}^{G} M_g^r\Big)^{2/r}}{M} \leq C < \infty. \tag{8}$$

The assumption in (8) becomes more restrictive when $r$ approaches 2 with a trade-off between cluster sizes and the order of moments. The intuition can be most easily seen with balanced clusters; the cluster sizes are required to be bounded when $r = 2$ but can grow uniformly at the rate $M_g = M^{\alpha}$ for $0 \leq \alpha \leq \frac{r-2}{2r-2}$ for any $r > 2$. In fact, Assumption 5 rules out clustered data with all clusters unbounded since

$$C \geq \frac{\sum_{g=1}^{G} M_g^2}{M} \geq \frac{\min_{g \leq G} M_g \Big(\sum_{g=1}^{G} M_g\Big)}{M} = \min_{g \leq G} M_g. \tag{9}$$

It is appropriate for Hansen and Lee (2019) to assume (8) since their main contribution is providing fundamental asymptotic distribution theory, such as the weak law of large numbers and the central limit theorem, for clustered data with a large number of independent groups, and potentially unbalanced and unbounded cluster sizes. Also, if we only consider linear models, Assumption 5 can be relaxed to (8), including the cases where all cluster sizes are unbounded. However, for nonlinear models, Assumption 5 is required for the sufficient conditions provided to apply the uniform laws of large numbers for the CRVE. Since I study the asymptotic properties of a general class of M-estimators in this paper, I impose the more restrictive assumption directly. Consequently, the asymptotic theory in this paper is most relevant to clustered data where some cluster sizes are small.

### 3.2 Asymptotic Distribution

The next theorem proves consistency of M-estimators. The limiting population estimand $\theta^*$ in the theorem is defined in the following way:

$$\theta^* = \lim_{M\to\infty} \theta_M^* = \arg\min_{\theta} Q(\theta), \tag{10}$$

where $Q(\theta) \equiv \lim_{M\to\infty} \frac{1}{M}\sum_{i=1}^{M} \mathbb{E}_X\big[q_{iM}(W_{iM},\theta)\big]$. The asymptotic theory relies on the number of clusters $G \to \infty$ as $M \to \infty$. In the meanwhile, heterogeneous and unbounded cluster sizes are allowed.

**Theorem 3.1.** *In addition to Assumptions 1-4, assume that: (i) $Q(\theta)$ is uniquely minimized at $\theta^*$; (ii) $\Theta$ is compact; (iii) $q_{iM}(w,\theta)$ is continuous in $\theta$ 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 $\tilde{\theta}, \theta \in \Theta$, $\big|q_{iM}(W_{iM},\tilde{\theta}) - q_{iM}(W_{iM},\theta)\big| \le b_{1,iM}(W_{iM})h(\|\tilde{\theta}-\theta\|)$. Then $\hat{\theta}_N - \theta^* \xrightarrow{p} \boldsymbol{0}$.*

The major difference between the regularity conditions here and those in the standard case is that the expectation here is taken only over the distribution of $X$ (besides the sampling indicator) while the expectation is taken over the joint distribution of $\{X, z, Y\}$ in the standard case. The introduction of $\theta^*$ 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. Theorem 3.1 implies that $\hat{\theta}_N - \theta_M^* \xrightarrow{p} \mathbf{0}$ since $\theta_M^* - \theta^* \to \mathbf{0}$ holds by definition. The finite population estimand $\theta_M^*$ may not be the true parameters in the potential outcome function since I do not impose correct specification of the objective function. Nevertheless, it generally provides the best approximation to the underlying parameters given the models specified.

I introduce the following notation to help guide the discussion of the asymptotic dis-

tribution of M-estimators:

$$\Delta_{ehw,M}(\theta) = \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_X\big[m_{iM}(W_{iM},\theta)m_{iM}(W_{iM},\theta)'\big], \tag{11}$$

$$\Delta_{E,M} = \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_X\big[m_{iM}(W_{iM},\theta^*_M)\big]\mathbb{E}_X\big[m_{iM}(W_{iM},\theta^*_M)\big]', \tag{12}$$

$$\Delta_{cluster,M}(\theta) = \frac{1}{M}\sum_{g=1}^{G}\sum_{i=1}^{M_g}\sum_{j\neq i}^{M_g}\mathbb{E}_X\big[m_{igM}(W_{igM},\theta)m_{jgM}(W_{jgM},\theta)'\big], \tag{13}$$

$$\Delta_{EC,M} = \frac{1}{M}\sum_{g=1}^{G}\sum_{i=1}^{M_g}\sum_{j\neq i}^{M_g}\mathbb{E}_X\big[m_{igM}(W_{igM},\theta^*_M)]\mathbb{E}_X[m_{jgM}(W_{jgM},\theta^*_M)\big]', \tag{14}$$

$$H_M(\theta) = \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_X\big[\nabla_\theta m_{iM}(W_{iM},\theta)\big], \tag{15}$$

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

$$V_M = H_M(\theta^*_M)^{-1}\big(\Delta_{ehw,M}(\theta^*_M)+\rho_{uM}\Delta_{cluster,M}(\theta^*_M)-\rho_{uM}\rho_{cM}\Delta_{E,M}-\rho_{uM}\rho_{cM}\Delta_{EC,M}\big)H_M(\theta^*_M)^{-1}. \tag{16}$$

Notice that all the matrices are denoted by a subscript $M$ to emphasize their dependence on the population size. Also, the middle part of the sandwich form in (16) is different from the standard case with two additional terms. I denote the conventional infinite population variance matrix and its estimator below:

$$V_{1M} = H_M(\theta^*_M)^{-1}\big(\Delta_{ehw,M}(\theta^*_M) + \rho_{uM}\Delta_{cluster,M}(\theta^*_M)\big)H_M(\theta^*_M)^{-1}, \tag{17}$$

$$\hat{H}_N(\theta) = \frac{1}{N}\sum_{i=1}^{M}R_{iM}\nabla_\theta m_{iM}(W_{iM},\theta), \tag{18}$$

$$\hat{\Delta}_{ehw,N}(\theta) = \frac{1}{N}\sum_{i=1}^{M} R_{iM} \cdot m_{iM}(W_{iM},\theta) m_{iM}(W_{iM},\theta)', \tag{19}$$

$$\hat{\Delta}_{cluster,N}(\theta) = \frac{1}{N}\sum_{g=1}^{G}\sum_{i=1}^{M_g}\sum_{j\neq i}^{M_g} R_{igM}R_{jgM} \cdot m_{igM}(W_{igM},\theta) m_{jgM}(W_{jgM},\theta)', \tag{20}$$

$$\hat{V}_{1N} = \hat{H}_N\big(\hat{\theta}_N\big)^{-1}\big(\hat{\Delta}_{ehw,N}(\hat{\theta}_N) + \hat{\Delta}_{cluster,N}(\hat{\theta}_N)\big)\hat{H}_N(\hat{\theta}_N)^{-1}. \tag{21}$$

The theorem below summarizes the regularity conditions required for asymptotic normality of M-estimators and consistency of the usual CRVE to $V_{1M}$.

**Theorem 3.2.** *Under Assumptions 1, 2, 3, 5, and conditions in Theorem 3.1, 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)* $\theta^* \in int(\Theta)$*; (ii)* $q_{iM}(w,\theta)$ *is twice continuously differentiable on* $int(\Theta)$ *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* $\tilde{\theta},\theta\in\Theta$*,* $\Big\|\nabla_\theta m_{iM}(W_{iM},\tilde{\theta}) - \nabla_\theta m_{iM}(W_{iM},\theta)\Big\| \leq b_{2,iM}(W_{iM})h(\|\tilde{\theta}-\theta\|)$*; (vii)* $H_M(\theta^*_M)$ *is nonsingular; (viii)* $\Delta_{ehw,M}(\theta^*_M)+\rho_{uM}\Delta_{cluster,M}(\theta^*_M)$ *is nonsingular; (ix) 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* $\tilde{\theta},\theta\in\Theta$*,* $\Big\|m_{iM}(W_{iM},\tilde{\theta}) - m_{iM}(W_{iM},\theta)\Big\| \leq b_{3,iM}(W_{iM})h(\|\tilde{\theta}-\theta\|)$*. Then (1)* $V_M^{-1/2}\sqrt{N}(\hat{\theta}_N - \theta^*_M) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, I_k)$*; (2)* $V_{1M}^{-1/2}\hat{V}_{1N}V_{1M}^{-1/2} \xrightarrow{p} I_k$*.*

Because I allow arbitrary within-cluster correlations of the score functions, the convergence rate of M-estimators is unknown. The typical convergence rates with clustered data are $\sqrt{N}$ or $\sqrt{G_N}$, but Hansen and Lee (2019) have given examples showing that the convergence rate can be in between or even slower than these rates. Since the rate of convergence can be calculated as the standard deviation of M-estimators, the composite $V_M^{-1/2}\sqrt{N}$ serves as the implicit rate in Theorem 3.2. As it turns out, it is not necessary

to know the convergence rate since $V_M/N$ gives the correct variance of $\hat{\theta}_N$.

In terms of the variance-covariance matrices, the term $\Delta_{cluster,M}(\theta_M^*)$ is scaled by the sampling probability $\rho_{uM}$ because of the two-stage sampling scheme. Nevertheless, the usual CRVE, $\hat{V}_{1N}$, converges to $V_{1M}$, in which the estimation of $\rho_{uM}$ has been accounted for.

**Corollary 1.** *Clustering is necessary when 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, $\rho_{uM}\big(\Delta_{cluster,M}(\theta_M^*) - \rho_{cM}\Delta_{EC,M}(\theta_M^*)\big)$, is nonzero unless we have both independent sampling and independent assignment. Corollary 1 states 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 give 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 Abadie et al. (2017) but in a much more generalized way. Abadie et al. (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. However, Corollary 1 mainly applies to finite populations, although the generalization to infinite populations would be a natural conjecture.

Let us revisit the example in MacKinnon (2019). Given education is a personal choice, the assignment of education levels is independent across individuals. On the other hand,

since CPS has a two-stage sampling design with the PSUs being either one county or contiguous counties, we need to adjust for clustering at the PSU level whenever the PSU identifier is available. When we have access to the same CPS data, we can consider a different problem, such as the effect of state minimum wage laws on individual wage rates. In this example, there is cluster assignment given the policy is imposed on states. As a result, we should cluster the standard errors at the state level. These two examples serve as the leading cases in practice.

**Corollary 2.** *Since*

$$\Delta_{E,M}+\Delta_{EC,M} = \frac{1}{M}\sum_{g=1}^{G}\left[\sum_{i=1}^{M_g}\mathbb{E}_X\left(m_{igM}(W_{igM},\theta_M^*)\right)\right]\left[\sum_{i=1}^{M_g}\mathbb{E}_X\left(m_{igM}(W_{igM},\theta_M^*)\right)\right]' \quad (22)$$

*is positive semidefinite, the infinite population CRAV of M-estimators is no less than the finite population CRAV, in the matrix sense.*

When clustering is necessary, I rewrite the two additional terms in (22) to compare the asymptotic variance of M-estimators obtained in Theorem 3.2 to the usual infinite population asymptotic variance. Corollary 2 is a generalization of Theorem 2.3 in Xu (2019) to clustered data, which accounts for both sampling-based and design-based uncertainties.

Although the usual CRVE is often overly conservative, there are exceptional cases where it is appropriate to use the usual CRVE for inference. The first scenario is summarized in the corollary below.

**Corollary 3.** *If few clusters are sampled from a large population of clusters, namely, $\rho_{cM} \to 0$, or there is at most one unit sampled from each cluster, i.e., $\rho_{uM} \to 0$, it is appropriate to use the usual CRVE of M-estimators for inference.*

Corollary 3 reaches the same conclusion of Corollary 2(ii) and 2(iii) in Abadie et al. (2017) but in a general framework of M-estimation. When $\rho_{cM}$ is small, which is the case

close to sampling from an infinite number of clusters, we are left with the usual expression of the CRAV. In other words, the CRAV of M-estimators in the infinite population setting is in general conservative unless few clusters are sampled. When $\rho_{uM}$ is close to zero, there is at most one unit sampled from each cluster. The CRAV then reduces to the EHW asymptotic variance, $\Delta_{ehw}(\theta_M^*) - \rho_{uM}\rho_{cM}\Delta_E$. Because $\rho_{uM}$ is close to zero, the composite sampling probability $\rho_{uM}\rho_{cM}$ is also small, which is again close to the case of sampling from an infinite population. As a result, the usual EHW variance estimator is appropriate, and so is the usual CRVE since clustering adjustment does not matter in this case.

Another special case for the usual CRVE to be appropriate 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 coefficient estimator on the assignment variables under the sufficient conditions provided by Abadie et al. (2019), including constant treatment effects, which is required for a correct specification of a linear regression function, and other linearity conditions. The latter holds if the finite population is composed of repetitions of the smallest cluster in terms of the potential outcomes, the fixed attributes, and assignment rules. With this kind of data structure, $\theta_M^*$ 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. The following example demonstrates why the finite population CRAV is generally smaller than the usual CRAV, in the matrix sense. In the case of linear regression, suppose we regress $Y_{iM}$ on $X_{iM}$. The population estimand has the following closed form:

$$\theta_M^* = \Big[\sum_{i=1}^{M} \mathbb{E}_X(X_{iM}' X_{iM})\Big]^{-1} \sum_{i=1}^{M} \mathbb{E}_X(X_{iM}' Y_{iM}). \tag{23}$$

The population residual is defined as $U_{iM} = Y_{iM} - X_{iM}\theta_M^* = y_{iM}(X_{iM}) - X_{iM}\theta_M^*$.

Though $\sum_{i=1}^{M} \mathbb{E}_X(X'_{iM}U_{iM}) = \mathbf{0}$ because of the first order condition in the population minimization problem, $\mathbb{E}_X(X'_{iM}U_{iM})$ is nonzero at least for some unit $i$. The underlying reason is that given $U_{iM}$ contains the fixed potential outcome function, the joint distribution of $\{X_{iM}, U_{iM}\}$ is necessarily nonidentical.[7] Consequently, the cluster summation $\sum_{i=1}^{M_g} \mathbb{E}_X(X'_{igM}U_{igM})$ will only equal to zero by chance, $\forall\ g = 1, \ldots, G$. Hence, when we derive the usual asymptotic variance of the least squares estimator, the positive definite term $\big[\sum_{i=1}^{M_g} \mathbb{E}_X(X'_{igM}U_{igM})\big]\big[\sum_{i=1}^{M_g} \mathbb{E}_X(X'_{igM}U_{igM})\big]'$ remains in the variance matrix, which resulted in a larger (in the matrix sense) asymptotic variance.

