Econometric Theory, **41**, 2025, 79–122.
doi:10.1017/S0266466623000166

ALGORITHMIC SUBSAMPLING UNDER MULTIWAY CLUSTERING

HAROLD D. CHIANG
University of Wisconsin–Madison

JIATONG LI
Vanderbilt University

YUYA SASAKI
Vanderbilt University

This paper proposes a novel method of algorithmic subsampling (data sketching) for multiway cluster-dependent data. We establish a new uniform weak law of large numbers and a new central limit theorem for multiway algorithmic subsample means. We show that algorithmic subsampling allows for robustness against potential degeneracy, and even non-Gaussian degeneracy, of the asymptotic distribution under multiway clustering at the cost of efficiency and power loss due to algorithmic subsampling. Simulation studies support this novel result, and demonstrate that inference with algorithmic subsampling entails more accuracy than that without algorithmic subsampling. We derive the consistency and the asymptotic normality for multiway algorithmic subsampling generalized method of moments estimator and for multiway algorithmic subsampling M-estimator. We illustrate with an application to scanner data for the analysis of differentiated products markets.

1. INTRODUCTION

In the present era of big data, it is not uncommon that datasets are so large that researchers may not need to use the whole sample for statistical inference to draw informative conclusions. Furthermore, computational bottlenecks in time and/or memory may even prohibit econometric analyses with