# 4 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. However, there is a menu of options valid under different circumstances to at least find a lower bound of the two extra terms.

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 proposed in Theorem 4.2 in Xu (2019). Consider the estimator,

$$\hat{\Delta}^Z_N = \frac{1}{N}\sum_{i=1}^{M} R_{iM}\hat{L}'_N z'_{iM} z_{iM} \hat{L}_N, \tag{24}$$

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 4.1.** *In addition to Assumptions 1-4 and conditions in Theorem 3.2, assume that*

[7] Even if $X_{iM}$ is identically distributed, the usual asymptotic variance is in general too large with the presence of the fixed potential outcome function.

$\frac{1}{M}\sum_{i=1}^{M} z_{iM}' z_{iM}$ *and* $\Delta_M^Z$ *are nonsingular. Then* $\boldsymbol{0} \leq \Delta_M^Z \leq \Delta_{E,M}$*, where* ${\Delta_M^Z}^{-1/2}\hat{\Delta}_N^Z{\Delta_M^Z}^{-1/2} \xrightarrow{p} I_k$ *(all inequalities are in the matrix sense).*

With clustered data, we can include cluster dummies as regressors in the linear projection of $m_{iM}(W_{iM}, \hat{\theta}_N)$ on the fixed attributes. Therefore, the estimation method proposed in Theorem 4.1 is applicable even when there are no other fixed attribute variables available, since the cluster dummies are always known. 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 $\mathbb{E}_X\left[m_{iM}(W_{iM}, \theta_M^*)\right]$ 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 CRVE. 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, we could sum $m_{igM}(W_{igM}, \hat{\theta}_N)$ within each cluster, and linearly project $\sum_{i=1}^{M_g} m_{igM}(W_{igM}, \hat{\theta}_N)$ on the fixed attributes. The number of observations in the linear projection would be the number of clusters in the sample. Hence, cluster dummies should be dropped from the regression. Otherwise, we would run out of degrees of freedom. To reduce the dimensionality of the regressors, the fixed attributes can also be summed within clusters as one way of aggregation. As a result, $\sum_{i=1}^{M_g} \mathbb{E}_X\left[m_{igM}(W_{igM}, \theta_M^*)\right]$ can be partially estimated by its predicted value from the linear projection. Let

$$\tilde{z}_{gM} = \sum_{i=1}^{M_g} z_{igM}, \tag{25}$$

$$\tilde{m}_{gM}(\theta) = \sum_{i=1}^{M_g} m_{igM}(W_{igM}, \theta), \tag{26}$$

and

$$\hat{P}_N = \Big( \sum_{g=1}^{G} R_{gM} \tilde{z}'_{gM} \tilde{z}_{gM} \Big)^{-1} \Big( \sum_{g=1}^{G} R_{gM} \tilde{z}'_{gM} \tilde{m}_{gM}(\hat{\theta}_N)' \Big). \tag{27}$$

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

$$\hat{\Delta}^{Z}_{CE,N} = \frac{1}{N} \sum_{g=1}^{G} R_{gM} \hat{P}'_N \tilde{z}'_{gM} \tilde{z}_{gM} \hat{P}_N. \tag{28}$$

**Theorem 4.2.** *In addition to Assumptions 1, 2, 3, 5, and conditions in Theorem 3.2, assume that (i)* $\rho_{uM} = 1$*; (ii)* $\sum_{g=1}^{G} \tilde{z}'_{gM} \tilde{z}_{gM}$ *is nonsingular; (iii)* $\Delta^{Z}_{CE,M}$ *is nonsingular. Then* $0 \leq \Delta^{Z}_{CE,M} \leq \big(\Delta_{E,M} + \Delta_{EC,M}\big)$*, where* ${\Delta^{Z}_{CE,M}}^{-1/2} \hat{\Delta}^{Z}_{CE,N} {\Delta^{Z}_{CE,M}}^{-1/2} \xrightarrow{p} I_k$ *(all inequalities are in the matrix sense).*

Theorem 4.2 proposes an easy way to partially remove $\Delta_{E,M} + \Delta_{EC,M}$ all at once. The sampling probability $\rho_{uM}\rho_{cM}$ can be estimated by $\frac{N}{M}$, 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,

$$\hat{\Delta}_{ehw,N}(\hat{\theta}_N) + \hat{\Delta}_{cluster,N}(\hat{\theta}_N) - \frac{N}{M} \hat{\Delta}^{Z}_{CE,N} \leq \hat{\Delta}_{ehw,N}(\hat{\theta}_N) + \hat{\Delta}_{cluster,N}(\hat{\theta}_N) \tag{29}$$

(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.

# 5 Asymptotic Distribution of 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 $\theta^*_M$. 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, $\gamma_M^*$ 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. The estimator of $\gamma_M^*$ is denoted by $\hat{\gamma}_N = \frac{1}{N}\sum_{i=1}^{M} R_{iM} f_{iM}(W_{iM},\hat{\theta}_N)$. The conditional variance of $\hat{\gamma}_N$ (conditional on $W$), such as the variance of the partial effect estimator, can be obtained by applying Theorem 3.2 and the delta method directly. The delta method can also be applied to functions $f(\theta_M^*)$ without $W$. When the randomness of $W$ is also taken into account, $\hat{\gamma}_N$ has the asymptotic distribution as shown in Theorem 5.1 (see below).

The notation used in the theorem is defined as follows:

$$F_M(\theta) = \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_X\big[\nabla_\theta f_{iM}(W_{iM},\theta)\big], \tag{30}$$

$$\hat{F}_N(\theta) = \frac{1}{N}\sum_{i=1}^{M} R_{iM}\nabla_\theta f_{iM}(W_{iM},\theta), \tag{31}$$

$$\begin{aligned}\Delta^f_{ehw,M} =& \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_X\Big\{\big[f_{iM}(W_{iM},\theta_M^*) - \gamma_M^* - F_M(\theta_M^*)H_M(\theta_M^*)^{-1}m_{iM}(W_{iM},\theta_M^*)\big]\cdot \\ & \big[f_{iM}(W_{iM},\theta_M^*) - \gamma_M^* - F_M(\theta_M^*)H_M(\theta_M^*)^{-1}m_{iM}(W_{iM},\theta_M^*)\big]'\Big\},\end{aligned} \tag{32}$$

$$\begin{aligned}\Delta^f_{E,M} =& \frac{1}{M}\sum_{i=1}^{M}\Big\{\mathbb{E}_X\big[f_{iM}(W_{iM},\theta_M^*) - \gamma_M^* - F_M(\theta_M^*)H_M(\theta_M^*)^{-1}m_{iM}(W_{iM},\theta_M^*)\big]\cdot \\ & \mathbb{E}_X\big[f_{iM}(W_{iM},\theta_M^*) - \gamma_M^* - F_M(\theta_M^*)H_M(\theta_M^*)^{-1}m_{iM}(W_{iM},\theta_M^*)\big]'\Big\},\end{aligned} \tag{33}$$

$$\begin{aligned}\Delta^f_{cluster,M} =& \frac{1}{M}\sum_{g=1}^{G}\sum_{i=1}^{M_g}\sum_{j\neq i}^{M_g}\mathbb{E}_X\Big\{\big[f_{igM}(W_{igM},\theta_M^*) - \gamma_M^* - F_M(\theta_M^*)H_M(\theta_M^*)^{-1}m_{igM}(W_{igM},\theta_M^*)\big]\cdot \\ & \big[f_{jgM}(W_{jgM},\theta_M^*) - \gamma_M^* - F_M(\theta_M^*)H_M(\theta_M^*)^{-1}m_{jgM}(W_{jgM},\theta_M^*)\big]'\Big\},\end{aligned} \tag{34}$$

$$\Delta^f_{EC,M} = \frac{1}{M}\sum_{g=1}^{G}\sum_{i=1}^{M_g}\sum_{j\neq i}^{M_g}\Big\{\mathbb{E}_X\big[f_{igM}(W_{igM},\theta^*_M) - \gamma^*_M - F_M(\theta^*_M)H_M(\theta^*_M)^{-1}m_{igM}(W_{igM},\theta^*_M)\big]\cdot$$
$$\mathbb{E}_X\Big[f_{jgM}(W_{jgM},\theta^*_M) - \gamma^*_M - F_M(\theta^*_M)H_M(\theta^*_M)^{-1}m_{jgM}(W_{jgM},\theta^*_M)\Big]'\Big\}. \tag{35}$$

The variance matrix of $\hat{\gamma}_N$ is then defined as

$$V_{f,M} = \Delta^f_{ehw,M} + \rho_{uM}\Delta^f_{cluster,M} - \rho_{uM}\rho_{cM}\Delta^f_{E,M} - \rho_{uM}\rho_{cM}\Delta^f_{EC,M}. \tag{36}$$

The usual CRVE in the setting of infinite populations is denoted by $\hat{\Delta}^f_{ehw,N} + \hat{\Delta}^f_{cluster,N}$, where

$$\hat{\Delta}^f_{ehw,N} = \frac{1}{N}\sum_{i=1}^{M} R_{iM}\big[f_{iM}(W_{iM},\hat{\theta}_N) - \hat{\gamma}_N - \hat{F}_N(\hat{\theta}_N)\hat{H}_N(\hat{\theta}_N)^{-1}m_{iM}(W_{iM},\hat{\theta}_N)\big]\cdot$$
$$\big[f_{iM}(W_{iM},\hat{\theta}_N) - \hat{\gamma}_N - \hat{F}_N(\hat{\theta}_N)\hat{H}_N(\hat{\theta}_N)^{-1}m_{iM}(W_{iM},\hat{\theta}_N)\big] \tag{37}$$

and

$$\hat{\Delta}^f_{cluster,N} = \frac{1}{N}\sum_{g=1}^{G}\sum_{i=1}^{M_g}\sum_{j\neq i}^{M_g} R_{igM}R_{jgM}\big[f_{igM}(W_{igM},\hat{\theta}_N) - \hat{\gamma}_N - \hat{F}_N(\hat{\theta}_N)\hat{H}_N(\hat{\theta}_N)^{-1}m_{igM}(W_{igM},\hat{\theta}_N)\big]\cdot$$
$$\big[f_{jgM}(W_{jgM},\hat{\theta}_N) - \hat{\gamma}_N - \hat{F}_N(\hat{\theta}_N)\hat{H}_N(\hat{\theta}_N)^{-1}m_{jgM}(W_{jgM},\hat{\theta}_N)\big]'. \tag{38}$$

**Theorem 5.1.** *Under Assumptions 1, 2, 3, 5, and conditions in Theorem 3.2, suppose that (i) $f_{iM}(w,\theta)$ is continuously differentiable on int$(\Theta)$ 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) $\Delta^f_{ehw,M} +$*

$\rho_{uM}\Delta^f_{cluster,M}$ *is nonsingular; (vii) there is* $h(u) \downarrow 0$ *as* $u \downarrow 0$ *and* $b_5(\cdot) : \mathcal{W} \to R$ *such that* $\sup_{i,M} \mathbb{E}_X\left[b_{5,iM}(W_{iM})^2\right] < \infty$*, and for all* $\tilde{\theta}, \theta \in \Theta$*,* $\left\| f_{iM}(W_{iM}, \tilde{\theta}) - f_{iM}(W_{iM}, \theta) \right\| \leq b_{5,iM}(W_{iM}) h(\|\tilde{\theta} - \theta\|)$*. Then (1)* $V_{f,M}^{-1/2}\sqrt{N}(\hat{\gamma}_N - \gamma^*_M) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, I_q)$*;*
*(2)* $(\Delta^f_{ehw,M} + \rho_{uM}\Delta^f_{cluster,M})^{-1/2}\big(\hat{\Delta}^f_{ehw,N} + \hat{\Delta}^f_{cluster,N}\big)(\Delta^f_{ehw,M} + \rho_{uM}\Delta^f_{cluster,M})^{-1/2} \xrightarrow{p} I_q$.

Theorem 5.1 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 4 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

$$f_{igM}(W_{igM}, \hat{\theta}_N) - \hat{\gamma}_N - \hat{F}_N(\hat{\theta}_N)\hat{H}_N(\hat{\theta}_N)^{-1} m_{igM}(W_{igM}, \hat{\theta}_N) \tag{39}$$

or the cluster sum of it rather than $m_{igM}(W_{igM}, \hat{\theta}_N)$ alone.

# 6 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

$$y_{igM}(x) = \mathbb{1}[x + 2z_{igM} \cdot x + c_{gM} + e_{igM} > 0]. \tag{40}$$

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 unobservables $e_{igM}$ 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 $\sigma^2$. In the second stage, $X_{igM}$ in cluster $g$ is assigned to 1 independently, with cluster specific probability $p_{gM}$. If $\sigma^2 > 0$, we have correlated assignment within each cluster but independent assignment 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. Since I want to show the finite-sample performance of the regression-based adjusted finite population CRVE proposed in Theorem 4.2, there is no second-stage 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.

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

| | | No Cluster Assignment | | | With Cluster Assignment | | |
|---|---|---|---|---|---|---|---|
| | | (1) | (2) | (3) | (4) | (5) | (6) |
| | | $\rho_c = 0.1$ | $\rho_c = 0.5$ | $\rho_c = 1$ | $\rho_c = 0.1$ | $\rho_c = 0.5$ | $\rho_c = 1$ |
| $G\rho_c = 50$ | $APE^*_M$ | 0.1007 | 0.1350 | 0.1300 | 0.1007 | 0.1350 | 0.1300 |
| | $\widehat{\widehat{APE}}$ | 0.1091 | 0.1413 | 0.1376 | 0.1090 | 0.1431 | 0.1388 |
| | $std$ | 0.0736 | 0.0618 | 0.0395 | 0.0809 | 0.0724 | 0.0572 |
| | $se_{limit}$ | 0.0745 | 0.0623 | 0.0390 | 0.0817 | 0.0717 | 0.0553 |
| | $\bar{se}_{cluster}$ | 0.0761 | 0.0783 | 0.0785 | 0.0829 | 0.0855 | 0.0874 |
| | $cov_{cluster}$ | (0.955) | (0.985) | (1.000) | (0.953) | (0.977) | (0.996) |
| | $\bar{se}_{adj}$ | 0.0740 | 0.0659 | 0.0507 | 0.0810 | 0.0744 | 0.0635 |
| | $cov_{adj}$ | (0.949) | (0.964) | (0.988) | (0.948) | (0.956) | (0.969) |
| | $\bar{se}_{ehw,adj}$ | 0.0513 | 0.0491 | 0.0468 | 0.0516 | 0.0495 | 0.0471 |
| $G\rho_c = 100$ | $APE^*_M$ | 0.1182 | 0.1075 | 0.1350 | 0.1182 | 0.1075 | 0.1350 |
| | $\widehat{\widehat{APE}}$ | 0.1263 | 0.1174 | 0.1418 | 0.1261 | 0.1165 | 0.1422 |
| | $std$ | 0.0523 | 0.0439 | 0.0270 | 0.0581 | 0.0501 | 0.0373 |
| | $se_{limit}$ | 0.0533 | 0.0442 | 0.0272 | 0.0591 | 0.0506 | 0.0371 |
| | $\bar{se}_{cluster}$ | 0.0550 | 0.0558 | 0.0558 | 0.0604 | 0.0608 | 0.0612 |
| | $cov_{cluster}$ | (0.961) | (0.985) | (1.000) | (0.954) | (0.982) | (0.999) |
| | $\bar{se}_{adj}$ | 0.0535 | 0.0471 | 0.0361 | 0.0590 | 0.0530 | 0.0440 |
| | $cov_{adj}$ | (0.955) | (0.963) | (0.989) | (0.949) | (0.962) | (0.979) |
| | $\bar{se}_{ehw,adj}$ | 0.0364 | 0.0349 | 0.0327 | 0.0365 | 0.0350 | 0.0329 |

[1] $G$ is the number of clusters in the population; $\rho_c$ is the sampling probability of clusters; thus, $\mathbb{E}(G_N) = G\rho_c$ 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; $\widehat{\widehat{APE}}$ stands for the average of the APE estimates across replications; $std$ stands for the Monte Carlo standard deviation; $se_{limit}$ stands for the analytical cluster-robust standard error with finite populations; $\bar{se}_{cluster}$ 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; $\bar{se}_{adj}$ 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; $\bar{se}_{ehw,adj}$ 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.

Estimates from the pooled probit regression of $Y_{igM}$ on 1, $X_{igM}$, and $z_{igM}$ are collected in Table 1. To report the analytical standard errors, $\theta_M^*$ is computed by minimizing the finite population objective function as in (5) 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 two rows in Table 1 report the APE of $X$ in the population obtained from the potential outcome function and the average of the APE estimates across the replications. The population APEs vary across columns because the population sizes are different in the design of each column. 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.[8] With quasi-MLE, the hope is to get the best approximation to the population APE given the model specified.

The third and fourth rows report the Monte Carlo standard deviation of the APE estimator and its analytical cluster-robust standard error with finite populations using the formula in Theorem 5.1.[9] In all designs, the analytical cluster-robust standard errors are

[8] The model is misspecified because the interaction term in the potential outcome function is not captured by the probit regression and $\{e_{igM},\ i = 1, \ldots, M_g,\ g = 1, \ldots, G\}$ has a discrete rather than standard normal distribution.

[9] It is equivalent to apply the delta method in this case since the individual partial effect does not contain stochastic assignment variables.

pretty close to the Monte Carlo standard deviations, confirming the correctness of the analytical formula at least in this population generating process.

The next two rows report the average of the usual cluster-robust standard error and the corresponding coverage rate of the 95% confidence interval. 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 larger than 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 coverage rates here, especially the ones in the top panel, should be interpreted with caution though due to the cluster heterogeneity and the relatively small cluster number in the sample. In the setting of infinite populations, Bester, Conley, and Hansen (2011), by adopting the fixed-G asymptotics, show that the cluster-robust t statistic follows a t distribution with $G_N - 1$ degrees of freedom under homogeneity of both the design matrices and the variance of the within-group scores. Simulation results in MacKinnon and Webb (2017) with wildly different cluster sizes show that using critical values from $t(G_N - 1)$ at least outperforms critical values from the standard normal distribution. As a result, the $97.5^{th}$ percentile of $t(G\rho_c - 1)$ is used as the critical value in constructing the confidence intervals. The appropriateness of using critical values as such is a conjecture without proof in the context of finite populations. 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.

The seventh and eighth 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).[10] 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 error components within clusters.

It is interesting to find that the adjusted finite population cluster-robust standard error can undo the unnecessary clustering to some extent, which supplements another reason to use the adjusted finite population standard errors whenever appropriate. This phenomenon is implied by the theory since if the extra terms in the finite population CRAV can be identified, then clustering makes no difference when it is unnecessary. However, it is usually undetermined whether the adjusted finite population EHW standard error or the adjusted finite population cluster-robust standard error is closer to the standard deviation. The limited evidence in column (3) shows that the former performs better, though.

The different standard errors of the coefficient estimator on the assignment variable perform in the same way as those of the APE estimator, as shown in Table 3 in Appendix B. When the interaction term is removed from the potential outcome function in (40), the probit specification can be considered as approximately correct as long as the population size is large enough for $\{e_{igM},\ i = 1, \ldots, M_g,\ g = 1, \ldots, G\}$ to approach the standard

[10] The adjusted EHW standard errors are obtained using the regression-based approach in Theorem 4.1 without the clustering term.

normal distribution; in the simulation, we do observe that the average APE estimates are pretty close to the population APEs. The simulation results in terms of the comparison of the standard errors show the same pattern (not reported here). The only major difference is that since now the attribute $z_{igM}$ is no longer correlated with the score function, the adjusted finite population clustered standard errors are equivalently conservative as the usual clustered standard errors.

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 a large number of clusters in the population. When there are fixed attributes available, they can be used to estimate a lower 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.

# 7 Application

The adjusted finite population CRVE proposed in Theorem 4.2 is applied to Antecol et al. (2018), who study the effect of tenure clock stopping policies on tenure rates among assistant professors. The unique dataset collected by the authors 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. The standard errors in the original paper are clustered at the policy university level, which is the correct level to cluster the standard errors implied by Corollary 1.

Since the dependent variable is a binary response, I analyze the linear probability model given in the original paper along with an additional probit model given in (41) below, which

adopts the same notation from the original paper.

$$
\begin{aligned}
&P(Y_{ugit} = 1|GN_{ut}, F_{ugit}, E_{ut}, FO_{ut}, X_{ugit}, Z_{ut}, \rho_{gt}, \psi_{ug}) = \\
&\Phi(\beta_0 + \beta_1 GN_{ut} + \beta_2 GN_{ut} \times F_{ugit} + \beta_3 GN_{ut} \times E_{ut} + \beta_4 GN_{ut} \times E_{ut} \times F_{ugit} \\
&+\beta_5 FO_{ut} + \beta_6 FO_{ut} \times F_{ugit} + \beta_7 FO_{ut} \times E_{ut} + \beta_8 FO_{ut} \times E_{ut} \times F_{ugit} \\
&+X_{ugit}\xi + Z_{ut}\eta + \rho_{gt} + \psi_{ug})
\end{aligned}
\tag{41}
$$

The dependent variable $Y$ is an indicator of obtaining tenure at the policy university. 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 $\rho$ captures gender-specific time trend and $\psi$ represents gender-specific university heterogeneity. The subscripts, $u$, $g$, $i$, $t$, are indicators for university, gender, individual, and the year the job started respectively.

The authors of the original paper 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 for 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 time-varying university characteristics, and their interactions with cluster sizes are included as additional controls.

[11] Please refer to the original paper for the details of the variables included as controls.

Given the probit model above is a nonlinear "difference-in-differences" model, the common 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 induced by the incremental effect of the coefficient on the treatment variables. For instance, the total effect of the female-only policy for men hired in years zero through three after policy adoption is defined in (42) below.

$$\Phi\big(\beta_{0c}+\beta_{5c}+\beta_{7c}+\xi_c X_{ugit}+\eta_c Z_{ut}+\rho_{mtc}+\bar{X}_{ug}\lambda_{1c}+\bar{Z}_u\lambda_{2c}+\lambda_{3c}M_g+(M_g\times\bar{X}_{ug})\lambda_{4c}+(M_g\times\bar{Z}_u)\lambda_{5c}\big)$$
$$-\Phi\big(\beta_{0c}+\xi_c X_{ugit}+\eta_c Z_{ut}+\rho_{mtc}+\bar{X}_{ug}\lambda_{1c}+\bar{Z}_u\lambda_{2c}+\lambda_{3c}M_g+(M_g\times\bar{X}_{ug})\lambda_{4c}+(M_g\times\bar{Z}_u)\lambda_{5c}\big) \tag{42}$$

Notice that the potential outcomes for males with or without treatment is obtained by imposing the male time trend, denoted by $\rho_{mtc}$, to both males and females. The average treatment effect on the treated is then calculated as the average of the treatment effect for those actually treated by the specific policy. Assume that $\psi$ conditional on the sufficient statistics (the additional controls included) follows a normal distribution. The subscript $c$ denotes the scaled parameters. Even though the parameters can only be identified up to scale via pooled probit, we can still obtain the APEs. The total effects of other treatment groups are defined similarly.

Table 2: The Effect of Clock Stopping Policies on the Probability of Tenure at the Policy University

| | 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 | 0.0574 | -0.0067 | 0.0610 | 0.0422 |
| Women FOCS | 0.1723 | 0.1405 | 0.1102 | 0.1491 | 0.1826 | 0.1241 |
| Men GNCS | 0.0511 | 0.0787 | 0.0690 | 0.0426 | 0.0693 | 0.0553 |
| Women GNCS | -0.0166 | 0.1071 | 0.0894 | 0.0256 | 0.1213 | 0.0857 |
| Panel B. Policy effects years 4+ | | | | | | |
| Men FOCS | 0.0023 | 0.0747 | 0.0639 | -0.0054 | 0.0638 | 0.0458 |
| Women FOCS | 0.0493 | 0.1015 | 0.0743 | 0.0433 | 0.0959 | 0.0643 |
| Men GNCS | 0.1757 | 0.0826 | 0.0650 | 0.1468 | 0.0733 | 0.0549 |
| Women GNCS | -0.1945 | 0.1057 | 0.0859 | -0.2207 | 0.0924 | 0.0660 |

[1] Standard errors are clustered at the policy 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 (coefficient estimators in the case of the linear probability model); columns (3) and (6) report the adjusted finite population cluster-robust standard errors of the APE estimators.

[3] Please refer to the original paper for detailed control variables.

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 linear probability model. Columns (1) and (2) report the total effects and the standard errors, as shown in column (1) in the original table labelled 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 (29) 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 individual and university characteristics as the fixed attributes in the linear projection, resulting in a linear regression with 49 observations and 12 independent variables. Compared with the usual cluster-robust standard errors, the finite population cluster-robust standard errors shrink by about 12% to 27% across the eight treatment groups. In terms of the statistical significance, the effect of gender-neutral policy for men hired three or more years after the policy adoption is significant at the 1% rather than the 5% level based on the adjusted finite population cluster-robust standard error. Also, the effect of gender-neutral policy for women hired in later years is now significant at the 5% instead of the 10% level. The same results hold 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 4.2 and the delta method. The reduction from the usual clustered standard error to the finite population clustered standard error is even more substantial in the nonlinear model, varying from 20% to 33%, partly because the fixed sufficient statistics are also included as regressors in the linear projection. Based on the critical values from $t(48)$, the effect of gender-neutral policy for men hired in later years is significant at the 5% level rather than the 10% level when the finite population clustered

standard error is adopted. In addition, the significance level of the effect of gender-neutral policy for women hired in later years changes from the 5% to the 1% level under finite population inference.

Table 4 in Appendix B 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 nearly at the 5% 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 25% where the score functions are regressed on the cluster sums of the cluster size dummies and the attribute variables.

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 linear probability model and the infinite population clustered standard error.

# 8 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. For wildly unbalanced clusters or a small number of clusters, the wild cluster bootstrap [12] 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.

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

## 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. (2019), Sampling-based vs. design-based uncertainty in regression analysis. *Econometrica* forthcoming.

Antecol, H., Bedard, K., and Stearns, J. (2018), Equal but inequitable: Who benefits from gender-neutral tenure clock stopping policies? *American Economic Review* 108(9), 2420–2441.

Arellano, M. (1987), Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics* 49(4), 431–434.

Bell, R.M. and McCaffrey, D.F. (2002), Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology* 28(2), 169–181.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004), How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics* 119(1), 249–275.

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

Bhattacharya, D. (2005), Asymptotic inference from multi-stage samples. *Journal of Econometrics* 126(1), 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(3), 414–427.

Cameron, A.C. and Miller, D.L. (2015), A practitioner's guide to cluster-robust inference. *Journal of Human Resources* 50(2), 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(4), 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(2), 393–412.

Fernández-Val, I. and Weidner, M. (2016), Individual and time effects in nonlinear panel models with large N, T. *Journal of Econometrics* 192(1), 291–312.

Freedman, D.A. (2008a), On regression adjustments in experiments with several treatments. *Annals of Applied Statistics* 2(1), 176–196.

Freedman, D.A. (2008b), On regression adjustments to experimental data. *Advances in Applied Mathematics* 40(2), 180–193.

Hansen, B.E. and Lee, S. (2019), Asymptotic theory for clustered samples. *Journal of Econometrics* 210(2), 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(2), 597–620.

Imbens, G.W. and Rubin, D.B. (2015), *Causal inference in statistics, social, and biomedical science*. Cambridge University Press.

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

Kloek, T. (1981), OLS estimation in a model where a microvariable is explained by aggregates and contemporaneous disturbances are equicorrelated. *Econometrica* 49(1), 205–207.

Lechner, M. (2011), The estimation of causal effects by difference-in-difference methods. *Foundations and Trends in Econometrics* 4(3), 165–224.

Liang, K. and Zeger, S.L. (1986), Longitudinal data analysis using generalized linear models. *Biometrika* 73(1), 13–22.

Lin, W. (2013), Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *Annals of Applied Statistics* 7(1), 295–318.

MacKinnon, J.G. (2019), How cluster-robust inference is changing applied econometrics. *Canadian Journal of Economics* 52(3), 851–881.

MacKinnon, J.G. and Webb, M.D. (2017), Wild bootstrap inference for wildly different cluster sizes. *Journal of Applied Econometrics* 32(2), 233–254.

Moulton, B.R. (1986), Random group effects and the precision of regression estimates. *Journal of Econometrics* 32(3), 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(2), 334–338.

Newey, W.K. (1991), Uniform convergence in probability and stochastic equicontinuity. *Econometrica* 59(4), 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.

Neyman, J. (1923), On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Annals of Agricultural Sciences* 10, 1–51.

Puhani, P.A. (2012), The treatment effect, the cross difference, and the interaction term in nonlinear "difference-in-differences" models. *Economics Letters* 115(1), 85–87.

Rao, C.R. (1973), *Linear statistical inference and its applications (2nd ed.)*. Wiley.

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(380), 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(2), 133–138.

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

Wooldridge, J.M. (2019), Correlated random effects models with unbalanced panels. *Journal of Econometrics* 211(1), 137–150.

Xu, R. (2019), Potential outcomes and finite population inference for M-estimators, working paper.

# A Proof

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

**Lemma A.1.** *Under Assumption 4, suppose* $N = \sum_{g=1}^{G}\sum_{i=1}^{M_g} R_{igM}$, *where* $R_{igM} = R_{gM}\tilde{R}_{igM}$, *and* $R_{gM}$ *and* $\tilde{R}_{igM}$ *follow Bernoulli distribution with probability* $\rho_{uM} > 0$ *and* $\rho_{cM} > 0$ *respectively. Then* $\frac{N}{M\rho_{uM}\rho_{cM}} \xrightarrow{p} 1$.

*Proof.* Since

$$\mathbb{E}\left(\left|\frac{R_{igM}}{\rho_{uM}\rho_{cM}}\right|\right) = 1 < \infty, \tag{A.1}$$

$$\frac{N}{M\rho_{uM}\rho_{cM}} = \frac{\sum_{g=1}^{G}\sum_{i=1}^{M_g} R_{igM}}{M\rho_{uM}\rho_{cM}} \xrightarrow{p} 1 \tag{A.2}$$

follows from Theorem 1 in Hansen and Lee (2019) under Assumption 4. ■

**Lemma A.2.** *Suppose there exists* $h(u) \downarrow 0$ *as* $u \downarrow 0$ *and* $b(\cdot) : \mathcal{W} \to R$ *such that* $\sup_{i,M} \mathbb{E}_X\left[b_{iM}(W_{iM})\right] < \infty$, *and for all* $\tilde{\theta}, \theta \in \Theta$, $\left\|a_{iM}(W_{iM},\tilde{\theta}) - a_{iM}(W_{iM},\theta)\right\| \leq b_{iM}(W_{iM})h(\|\tilde{\theta}-\theta\|)$. *Then* $B_N \equiv \frac{1}{N}\sum_{i=1}^{M} R_{iM}\cdot b_{iM}(W_{iM}) = O_p(1)$ *and* $\left\|A_N(\tilde{\theta}) - A_N(\theta)\right\| \leq B_N h(\|\tilde{\theta}-\theta\|)$, *where* $A_N(\theta) \equiv \frac{1}{N}\sum_{i=1}^{M} R_{iM}\cdot a_{iM}(W_{iM},\theta)$.

*Proof.* The proof is modifications of the proof of Corollary 3.1 in Newey (1991).

$$\begin{aligned}
&\left\|A_N(\tilde{\theta}) - A_N(\theta)\right\| \\
=&\left\|\frac{1}{N}\sum_{i=1}^{M} R_{iM}\left[a_{iM}(W_{iM},\tilde{\theta}) - a_{iM}(W_{iM},\theta)\right]\right\| \\
\leq&\frac{1}{N}\sum_{i=1}^{M} R_{iM}\left\|a_{iM}(W_{iM},\tilde{\theta}) - a_{iM}(W_{iM},\theta)\right\| \\
\leq&\frac{1}{N}\sum_{i=1}^{M} R_{iM}\cdot b_{iM}(W_{iM})h(\|\tilde{\theta}-\theta\|)
\end{aligned}$$

$$= B_N h(\|\tilde{\theta} - \theta\|) \tag{A.3}$$

$$B_N \equiv \frac{1}{N}\sum_{i=1}^{M} R_{iM} \cdot b_{iM}(W_{iM}) = \frac{M\rho_{uM}\rho_{cM}}{N}\frac{1}{M}\sum_{i=1}^{M}\frac{R_{iM}}{\rho_{uM}\rho_{cM}} b_{iM}(W_{iM}) \tag{A.4}$$

Because of Lemma A.1 and the continuous mapping theorem, $\frac{M\rho_{uM}\rho_{cM}}{N} \xrightarrow{p} 1$. As a result, it is sufficient to prove $\frac{1}{M}\sum_{i=1}^{M}\frac{R_{iM}}{\rho_{uM}\rho_{cM}} b_{iM}(W_{iM}) = O_p(1)$. For all $\epsilon > 0$, let $b_\epsilon = C/\epsilon$,

$$\begin{aligned}
&P\left(\left|\frac{1}{M}\sum_{i=1}^{M}\frac{R_{iM}}{\rho_{uM}\rho_{cM}} b_{iM}(W_{iM})\right| \geq b_\epsilon\right) \\
\leq &\mathbb{E}_X\left(\left|\frac{1}{M}\sum_{i=1}^{M}\frac{R_{iM}}{\rho_{uM}\rho_{cM}} b_{iM}(W_{iM})\right|\right)/b_\epsilon \\
\leq &\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}\Big(\frac{R_{iM}}{\rho_{uM}\rho_{cM}}\Big)\mathbb{E}_X\Big[\big|b_{iM}(W_{iM})\big|\Big]/b_\epsilon \\
= &\sup_{i,M}\mathbb{E}_X\Big[\big|b_{iM}(W_{iM})\big|\Big]/b_\epsilon < C/b_\epsilon = \epsilon.
\end{aligned} \tag{A.5}$$

Hence, $B_N = O_p(1)$. ■

**Proof of Theorem 3.1**

*Proof.* To prove Theorem 3.1, 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 Theorem 3.1. Their condition (iii) holds under conditions (iii) and (iv) in Theorem 3.1 by the dominated convergence theorem (DCT) and Jensen's inequality. To show their condition (iv) holds under the conditions in Theorem 3.1, first note that

$$\frac{1}{N}\sum_{i=1}^{M} R_{iM} q_{iM}(W_{iM}, \theta) = \frac{M\rho_{uM}\rho_{cM}}{N}\frac{1}{M}\sum_{i=1}^{M}\frac{R_{iM}}{\rho_{uM}\rho_{cM}} q_{iM}(W_{iM}, \theta). \tag{A.6}$$

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

$$\left\| \frac{1}{M}\sum_{i=1}^{M} \frac{R_{iM}}{\rho_{uM}\rho_{cM}} q_{iM}(W_{iM},\theta) - \frac{1}{M}\sum_{i=1}^{M} \mathbb{E}_X\big[q_{iM}(W_{iM},\theta)\big] \right\| \xrightarrow{p} 0. \tag{A.7}$$

Condition (iii) in Theorem 3.1 implies $\forall\ \theta \in \Theta$

$$\sup_{i,M} \mathbb{E}_X\left[ \left| \frac{R_{iM}}{\rho_{uM}\rho_{cM}} q_{iM}(W_{iM},\theta) \right|^r \right] \leq \frac{1}{(\rho_{uM}\rho_{cM})^{r-1}} \sup_{i,M} \mathbb{E}_X\Big[ \sup_{\theta\in\Theta} |q_{iM}(W_{iM},\theta)|^r \Big] < \infty \tag{A.8}$$

for some $r > 1$, which further implies

$$\lim_{C\to\infty} \sup_{i,M} \left\{ \mathbb{E}\left[ \left| \frac{R_{iM}}{\rho_{uM}\rho_{cM}} q_{iM}(W_{iM},\theta) \right| \cdot \mathbb{1}\left( \left| \frac{R_{iM}}{\rho_{uM}\rho_{cM}} q_{iM}(W_{iM},\theta) \right| > C \right) \right] \right\} = 0. \tag{A.9}$$

(A.7) 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 and Corollary 2.2 in Newey (1991) under condition (v) in Theorem 3.1.

■

**Proof of Theorem 3.2**

*Proof.* The proof is modifications of the proof of Theorem 11 in Hansen and Lee (2019) to M-estimators with smooth objective functions under finite populations.

I start by showing that

$$\sum_{i=1}^{M} \mathbb{E}_X\big[m_{iM}(W_{iM},\theta_M^*)\big] = \mathbf{0}, \tag{A.10}$$

which holds by Lemma 3.6 in Newey and McFadden (1994) and Jensen's inequality under conditions (ii) and (iii) in Theorem 3.2.

By the element-by-element mean value expansion around $\theta_M^*$,

$$\begin{aligned} o_p(N^{-1/2}) &= V_M^{-1/2}\frac{1}{N}\sum_{i=1}^{M} R_{iM}\cdot m_{iM}(W_{iM},\hat{\theta}_N) \\ =&V_M^{-1/2}\frac{1}{N}\sum_{i=1}^{M} R_{iM}\cdot m_{iM}(W_{iM},\theta_M^*) + V_M^{-1/2}\frac{1}{N}\sum_{i=1}^{M} R_{iM}\nabla_\theta m_{iM}(W_{iM},\check{\theta})(\hat{\theta}_N-\theta_M^*), \end{aligned} \tag{A.11}$$

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

I first show

$$\hat{H}_N(\check{\theta}) = H_M(\theta_M^*)\big(I_k + o_p(1)\big). \tag{A.12}$$

Since we can write

$$\hat{H}_N(\check{\theta}) = H_M(\theta_M^*)\Big[I_k + H_M(\theta_M^*)^{-1}\big(\hat{H}_N(\check{\theta}) - H_M(\theta_M^*)\big)\Big], \tag{A.13}$$

it suffices to show

$$\Big\|H_M(\theta_M^*)^{-1}\big(\hat{H}_N(\check{\theta}) - H_M(\theta_M^*)\big)\Big\| \xrightarrow{p} 0. \tag{A.14}$$

We can write

$$\begin{aligned} \hat{H}_N(\theta) =&\frac{M\rho_{uM}\rho_{cM}}{N}\frac{1}{M}\sum_{i=1}^{M}\frac{R_{iM}}{\rho_{uM}\rho_{cM}}\nabla_\theta m_{iM}(W_{iM},\theta) \\ =&\big(1+o_p(1)\big)\frac{1}{M}\sum_{i=1}^{M}\frac{R_{iM}}{\rho_{uM}\rho_{cM}}\nabla_\theta m_{iM}(W_{iM},\theta). \end{aligned} \tag{A.15}$$

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

$$\begin{aligned} &\sup_{i,M}\mathbb{E}_X\left[\left\|\frac{R_{iM}}{\rho_{uM}\rho_{cM}}\nabla_\theta m_{iM}(W_{iM},\theta)\right\|^r\right] \\ \le&\frac{1}{(\rho_{uM}\rho_{cM})^{r-1}}\sup_{i,M}\mathbb{E}_X\Big[\sup_{\theta\in\Theta}\|\nabla_\theta m_{iM}(W_{iM},\theta)\|^r\Big] < \infty \end{aligned} \tag{A.16}$$

for some $r > 1$,

$$\left\| \frac{1}{M} \sum_{i=1}^{M} \frac{R_{iM}}{\rho_{uM}\rho_{cM}} \nabla_\theta m_{iM}(W_{iM}, \theta) - H_M(\theta) \right\| \xrightarrow{p} 0 \tag{A.17}$$

by Theorem 1 in Hansen and Lee (2019) under Assumption 4 (implied by Assumption 5) and condition (v) in Theorem 3.2. Note that $H_M(\theta)$ is continuous in $\theta$ by the DCT and Jensen's inequality under conditions (ii) and (v) in Theorem 3.2. By Corollary 2.2 in Newey (1991) and Lemma A.2,

$$\begin{aligned} &\left\| H_M(\theta_M^*)^{-1} \big( \hat{H}_N(\check{\theta}) - H_M(\theta_M^*) \big) \right\| \\ \leq & C \left( \sup_{\theta \in \Theta} \left\| \hat{H}_N(\theta) - H_M(\theta) \right\| + \left\| H_M(\check{\theta}) - H_M(\theta_M^*) \right\| \right) \xrightarrow{p} 0 \end{aligned} \tag{A.18}$$

under conditions (vi) and (vii) in Theorem 3.2.

(A.12) implies

$$\hat{H}_N(\check{\theta})^{-1} = H_M(\theta_M^*)^{-1}(I_k + o_p(1)). \tag{A.19}$$

Using (A.19), (A.11) can be written as

$$\begin{aligned} V_M^{-1/2}\sqrt{N}(\hat{\theta}_N - \theta_M^*) = & - V_M^{-1/2} H_M(\theta_M^*)^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{M} R_{iM} \cdot m_{iM}(W_{iM}, \theta_M^*) \\ & - V_M^{-1/2} H_M(\theta_M^*)^{-1} o_p(1) \frac{1}{\sqrt{N}} \sum_{i=1}^{M} R_{iM} \cdot m_{iM}(W_{iM}, \theta_M^*) + o_p(1). \end{aligned} \tag{A.20}$$

We can write

$$\begin{aligned} \frac{1}{\sqrt{N}} \sum_{i=1}^{M} R_{iM} \cdot m_{iM}(W_{iM}, \theta_M^*) &= \sqrt{\frac{M\rho_{uM}\rho_{cM}}{N}} \frac{1}{\sqrt{M}} \sum_{i=1}^{M} \frac{R_{iM}}{\sqrt{\rho_{uM}\rho_{cM}}} m_{iM}(W_{iM}, \theta_M^*) \\ &= \big(1 + o_p(1)\big) \frac{1}{\sqrt{M}} \sum_{i=1}^{M} \frac{R_{iM}}{\sqrt{\rho_{uM}\rho_{cM}}} m_{iM}(W_{iM}, \theta_M^*). \end{aligned} \tag{A.21}$$

Plug (A.21) into (A.20), we have

$$
\begin{aligned}
&V_M^{-1/2}\sqrt{N}(\hat{\theta}_N - \theta_M^*) \\
=&-V_M^{-1/2}H_M(\theta_M^*)^{-1}\frac{1}{\sqrt{M}}\sum_{i=1}^{M}\frac{R_{iM}}{\sqrt{\rho_{uM}\rho_{cM}}}m_{iM}(W_{iM},\theta_M^*) \\
&-V_M^{-1/2}H_M(\theta_M^*)^{-1}\frac{1}{\sqrt{M}}\sum_{i=1}^{M}\frac{R_{iM}}{\sqrt{\rho_{uM}\rho_{cM}}}m_{iM}(W_{iM},\theta_M^*)\cdot o_p(1) + o_p(1). \qquad \text{(A.22)}
\end{aligned}
$$

Since

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

$$= \Delta_{ehw,M}(\theta_M^*) - \rho_{uM}\rho_{cM}\Delta_{E,M} + \rho_{uM}\Delta_{cluster,M}(\theta_M^*) - \rho_{uM}\rho_{cM}\Delta_{EC,M}, \tag{A.23}$$

we have

$$\mathbb{V}_X\left(V_M^{-1/2}H_M(\theta_M^*)^{-1}\frac{1}{\sqrt{M}}\sum_{i=1}^{M}\frac{R_{iM}}{\sqrt{\rho_{uM}\rho_{cM}}}m_{iM}(W_{iM},\theta_M^*)\right) = I_k. \tag{A.24}$$

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

$$\sup_{i,M}\mathbb{E}_X\left[\left\|\frac{R_{iM}}{\sqrt{\rho_{uM}\rho_{cM}}}m_{iM}(W_{iM},\theta)\right\|^r\right] \leq \frac{1}{(\rho_{uM}\rho_{cM})^{r/2-1}}\sup_{i,M}\mathbb{E}_X\left[\sup_{\theta\in\Theta}\|m_{iM}(W_{iM},\theta)\|^r\right] < \infty \tag{A.25}$$

for some $r > 2$ under condition (iii) in Theorem 3.2,

$$V_M^{-1/2}H_M(\theta_M^*)^{-1}\frac{1}{\sqrt{M}}\sum_{i=1}^{M}\frac{R_{iM}}{\sqrt{\rho_{uM}\rho_{cM}}}m_{iM}(W_{iM},\theta_M^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, I_k) \tag{A.26}$$

by Theorem 2 in Hansen and Lee (2019) under Assumption 5 and condition (iv) in Theorem 3.2.

Because of (A.26),

$$\begin{aligned} V_M^{-1/2}\sqrt{N}(\hat{\theta}_N - \theta_M^*) = &- V_M^{-1/2}H_M(\theta_M^*)^{-1}\frac{1}{\sqrt{M}}\sum_{i=1}^{M}\frac{R_{iM}}{\sqrt{\rho_{uM}\rho_{cM}}}m_{iM}(W_{iM},\theta_M^*) \\ &+ o_p(1)O_p(1) + o_p(1) \xrightarrow{d} \mathcal{N}(\mathbf{0}, I_k). \end{aligned} \tag{A.27}$$

As for Theorem 3.2(2), it is equivalent to show $\left\|V_{1M}^{-1/2}\hat{V}_{1N}V_{1M}^{-1/2} - I_k\right\| \xrightarrow{p} 0$.

Since (A.19) holds by replacing $\check{\theta}$ with $\hat{\theta}_N$,

$$\hat{H}_N(\hat{\theta}_N)^{-1} = H_M(\theta_M^*)^{-1}\big(I_k + o_p(1)\big). \tag{A.28}$$

We can write

$$
\begin{aligned}
&\hat{\Delta}_{ehw,N}(\theta)+\hat{\Delta}_{cluster,N}(\theta)\\
=&\frac{1}{N}\sum_{g=1}^{G}\left[\sum_{i=1}^{M_g}R_{igM}\cdot m_{igM}(W_{igM},\theta)\right]\left[\sum_{i=1}^{M_g}R_{igM}\cdot m_{igM}(W_{igM},\theta)\right]'\\
=&\frac{M\rho_{uM}\rho_{cM}}{N}\frac{1}{M}\sum_{g=1}^{G}\left[\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}}m_{igM}(W_{igM},\theta)\right]\left[\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}}m_{igM}(W_{igM},\theta)\right]'\\
=&\big(1+o_p(1)\big)\frac{1}{M}\sum_{g=1}^{G}\left[\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}}m_{igM}(W_{igM},\theta)\right]\left[\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}}m_{igM}(W_{igM},\theta)\right]'.
\end{aligned}
\tag{A.29}
$$

Note that

$$
\begin{aligned}
&\mathbb{E}_X\left\{\frac{1}{M}\sum_{g=1}^{G}\left[\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}}m_{igM}(W_{igM},\theta)\right]\left[\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}}m_{igM}(W_{igM},\theta)\right]'\right\}\\
=&\mathbb{E}_X\left[\frac{1}{M}\sum_{i=1}^{M}\frac{R_{iM}}{\rho_{uM}\rho_{cM}}m_{iM}(W_{iM},\theta)m_{iM}(W_{iM},\theta)'\right]\\
&+\mathbb{E}_X\left[\frac{1}{M}\sum_{g=1}^{G}\sum_{i=1}^{M_g}\sum_{j\neq i}^{M_g}\frac{R_{igM}R_{jgM}}{\rho_{uM}\rho_{cM}}m_{igM}(W_{igM},\theta)m_{jgM}(W_{jgM},\theta)'\right]\\
=&\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_X\big[m_{iM}(W_{iM},\theta)m_{iM}(W_{iM},\theta)'\big]\\
&+\frac{1}{M}\sum_{g=1}^{G}\sum_{i=1}^{M_g}\sum_{j\neq i}^{M_g}\rho_{uM}\mathbb{E}_X\big[m_{igM}(W_{igM},\theta)m_{jgM}(W_{jgM},\theta)'\big]\\
=&\Delta_{ehw,M}(\theta)+\rho_{uM}\Delta_{cluster,M}(\theta).
\end{aligned}
\tag{A.30}
$$

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

$$
\left\|\frac{1}{M}\sum_{g=1}^{G}\left[\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}}m_{igM}(W_{igM},\theta)\right]\left[\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}}m_{igM}(W_{igM},\theta)\right]'\right.
$$

$$- \big(\Delta_{ehw,M}(\theta) + \rho_{uM}\Delta_{cluster,M}(\theta)\big)\Big\| \xrightarrow{p} 0 \tag{A.31}$$

follows by (A.25) 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 $\theta$ by the DCT, Jensen's inequality, and Cauchy-Schwarz Inequality under conditions (ii) and (iii) in Theorem 3.2. In addition,

$$\begin{aligned}
&\Big\|\hat{\Delta}_{ehw,N}(\tilde{\theta}) + \hat{\Delta}_{cluster,N}(\tilde{\theta}) - \big(\hat{\Delta}_{ehw,N}(\theta) + \hat{\Delta}_{cluster,N}(\theta)\big)\Big\| \\
\leq &\frac{1}{N}\sum_{g=1}^{G}\Bigg\|\Bigg[\sum_{i=1}^{M_g} R_{igM}\cdot m_{igM}(W_{igM},\tilde{\theta})\Bigg]\Bigg[\sum_{i=1}^{M_g} R_{igM}\cdot m_{igM}(W_{igM},\tilde{\theta})\Bigg]' \\
&- \Bigg[\sum_{i=1}^{M_g} R_{igM}\cdot m_{igM}(W_{igM},\theta)\Bigg]\Bigg[\sum_{i=1}^{M_g} R_{igM}\cdot m_{igM}(W_{igM},\theta)\Bigg]'\Bigg\| \\
\leq &\frac{1}{N}\sum_{g=1}^{G} 2\sup_{\theta\in\Theta}\Bigg\|\sum_{i=1}^{M_g} R_{igM}\cdot m_{igM}(W_{igM},\theta)\Bigg\|\cdot\Bigg\|\sum_{i=1}^{M_g} R_{igM}\cdot m_{igM}(W_{igM},\tilde{\theta}) - \sum_{i=1}^{M_g} R_{igM}\cdot m_{igM}(W_{igM},\theta)\Bigg\| \\
\leq &\frac{2}{N}\sum_{g=1}^{G}\sup_{\theta\in\Theta}\Bigg\|\sum_{i=1}^{M_g} R_{igM}\cdot m_{igM}(W_{igM},\theta)\Bigg\|\sum_{i=1}^{M_g} R_{igM} b_{3,igM}(W_{igM})h(\|\tilde{\theta}-\theta\|).
\end{aligned} \tag{A.32}$$

under condition (ix) in Theorem 3.2. Let

$$\begin{aligned}
B_N^1 \equiv &\frac{2}{N}\sum_{g=1}^{G}\sup_{\theta\in\Theta}\Bigg\|\sum_{i=1}^{M_g} R_{igM}\cdot m_{igM}(W_{igM},\theta)\Bigg\|\sum_{i=1}^{M_g} R_{igM} b_{3,igM}(W_{igM}) \\
= &2\frac{M\rho_{uM}\rho_{cM}}{N}\frac{1}{M}\sum_{g=1}^{G}\sup_{\theta\in\Theta}\Bigg\|\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}} m_{igM}(W_{igM},\theta)\Bigg\|\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}} b_{3,igM}(W_{igM}) \\
= &\big(1+o_p(1)\big)\frac{2}{M}\sum_{g=1}^{G}\sup_{\theta\in\Theta}\Bigg\|\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}} m_{igM}(W_{igM},\theta)\Bigg\|\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}} b_{3,igM}(W_{igM}).
\end{aligned} \tag{A.33}$$

Since

$$\mathbb{E}_X\left[\sup_{\theta\in\Theta}\left\|\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}}m_{igM}(W_{igM},\theta)\right\|^2\right] < CM_g^2 \tag{A.34}$$

by Cr inequality and Jensen's inequality under condition (iii) in Theorem 3.2,

$$\begin{aligned}
&\mathbb{E}_X\left[\frac{2}{M}\sum_{g=1}^{G}\sup_{\theta\in\Theta}\left\|\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}}m_{igM}(W_{igM},\theta)\right\|\left\|\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}}b_{3,igM}(W_{igM})\right\|\right]\\
\leq&\frac{2}{M}\sum_{g=1}^{G}\sum_{i=1}^{M_g}\left\{\mathbb{E}_X\left[\sup_{\theta\in\Theta}\left\|\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}}m_{igM}(W_{igM},\theta)\right\|^2\right]\right\}^{1/2}\left\{\mathbb{E}_X\left[\frac{R_{igM}}{\rho_{uM}\rho_{cM}}b_{3,igM}(W_{igM})^2\right]\right\}^{1/2}\\
\leq&2C\frac{1}{M}\sum_{g=1}^{G}M_g^2<\infty
\end{aligned} \tag{A.35}$$

by Cauchy-Schwarz inequality under Assumption 5 and condition (ix) in Theorem 3.2. As a result, $B_N^1 = O_p(1)$ by Markov's inequality. Therefore, given condition (viii) in Theorem 3.2,

$$\begin{aligned}
&\left\|\left[\Delta_{ehw,M}(\theta_M^*)+\rho_{uM}\Delta_{cluster,M}(\theta_M^*)\right]^{-1}\right.\\
&\left.\left[\hat{\Delta}_{ehw,N}(\hat{\theta}_N)+\hat{\Delta}_{cluster,N}(\hat{\theta}_N)-\Delta_{ehw,M}(\theta_M^*)-\rho_{uM}\Delta_{cluster,M}(\theta_M^*)\right]\right\|\\
\leq&C\Bigg(\sup_{\theta\in\Theta}\|\hat{\Delta}_{ehw,N}(\theta)+\hat{\Delta}_{cluster,N}(\theta)-\Delta_{ehw,M}(\theta)-\rho_{uM}\Delta_{cluster,M}(\theta)\|\\
&+\left\|\Delta_{ehw,M}(\hat{\theta}_N)+\rho_{uM}\Delta_{cluster,M}(\hat{\theta}_N)-\Delta_{ehw,M}(\theta_M^*)-\rho_{uM}\Delta_{cluster,M}(\theta_M^*)\right\|\Bigg)=o_p(1)
\end{aligned} \tag{A.36}$$

by Corollary 2.2 in Newey (1991) under $\hat{\theta}_N-\theta_M^* \xrightarrow{p} \mathbf{0}$ (implied by Theorem 3.1). Hence,

$$\hat{\Delta}_{ehw,N}(\hat{\theta}_N)+\hat{\Delta}_{cluster,N}(\hat{\theta}_N)$$

$$
\begin{aligned}
=&\big(\Delta_{ehw,M}(\theta_M^*)+\rho_{uM}\Delta_{cluster,M}(\theta_M^*)\big)\Big[I_k+\big(\Delta_{ehw,M}(\theta_M^*)+\rho_{uM}\Delta_{cluster,M}(\theta_M^*)\big)^{-1}\\
&\big(\hat{\Delta}_{ehw,N}(\hat{\theta}_N)+\hat{\Delta}_{cluster,N}(\hat{\theta}_N)-\Delta_{ehw,M}(\theta_M^*)-\rho_{uM}\Delta_{cluster,M}(\theta_M^*)\big)\Big]\\
=&\big(\Delta_{ehw,M}(\theta_M^*)+\rho_{uM}\Delta_{cluster,M}(\theta_M^*)\big)\big(I_k+o_p(1)\big). \qquad \text{(A.37)}
\end{aligned}
$$

Using (A.28) and (A.37),

$$
\begin{aligned}
&\Big\|V_{1M}^{-1/2}\hat{V}_{1N}V_{1M}^{-1/2}-I_k\Big\|\\
=&\Big\|V_{1M}^{-1/2}\hat{H}_N(\hat{\theta}_N)^{-1}\big(\hat{\Delta}_{ehw,N}(\hat{\theta}_N)+\hat{\Delta}_{cluster,N}(\hat{\theta}_N)\big)\hat{H}_N(\hat{\theta}_N)^{-1}V_{1M}^{-1/2}-I_k\Big\|\\
=&\Big\|V_{1M}^{-1/2}H_M(\theta_M^*)^{-1}\big(I_k+o_p(1)\big)\big(\Delta_{ehw,M}(\theta_M^*)+\rho_{uM}\Delta_{cluster,M}(\theta_M^*)\big)\big(I_k+o_p(1)\big)\cdot\\
&H_M(\theta_M^*)^{-1}\big(I_k+o_p(1)\big)V_{1M}^{-1/2}-I_k\Big\| \qquad \text{(A.38)}\\
\le&\Big\|V_{1M}^{-1/2}V_{1M}V_{1M}^{-1/2}-I_k\Big\|+\Big\|V_{1M}^{-1/2}V_{1M}V_{1M}^{-1/2}\Big\|\,o_p(1)\\
=&o_p(1).
\end{aligned}
$$

Hence the result.

■

**Proof of Theorem 4.1**

*Proof.* Let

$$
L_M=\Big(\sum_{i=1}^{M}z_{iM}'z_{iM}\Big)^{-1}\Big\{\sum_{i=1}^{M}z_{iM}'\mathbb{E}_X\big[m_{iM}(W_{iM},\theta_M^*)\big]'\Big\}. \qquad \text{(A.39)}
$$

To show $\Big\|\hat{L}_N-L_M\Big\|\xrightarrow{p}0$, I first show

$$
\Big\|\frac{1}{N}\sum_{i=1}^{M}R_{iM}z_{iM}'z_{iM}-\frac{1}{M}\sum_{i=1}^{M}z_{iM}'z_{iM}\Big\|\xrightarrow{p}0. \qquad \text{(A.40)}
$$

We can write

$$\frac{1}{N}\sum_{i=1}^{M} R_{iM} z_{iM}' z_{iM} = \frac{M\rho_{uM}\rho_{cM}}{N}\frac{1}{M}\sum_{i=1}^{M}\frac{R_{iM}}{\rho_{uM}\rho_{cM}} z_{iM}' z_{iM}. \tag{A.41}$$

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

$$\left\| \frac{1}{M}\sum_{i=1}^{M}\frac{R_{iM}}{\rho_{uM}\rho_{cM}} z_{iM}' z_{iM} - \frac{1}{M}\sum_{i=1}^{M} z_{iM}' z_{iM} \right\| \xrightarrow{p} 0. \tag{A.42}$$

Given for some $r > 1$

$$\sup_{i,M} \mathbb{E}_X \left[ \left\| \frac{R_{iM}}{\rho_{uM}\rho_{cM}} z_{iM}' z_{iM} \right\|^r \right] = \frac{1}{(\rho_{uM}\rho_{cM})^{r-1}} \|z_{iM}\|^{2r} < \infty, \tag{A.43}$$

(A.42) is implied by Theorem 1 in Hansen and Lee (2019) under Assumption 4.

Next, I show

$$\left\| \frac{1}{N}\sum_{i=1}^{M} R_{iM} \cdot m_{iM}(W_{iM}, \hat{\theta}_N) z_{iM} - \frac{1}{M}\sum_{i=1}^{M} \mathbb{E}_X\big[m_{iM}(W_{iM}, \theta_M^*)\big] z_{iM} \right\| \xrightarrow{p} 0. \tag{A.44}$$

Again, we can write

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{M} R_{iM} \cdot m_{iM}(W_{iM}, \hat{\theta}_N) z_{iM} =& \frac{M\rho_{uM}\rho_{cM}}{N}\frac{1}{M}\sum_{i=1}^{M}\frac{R_{iM}}{\rho_{uM}\rho_{cM}} m_{iM}(W_{iM}, \hat{\theta}_N) z_{iM} \\ =& \big(1 + o_p(1)\big)\frac{1}{M}\sum_{i=1}^{M}\frac{R_{iM}}{\rho_{uM}\rho_{cM}} m_{iM}(W_{iM}, \hat{\theta}_N) z_{iM} \end{aligned} \tag{A.45}$$

I first show $\forall\ \theta \in \Theta$

$$\left\| \frac{1}{M}\sum_{i=1}^{M}\frac{R_{iM}}{\rho_{uM}\rho_{cM}} m_{iM}(W_{iM}, \theta) z_{iM} - \frac{1}{M}\sum_{i=1}^{M} \mathbb{E}_X\big[m_{iM}(W_{iM}, \theta)\big] z_{iM} \right\| \xrightarrow{p} 0. \tag{A.46}$$

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

$$\begin{aligned}
&\sup_{i,M} \mathbb{E}_X \left[ \left\| \frac{R_{iM}}{\rho_{uM}\rho_{cM}} m_{iM}(W_{iM}, \theta) z_{iM} \right\|^r \right] \\
\leq &\frac{1}{(\rho_{uM}\rho_{cM})^{r-1}} \left\{ \sup_{i,M} \mathbb{E}_X \left[ \sup_{\theta \in \Theta} \| m_{iM}(W_{iM}, \theta) \|^{2r} \right] \right\}^{1/2} \sup_{i,M} \| z_{iM} \|^r < \infty
\end{aligned} \tag{A.47}$$

for some $r > 1$ by Jensen's inequality under condition (iii) in Theorem 3.2, (A.46) holds by Theorem 1 in Hansen and Lee (2019) under Assumption 4. Next, I show the Lipschitz condition. $\forall\ \tilde{\theta}, \theta \in \Theta$

$$\begin{aligned}
\left\| m_{iM}(W_{iM}, \tilde{\theta}) z_{iM} - m_{iM}(W_{iM}, \theta) z_{iM} \right\| &\leq \| z_{iM} \| \cdot \left\| m_{iM}(W_{iM}, \tilde{\theta}) - m_{iM}(W_{iM}, \theta) \right\| \\
&\leq \| z_{iM} \|\, b_{3,iM}(W_{iM}) h(\| \tilde{\theta} - \theta \|),
\end{aligned} \tag{A.48}$$

and

$$\sup_{i,M} \mathbb{E}_X \left[ \| z_{iM} \|\, b_{3,iM}(W_{iM}) \right] \leq \sup_{i,M} \| z_{iM} \| \sup_{i,M} \left\{ \mathbb{E}_X \left[ b_{3,iM}(W_{iM})^2 \right] \right\}^{1/2} < \infty \tag{A.49}$$

by Jensen's inequality under condition (ix) in Theorem 3.2. Also, $\frac{1}{M} \sum_{i=1}^M \mathbb{E}_X \left[ m_{iM}(W_{iM}, \theta) \right] z_{iM}$ is continuous in $\theta$ by the DCT and Jensen's inequality under conditions (ii) and (iii) in Theorem 3.2. As a result,

$$\begin{aligned}
&\left\| \frac{1}{M} \sum_{i=1}^M \frac{R_{iM}}{\rho_{uM}\rho_{cM}} m_{iM}(W_{iM}, \hat{\theta}_N) z_{iM} - \frac{1}{M} \sum_{i=1}^M \mathbb{E}_X \left[ m_{iM}(W_{iM}, \theta_M^*) \right] z_{iM} \right\| \\
\leq &\sup_{\theta \in \Theta} \left\| \frac{1}{M} \sum_{i=1}^M \frac{R_{iM}}{\rho_{uM}\rho_{cM}} m_{iM}(W_{iM}, \theta) z_{iM} - \frac{1}{M} \sum_{i=1}^M \mathbb{E}_X \left[ m_{iM}(W_{iM}, \theta) \right] z_{iM} \right\| \\
&+ \left\| \frac{1}{M} \sum_{i=1}^M \mathbb{E}_X \left[ m_{iM}(W_{iM}, \hat{\theta}_N) \right] z_{iM} - \frac{1}{M} \sum_{i=1}^M \mathbb{E}_X \left[ m_{iM}(W_{iM}, \theta_M^*) \right] z_{iM} \right\| \xrightarrow{p} 0
\end{aligned} \tag{A.50}$$

by Lemma A.2 and Corollary 2.2 in Newey (1991) under $\hat{\theta}_N - \theta^*_M \xrightarrow{p} \mathbf{0}$. Combining (A.40) and (A.44), we conclude that $\left\|\hat{L}_N - L_M\right\| \xrightarrow{p} 0$.

Hence,

$$\begin{aligned}
\hat{\Delta}^Z_N =& \frac{M\rho_{uM}\rho_{cM}}{N}\frac{1}{M}\sum_{i=1}^{M}\frac{R_{iM}}{\rho_{uM}\rho_{cM}}\big(L'_M + o_p(1)\big)z'_{iM}z_{iM}\big(L_M + o_p(1)\big) \\
=& \big(1 + o_p(1)\big)\frac{1}{M}\sum_{i=1}^{M}\frac{R_{iM}}{\rho_{uM}\rho_{cM}}\big(L'_M + o_p(1)\big)z'_{iM}z_{iM}\big(L_M + o_p(1)\big) \\
=& \frac{1}{M}\sum_{i=1}^{M}\frac{R_{iM}}{\rho_{uM}\rho_{cM}}L'_M z'_{iM}z_{iM}L_M + \frac{1}{M}\sum_{i=1}^{M}\frac{R_{iM}}{\rho_{uM}\rho_{cM}}L'_M z'_{iM}z_{iM}L_M \cdot o_p(1) \\
&+ \frac{1}{M}\sum_{i=1}^{M}\frac{R_{iM}}{\rho_{uM}\rho_{cM}}L'_M z'_{iM}z_{iM} \cdot o_p(1) + \frac{1}{M}\sum_{i=1}^{M}\frac{R_{iM}}{\rho_{uM}\rho_{cM}}z'_{iM}z_{iM}L_M \cdot o_p(1) \\
&+ \frac{1}{M}\sum_{i=1}^{M}\frac{R_{iM}}{\rho_{uM}\rho_{cM}}z'_{iM}z_{iM} \cdot o_p(1). \qquad \text{(A.51)}
\end{aligned}$$

Let

$$\Delta^Z_M = \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_X\big[m_{iM}(W_{iM},\theta^*_M)\big]z_{iM}\left(\frac{1}{M}\sum_{i=1}^{M}z'_{iM}z_{iM}\right)^{-1}\frac{1}{M}\sum_{i=1}^{M}z'_{iM}\mathbb{E}_X\big[m_{iM}(W_{iM},\theta^*_M)\big]'. \qquad \text{(A.52)}$$

$$\left\|{\Delta^Z_M}^{-1/2}\hat{\Delta}^Z_N{\Delta^Z_M}^{-1/2} - I_k\right\| = o_p(1) \qquad \text{(A.53)}$$

follows by (A.43) and Theorem 1 in Hansen and Lee (2019).

Let $A_M$ and $D_M$ be the matrices with $i$-th rows equal to $\mathbb{E}_X\big[m_{iM}(W_{iM},\theta^*_M)\big]'/\sqrt{M}$ and $z_{iM}/\sqrt{M}$ respectively. Let $I_M$ be the identity matrix of size $M$. Then,

$$\Delta_{E,M} - \Delta^Z_M = A'_M\big(I_M - D_M(D'_M D_M)^{-1}D'_M\big)A_M, \qquad \text{(A.54)}$$

which is positive semidefinite. Hence, the result. ■

**Proof of Theorem 4.2**

*Proof.* Let

$$P_M = \left[\sum_{g=1}^{G} \tilde{z}'_{gM}\tilde{z}_{gM}\right]^{-1} \sum_{g=1}^{G} \tilde{z}'_{gM}\mathbb{E}_X\left[\tilde{m}_{gM}(\theta^*_M)\right]'. \tag{A.55}$$

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

$$\left\|\frac{1}{N}\sum_{g=1}^{G} R_{gM}\tilde{z}'_{gM}\tilde{z}_{gM} - \frac{1}{M}\sum_{g=1}^{G} \tilde{z}'_{gM}\tilde{z}_{gM}\right\| \xrightarrow{p} 0. \tag{A.56}$$

We can write

$$\frac{1}{N}\sum_{g=1}^{G} R_{gM}\tilde{z}'_{gM}\tilde{z}_{gM} = \frac{M\rho_{cM}}{N}\frac{1}{M}\sum_{g=1}^{G}\frac{R_{gM}}{\rho_{cM}}\tilde{z}'_{gM}\tilde{z}_{gM}. \tag{A.57}$$

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

$$\left\|\frac{1}{M}\sum_{g=1}^{G}\frac{R_{gM}}{\rho_{cM}}\tilde{z}'_{gM}\tilde{z}_{gM} - \frac{1}{M}\sum_{g=1}^{G} \tilde{z}'_{gM}\tilde{z}_{gM}\right\| \xrightarrow{p} 0, \tag{A.58}$$

Because for some $r > 2$

$$\sup_{i,g,M} \mathbb{E}_X\left(\left\|\frac{R_{gM}}{\sqrt{\rho_{cM}}} z_{igM}\right\|^r\right) < \infty, \tag{A.59}$$

(A.58) follows by the proof of (62) in Hansen and Lee (2019)

Next, I show

$$\left\|\frac{1}{N}\sum_{g=1}^{G} R_{gM}\tilde{m}_{gM}(\hat{\theta}_N)\tilde{z}_{gM} - \frac{1}{M}\sum_{g=1}^{G}\mathbb{E}_X\left[\tilde{m}_{gM}(\theta^*_M)\right]\tilde{z}_{gM}\right\| \xrightarrow{p} 0. \tag{A.60}$$

Again, we can write

$$\begin{aligned}\frac{1}{N}\sum_{g=1}^{G} R_{gM}\tilde{m}_{gM}(\hat{\theta}_N)\tilde{z}_{gM} =& \frac{M\rho_{cM}}{N}\frac{1}{M}\sum_{g=1}^{G}\frac{R_{gM}}{\rho_{cM}}\tilde{m}_{gM}(\hat{\theta}_N)\tilde{z}_{gM} \\ =& \big(1+o_p(1)\big)\frac{1}{M}\sum_{g=1}^{G}\frac{R_{gM}}{\rho_{cM}}\tilde{m}_{gM}(\hat{\theta}_N)\tilde{z}_{gM}\end{aligned} \tag{A.61}$$

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

$$\left\| \frac{1}{M}\sum_{g=1}^{G}\frac{R_{gM}}{\rho_{cM}}\tilde{m}_{gM}(\theta)\tilde{z}_{gM} - \frac{1}{M}\sum_{g=1}^{G}\mathbb{E}_X\big[\tilde{m}_{gM}(\theta)\big]\tilde{z}_{gM}\right\| \xrightarrow{p} 0. \tag{A.62}$$

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

$$\tilde{l}_{gM} = \frac{R_{gM}}{\rho_{cM}}\tilde{m}_{gM}(\theta)\tilde{z}_{gM}\mathbb{1}\left(\frac{R_{gM}}{\rho_{cM}}\left\|\tilde{m}_{gM}(\theta)\tilde{z}_{gM}\right\| \le M\epsilon\right). \tag{A.63}$$

Then

$$\begin{aligned}&\mathbb{E}_X\left[\left\|\frac{1}{M}\sum_{g=1}^{G}\frac{R_{gM}}{\rho_{cM}}\tilde{m}_{gM}(\theta)\tilde{z}_{gM} - \frac{1}{M}\sum_{g=1}^{G}\mathbb{E}_X\big[\tilde{m}_{gM}(\theta)\big]\tilde{z}_{gM}\right\|\right] \\ \le& \frac{1}{M}\mathbb{E}_X\left\{\left\|\sum_{g=1}^{G}\Big[\tilde{l}_{gM} - \mathbb{E}_X\big(\tilde{l}_{gM}\big)\Big]\right\|\right\} \\ &+ \frac{2}{M}\sum_{g=1}^{G}\mathbb{E}_X\left[\left\|\tilde{m}_{gM}(\theta)\tilde{z}_{gM}\right\|\mathbb{1}\left(\frac{R_{gM}}{\rho_{cM}}\left\|\tilde{m}_{gM}(\theta)\tilde{z}_{gM}\right\| > M\epsilon\right)\right].\end{aligned} \tag{A.64}$$

Observe that

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

$$\leq \frac{1}{M}\left\{\sum_{g=1}^{G}\mathbb{E}_X\left[\left\|\tilde{l}_{gM}\right\|^2\right]\right\}^{1/2} \leq (\epsilon C)^{1/2}\left(\frac{1}{M}\sum_{g=1}^{G}M_g^2\right)^{1/2} \leq \delta \tag{A.65}$$

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

$$\begin{aligned}&\sup_{g,M}\mathbb{E}_X\left[\left\|\frac{R_{gM}}{\rho_{cM}}\tilde{m}_{gM}(\theta)\tilde{z}_{gM}/M_g^2\right\|^r\right]\\ \leq &\frac{1}{\rho_{cM}^{r-1}}\left\{\sup_{g,M}\mathbb{E}_X\left[\sup_{\theta\in\Theta}\|\tilde{m}_{gM}(\theta)/M_g\|^{2r}\right]\right\}^{1/2}\sup_{g,M}\|\tilde{z}_{gM}/M_g\|^r \leq \infty\end{aligned} \tag{A.66}$$

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

$$\sup_{g,M}\mathbb{E}_X\left[\left\|\frac{R_{gM}}{\rho_{cM}}\tilde{m}_{gM}(\theta)\tilde{z}_{gM}/M_g^2\right\|\mathbb{1}\left(\left\|\frac{R_{gM}}{\rho_{cM}}\tilde{m}_{gM}(\theta)\tilde{z}_{gM}/M_g^2\right\| > B\right)\right] \leq \frac{\delta}{C}. \tag{A.67}$$

Pick $M$ large enough so that

$$\max_{g\leq G}\frac{M_g^2}{M} \leq \frac{\epsilon}{B}, \tag{A.68}$$

which is feasible under Assumption 5. Then,

$$\frac{2}{M}\sum_{g=1}^{G}\mathbb{E}_X\left[\|\tilde{m}_{gM}(\theta)\tilde{z}_{gM}\|\,\mathbb{1}\left(\frac{R_{gM}}{\rho_{cM}}\|\tilde{m}_{gM}(\theta)\tilde{z}_{gM}\| > M\epsilon\right)\right] \leq \frac{2}{M}\sum_{g=1}^{G}M_g^2\frac{\delta}{C} \leq 2\delta. \tag{A.69}$$

Combining (A.65) and (A.69), (A.62) holds by Markov's inequality.

Next,

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

$$
\begin{aligned}
=&\frac{1}{N}\sum_{g=1}^{G} R_{gM}\left\|\sum_{i=1}^{M_g}\sum_{j=1}^{M_g} m_{igM}(W_{igM},\tilde{\theta})z_{jgM}-\sum_{i=1}^{M_g}\sum_{j=1}^{M_g} m_{igM}(W_{igM},\theta)z_{jgM}\right\| \\
\leq&\frac{1}{N}\sum_{g=1}^{G} R_{gM}\sum_{i=1}^{M_g}\sum_{j=1}^{M_g}\left\|m_{igM}(W_{igM},\tilde{\theta})-m_{igM}(W_{igM},\theta)\right\|\cdot\|z_{jgM}\| \\
\leq&\frac{1}{N}\sum_{g=1}^{G} R_{gM}\sum_{i=1}^{M_g}\sum_{j=1}^{M_g} b_{3,igM}(W_{igM})\cdot\|z_{jgM}\|\cdot h(\|\tilde{\theta}-\theta\|)
\end{aligned}
\tag{A.70}
$$

Let

$$
\begin{aligned}
B_N^2 \equiv&\frac{1}{N}\sum_{g=1}^{G} R_{gM}\sum_{i=1}^{M_g}\sum_{j=1}^{M_g} b_{3,igM}(W_{igM})\cdot\|z_{jgM}\| \\
=&\frac{M\rho_{cM}}{N}\frac{1}{M}\sum_{g=1}^{G}\frac{R_{gM}}{\rho_{cM}}\sum_{i=1}^{M_g}\sum_{j=1}^{M_g} b_{3,igM}(W_{igM})\cdot\|z_{jgM}\| \\
=&\big(1+o_p(1)\big)\frac{1}{M}\sum_{g=1}^{G}\frac{R_{gM}}{\rho_{cM}}\sum_{i=1}^{M_g}\sum_{j=1}^{M_g} b_{3,igM}(W_{igM})\cdot\|z_{jgM}\|
\end{aligned}
\tag{A.71}
$$

Since

$$
\begin{aligned}
&\mathbb{E}_X\left[\frac{1}{M}\sum_{g=1}^{G}\frac{R_{gM}}{\rho_{cM}}\sum_{i=1}^{M_g}\sum_{j=1}^{M_g} b_{3,igM}(W_{igM})\cdot\|z_{igM}\|\right] \\
\leq&\frac{1}{M}\sum_{g=1}^{G}\mathbb{E}\Big(\frac{R_{gM}}{\rho_{cM}}\Big)\sum_{i=1}^{M_g}\sum_{j=1}^{M_g}\mathbb{E}_X\big[b_{3,igM}(W_{igM})\big]\,\|z_{igM}\| \\
\leq&\frac{1}{M}\sum_{g=1}^{G} M_g^2\sup_{i,g,M}\Big\{\mathbb{E}_X\big[b_{3,igM}(W_{igM})^2\big]\Big\}^{1/2}\sup_{i,g,M}\|z_{igM}\|<\infty
\end{aligned}
\tag{A.72}
$$

by Jensen's inequality under condition (ix) 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 $\theta$ by the DCT and Jensen's inequality under Assumption 5 and conditions (ii) and (iii) in Theorem 3.2. As a result,

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

$$\leq \sup_{\theta\in\Theta} \left\| \frac{1}{N}\sum_{g=1}^{G} R_{gM}\tilde{m}_{gM}(\theta)\tilde{z}_{gM} - \frac{1}{M}\sum_{g=1}^{G}\mathbb{E}_X\big[\tilde{m}_{gM}(\theta)\big]\tilde{z}_{gM} \right\|$$
$$+ \left\| \frac{1}{M}\sum_{g=1}^{G}\mathbb{E}_X\big[\tilde{m}_{gM}(\hat{\theta}_N)\big]\tilde{z}_{gM} - \frac{1}{M}\sum_{g=1}^{G}\mathbb{E}_X\big[\tilde{m}_{gM}(\theta^*_M)\big]\tilde{z}_{gM} \right\| \xrightarrow{p} 0. \tag{A.73}$$

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 (A.56) and (A.60) 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

$$\begin{aligned}
\hat{\Delta}^Z_{CE,N} =& \frac{M\rho_{cM}}{N}\frac{1}{M}\sum_{g=1}^{G}\frac{R_{gM}}{\rho_{cM}}\hat{P}'_N\tilde{z}'_{gM}\tilde{z}_{gM}\hat{P}_N \\
=& \big(1+o_p(1)\big)\frac{1}{M}\sum_{g=1}^{G}\frac{R_{gM}}{\rho_{cM}}\big(P'_M+o_p(1)\big)\tilde{z}'_{gM}\tilde{z}_{gM}\big(P_M+o_p(1)\big) \\
=& \frac{1}{M}\sum_{g=1}^{G}\frac{R_{gM}}{\rho_{cM}}P'_M\tilde{z}'_{gM}\tilde{z}_{gM}P_M + \frac{1}{M}\sum_{g=1}^{G}\frac{R_{gM}}{\rho_{cM}}P'_M\tilde{z}'_{gM}\tilde{z}_{gM}P_M\cdot o_p(1) \\
&+ \frac{1}{M}\sum_{g=1}^{G}\frac{R_{gM}}{\rho_{cM}}P'_M\tilde{z}'_{gM}\tilde{z}_{gM}\cdot o_p(1) + \frac{1}{M}\sum_{g=1}^{G}\frac{R_{gM}}{\rho_{cM}}\tilde{z}'_{gM}\tilde{z}_{gM}P_M\cdot o_p(1) \\
&+ \frac{1}{M}\sum_{g=1}^{G}\frac{R_{gM}}{\rho_{cM}}\tilde{z}'_{gM}\tilde{z}_{gM}\cdot o_p(1).
\end{aligned} \tag{A.74}$$

$$\left\|{\Delta^Z_{CE,M}}^{-1/2}\hat{\Delta}^Z_{CE,N}{\Delta^Z_{CE,M}}^{-1/2} - I_k\right\| = o_p(1) \tag{A.75}$$

follows from the similar arguments of the proof of (62) in Hansen and Lee (2019) under Assumption 5.

To show the ordering of the variance-covariance matrices in Theorem 4.2, notice that

$$\Delta_{E,M} + \Delta_{EC,M} - \Delta^Z_{CE,M}$$

$$
\begin{aligned}
=&\frac{1}{M}\sum_{g=1}^{G}\mathbb{E}_X\left[\tilde{m}_{gM}(\theta_M^*)\right]\mathbb{E}_X\left[\tilde{m}_{gM}(\theta_M^*)\right]' \\
&-\frac{1}{M}\sum_{g=1}^{G}\mathbb{E}_X\left[\tilde{m}_{gM}(\theta_M^*)\right]\tilde{z}_{gM}\left[\frac{1}{M}\sum_{g=1}^{G}\tilde{z}_{gM}'\tilde{z}_{gM}\right]^{-1}\frac{1}{M}\sum_{g=1}^{G}\tilde{z}_{gM}'\mathbb{E}_X\left[\tilde{m}_{gM}(\theta_M^*)\right]', \qquad \text{(A.76)}
\end{aligned}
$$

which is positive semidefinite.

Hence, the result. ■

**Proof of Theorem 5.1**

*Proof.* First, using similar arguments in the proof of Theorem 3.2,

$$
\hat{\gamma}_N - \gamma_M^* \xrightarrow{p} \mathbf{0} \qquad \text{(A.77)}
$$

under conditions (i), (ii), and (vii) in Theorem 5.1.

By a mean value expansion around $\theta_M^*$,

$$
\begin{aligned}
&V_{f,M}^{-1/2}\frac{1}{\sqrt{N}}\sum_{i=1}^{M}R_{iM}f_{iM}(W_{iM},\hat{\theta}_N) \\
=&V_{f,M}^{-1/2}\frac{1}{\sqrt{N}}\sum_{i=1}^{M}R_{iM}f_{iM}(W_{iM},\theta_M^*) \qquad \text{(A.78)} \\
&+V_{f,M}^{-1/2}\frac{1}{N}\sum_{i=1}^{M}R_{iM}\nabla_\theta f_{iM}(W_{iM},\check{\theta})\sqrt{N}(\hat{\theta}_N-\theta_M^*),
\end{aligned}
$$

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

Given Theorem 3.2, $V_M^{-1/2}\sqrt{N}(\hat{\theta}_N-\theta_M^*)=O_p(1)$. Further,

$$
\hat{F}_N(\check{\theta}) = F_M(\theta_M^*) + o_p(1) \qquad \text{(A.79)}
$$

under conditions (i), (iv), and (v) in Theorem 5.1. Therefore,

$$V_{f,M}^{-1/2}\frac{1}{N}\sum_{i=1}^{M}R_{iM}\nabla_\theta f_{iM}(W_{iM},\tilde{\theta})\sqrt{N}(\hat{\theta}_N-\theta_M^*)=V_{f,M}^{-1/2}F_M(\theta_M^*)\sqrt{N}(\hat{\theta}_N-\theta_M^*)+o_p(1). \tag{A.80}$$

According to the mean value expansion in the proof of the asymptotic normality of $V_M^{-1/2}\sqrt{N}(\hat{\theta}_N-\theta_M^*)$,

$$V_M^{-1/2}\sqrt{N}(\hat{\theta}_N-\theta_M^*)=-V_M^{-1/2}\frac{1}{\sqrt{N}}\sum_{i=1}^{M}R_{iM}H_M(\theta_M^*)^{-1}m_{iM}(W_{iM},\theta_M^*)+o_p(1). \tag{A.81}$$

Combining (A.78), (A.80), and (A.81),

$$\begin{aligned}&V_{f,M}^{-1/2}\frac{1}{\sqrt{N}}\sum_{i=1}^{M}R_{iM}f_{iM}(W_{iM},\hat{\theta}_N)\\ =&V_{f,M}^{-1/2}\frac{1}{\sqrt{N}}\sum_{i=1}^{M}R_{iM}\big[f_{iM}(W_{iM},\theta_M^*)-F_M(\theta_M^*)H_M(\theta_M^*)^{-1}m_{iM}(W_{iM},\theta_M^*)\big]+o_p(1).\end{aligned} \tag{A.82}$$

Subtract $V_{f,M}^{-1/2}\sqrt{N}\gamma_M^*$ from both sides of (A.82).

$$\begin{aligned}&V_{f,M}^{-1/2}\sqrt{N}\big(\hat{\gamma}-\gamma_M^*\big)\\ =&V_{f,M}^{-1/2}\frac{1}{\sqrt{N}}\sum_{i=1}^{M}R_{iM}\big[f_{iM}(W_{iM},\theta_M^*)-\gamma_M^*-F_M(\theta_M^*)H_M(\theta_M^*)^{-1}m_{iM}(W_{iM},\theta_M^*)\big]+o_p(1)\\ =&V_{f,M}^{-1/2}\sqrt{\frac{M\rho_{uM}\rho_{cM}}{N}}\frac{1}{\sqrt{M}}\sum_{i=1}^{M}\frac{R_{iM}}{\sqrt{\rho_{uM}\rho_{cM}}}\big[f_{iM}(W_{iM},\theta_M^*)-\gamma_M^*-F_M(\theta_M^*)H_M(\theta_M^*)^{-1}m_{iM}(W_{iM},\theta_M^*)\big]+o_p(1)\\ =&\big(1+o_p(1)\big)V_{f,M}^{-1/2}\frac{1}{\sqrt{M}}\sum_{i=1}^{M}\frac{R_{iM}}{\sqrt{\rho_{uM}\rho_{cM}}}\big[f_{iM}(W_{iM},\theta_M^*)-\gamma_M^*-F_M(\theta_M^*)H_M(\theta_M^*)^{-1}m_{iM}(W_{iM},\theta_M^*)\big]+o_p(1)\end{aligned} \tag{A.83}$$

Observe that $\forall\ \theta \in \Theta$

$$\begin{aligned}
&\sup_{i,M} \mathbb{E}_X \left\{ \left\| \frac{R_{iM}}{\sqrt{\rho_{uM}\rho_{cM}}} \left[ f_{iM}(W_{iM},\theta) - \gamma_M^* - F_M(\theta_M^*) H_M(\theta_M^*)^{-1} m_{iM}(W_{iM},\theta) \right] \right\|^r \right\} \\
\leq & \frac{1}{(\rho_{uM}\rho_{cM})^{r/2-1}} \left\{ \left[ \sup_{i,M} \mathbb{E}_X \Big( \sup_{\theta\in\Theta} \| f_{iM}(W_{iM},\theta) \|^r \Big) \right]^{1/r} + \|\gamma_M^*\| \right. \\
&\left. + C \left\| F_M(\theta_M^*) \right\| \left[ \sup_{i,M} \mathbb{E}_X \Big( \sup_{\theta\in\Theta} \| m_{iM}(W_{iM},\theta) \|^r \Big) \right]^{1/r} \right\}^r < \infty
\end{aligned} \tag{A.84}$$

for some $r > 2$ by Minkowski's inequality and Jensen's inequality under conditions (ii) and (iv) in Theorem 5.1, and condition (iii) in Theorem 3.2. Also,

$$\begin{aligned}
&\mathbb{V}_X \left\{ \frac{1}{\sqrt{M}} \sum_{i=1}^{M} \frac{R_{iM}}{\sqrt{\rho_{uM}\rho_{cM}}} \left[ f_{iM}(W_{iM},\theta_M^*) - \gamma_M^* - F_M(\theta_M^*) H_M(\theta_M^*)^{-1} m_{iM}(W_{iM},\theta_M^*) \right] \right\} \\
=& \Delta^f_{ehw,M} - \rho_{uM}\rho_{cM}\Delta^f_{E,M} + \rho_{uM}\Delta^f_{cluster,M} - \rho_{uM}\rho_{cM}\Delta^f_{EC,M}.
\end{aligned} \tag{A.85}$$

By Theorem 2 in Hansen and Lee (2019)

$$V_{f,M}^{-1/2} \frac{1}{\sqrt{M}} \sum_{i=1}^{M} \frac{R_{iM}}{\sqrt{\rho_{uM}\rho_{cM}}} \left[ f_{iM}(W_{iM},\theta_M^*) - \gamma_M^* - F_M(\theta_M^*) H_M(\theta_M^*)^{-1} m_{iM}(W_{iM},\theta_M^*) \right] \xrightarrow{d} \mathcal{N}(\mathbf{0}, I_q) \tag{A.86}$$

under condition (iii) in Theorem 5.1 and Assumption 5.

To show Theorem 5.1(2), observe that

$$\begin{aligned}
&\hat{\Delta}^f_{ehw,N} + \hat{\Delta}^f_{cluster,N} \\
=& \frac{M\rho_{uM}\rho_{cM}}{N} \frac{1}{M} \sum_{g=1}^{G} \left\{ \sum_{i=1}^{M_g} \frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}} \left[ f_{igM}(W_{igM},\hat{\theta}_N) - \hat{\gamma}_N - \hat{F}_N(\hat{\theta}_N) \hat{H}_N(\hat{\theta}_N)^{-1} m_{igM}(W_{igM},\hat{\theta}_N) \right] \right\} \cdot \\
&\left\{ \sum_{i=1}^{M_g} \frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}} \left[ f_{igM}(W_{igM},\hat{\theta}_N) - \hat{\gamma}_N - \hat{F}_N(\hat{\theta}_N) \hat{H}_N(\hat{\theta}_N)^{-1} m_{igM}(W_{igM},\hat{\theta}_N) \right] \right\}'
\end{aligned}$$

$$
\begin{aligned}
=&\big(1+o_p(1)\big)\frac{1}{M}\sum_{g=1}^{G}\Bigg\{\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}}\Big[f_{igM}(W_{igM},\hat{\theta}_N)-\gamma_M^*+o_p(1)\\
&-\big(F_M(\theta_M^*)+o_p(1)\big)H_M(\theta_M^*)^{-1}\big(I_k+o_p(1)\big)m_{igM}(W_{igM},\hat{\theta}_N)\Big]\Bigg\}\cdot\\
&\Bigg\{\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}}\Big[f_{igM}(W_{igM},\hat{\theta}_N)-\gamma_M^*+o_p(1)\\
&-\big(F_M(\theta_M^*)+o_p(1)\big)H_M(\theta_M^*)^{-1}\big(I_k+o_p(1)\big)m_{igM}(W_{igM},\hat{\theta}_N)\Big]\Bigg\}'\\
=&\big(1+o_p(1)\big)\frac{1}{M}\sum_{g=1}^{G}\Bigg\{\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}}\big[f_{igM}(W_{igM},\hat{\theta}_N)-\gamma_M^*-F_M(\theta_M^*)H_M(\theta_M^*)^{-1}m_{igM}(W_{igM},\hat{\theta}_N)\big]\Bigg\}\cdot\\
&\Bigg\{\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}}\big[f_{igM}(W_{igM},\hat{\theta}_N)-\gamma_M^*-F_M(\theta_M^*)H_M(\theta_M^*)^{-1}m_{igM}(W_{igM},\hat{\theta}_N)\big]\Bigg\}'+o_p(1)
\end{aligned}
\tag{A.87}
$$

under condition (ii) in Theorem 5.1, condition (iii) in Theorem 3.2, and Assumption 5. Denote

$$
\begin{aligned}
\tilde{\Delta}(\theta)=&\frac{1}{M}\sum_{g=1}^{G}\Bigg\{\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}}\big[f_{igM}(W_{igM},\theta)-\gamma_M^*-F_M(\theta_M^*)H_M(\theta_M^*)^{-1}m_{igM}(W_{igM},\theta)\big]\Bigg\}\cdot\\
&\Bigg\{\sum_{i=1}^{M_g}\frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}}\big[f_{igM}(W_{igM},\theta)-\gamma_M^*-F_M(\theta_M^*)H_M(\theta_M^*)^{-1}m_{igM}(W_{igM},\theta)\big]\Bigg\}'.
\end{aligned}
\tag{A.88}
$$

It suffices to show

$$
\Big\|(\Delta_{ehw,M}^f+\rho_{uM}\Delta_{cluster,M}^f)^{-1/2}\tilde{\Delta}(\hat{\theta}_N)(\Delta_{ehw,M}^f+\rho_{uM}\Delta_{cluster,M}^f)^{-1/2}-I_q\Big\|=o_p(1).
\tag{A.89}
$$

Note that

$$\begin{aligned}
&\left\| \frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}} \Big\{ \big[ f_{igM}(W_{igM}, \tilde{\theta}) - \gamma_M^* - F_M(\theta_M^*) H_M(\theta_M^*)^{-1} m_{igM}(W_{igM}, \tilde{\theta}) \big] \right. \\
&\left. - \big[ f_{igM}(W_{igM}, \theta) - \gamma_M^* - F_M(\theta_M^*) H_M(\theta_M^*)^{-1} m_{igM}(W_{igM}, \theta) \big] \Big\} \right\| \\
\leq & \frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}} \left[ \left\| f_{igM}(W_{igM}, \tilde{\theta}) - f_{igM}(W_{igM}, \theta) \right\| + C \left\| F_M(\theta_M^*) \right\| \left\| m_{igM}(W_{igM}, \tilde{\theta}) - m_{igM}(W_{igM}, \theta) \right\| \right] \\
\leq & \frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}} \big[ b_{5,igM}(W_{igM}) h(\|\tilde{\theta} - \theta\|) + C \left\| F_M(\theta_M^*) \right\| b_{3,igM}(W_{igM}) h(\|\tilde{\theta} - \theta\|) \big].
\end{aligned} \tag{A.90}$$

Let

$$b_{6,igM}(W_{igM}) = \frac{R_{igM}}{\sqrt{\rho_{uM}\rho_{cM}}} \big[ b_{5,igM}(W_{igM}) + C \left\| F_M(\theta_M^*) \right\| b_{3,igM}(W_{igM}) \big]. \tag{A.91}$$

Observe that

$$\begin{aligned}
\sup_{i,M} \mathbb{E}_X \big[ b_{6,igM}(W_{igM})^2 \big] \leq & \sup_{i,M} \mathbb{E}_X \big[ b_{5,igM}(W_{igM})^2 \big] + C \sup_{i,M} \mathbb{E}_X \big[ b_{3,igM}(W_{igM})^2 \big] \\
& + C \Big\{ \sup_{i,M} \mathbb{E}_X \big[ b_{5,igM}(W_{igM})^2 \big] \sup_{i,M} \mathbb{E}_X \big[ b_{3,igM}(W_{igM})^2 \big] \Big\}^{1/2} < \infty
\end{aligned} \tag{A.92}$$

by Cauchy-Schwarz inequality under condition (ix) in Theorem 3.2 and condition (vii) in Theorem 5.1. Therefore, (A.89) follows from similar arguments in the proof of Theorem 3.2(2). ■

# B Additional Tables

Table 3: Standard Errors and Coverage Rates for Probit: the Coefficient Estimator

| | | No Cluster Assignment | | | With Cluster Assignment | | |
|---|---|---|---|---|---|---|---|
| | | (1) | (2) | (3) | (4) | (5) | (6) |
| | | $\rho_c = 0.1$ | $\rho_c = 0.5$ | $\rho_c = 1$ | $\rho_c = 0.1$ | $\rho_c = 0.5$ | $\rho_c = 1$ |
| $G\rho_c = 50$ | $std$ | 0.2199 | 0.1831 | 0.1137 | 0.2443 | 0.2154 | 0.1680 |
| | $se_{limit}$ | 0.2126 | 0.1773 | 0.1118 | 0.2340 | 0.2052 | 0.1600 |
| | $\bar{se}_{cluster}$ | 0.2242 | 0.2279 | 0.2255 | 0.2460 | 0.2511 | 0.2541 |
| | $cov_{cluster}$ | (0.957) | (0.985) | (1.000) | (0.954) | (0.978) | (0.997) |
| | $\bar{se}_{adj}$ | 0.2180 | 0.1912 | 0.1447 | 0.2403 | 0.2182 | 0.1850 |
| | $cov_{adj}$ | (0.950) | (0.963) | (0.988) | (0.951) | (0.958) | (0.969) |
| | $\bar{se}_{ehw,adj}$ | 0.1510 | 0.1433 | 0.1339 | 0.1530 | 0.1452 | 0.1358 |
| $G\rho_c = 100$ | $std$ | 0.1544 | 0.1287 | 0.0781 | 0.1723 | 0.1472 | 0.1085 |
| | $se_{limit}$ | 0.1512 | 0.1258 | 0.0785 | 0.1682 | 0.1446 | 0.1073 |
| | $\bar{se}_{cluster}$ | 0.1587 | 0.1603 | 0.1596 | 0.1753 | 0.1758 | 0.1760 |
| | $cov_{cluster}$ | (0.958) | (0.986) | (1.000) | (0.952) | (0.982) | (0.999) |
| | $\bar{se}_{adj}$ | 0.1542 | 0.1351 | 0.1019 | 0.1712 | 0.1532 | 0.1261 |
| | $cov_{adj}$ | (0.952) | (0.965) | (0.989) | (0.947) | (0.960) | (0.978) |
| | $\bar{se}_{ehw,adj}$ | 0.1052 | 0.1006 | 0.0939 | 0.1059 | 0.1012 | 0.0946 |

[1] See notes under Table 1.

Table 4: The Effect of Clock Stopping Policies: an Alternative Probit Specification

| | APE | standard error | |
|---|---|---|---|
| | | inf pop | finite pop |
| Panel A. Policy effects years 0-3 | | | |
| Men FOCS | -0.0137 | 0.0580 | 0.0473 |
| Women FOCS | 0.2467 | 0.1610 | 0.1234 |
| Men GNCS | 0.0451 | 0.0649 | 0.0542 |
| Women GNCS | 0.0253 | 0.1323 | 0.1132 |
| | | | |
| Panel B. Policy effects years 4+ | | | |
| Men FOCS | 0.0004 | 0.0628 | 0.0532 |
| Women FOCS | 0.0458 | 0.0972 | 0.0727 |
| Men GNCS | 0.1537 | 0.0705 | 0.0551 |
| Women GNCS | -0.2758 | 0.1137 | 0.0888 |

[1] See notes under Table 2.

[2] Estimates are obtained from the correlated random effects probit model, where cluster sizes are grouped into three bins: $M_{g1} = \mathbb{1}(M_g <= 25)$, $M_{g2} = \mathbb{1}(25 < M_g <= 35)$, and $M_{g3} = \mathbb{1}(M_g > 35)$ with $M_{g3}$ omitted as the base group.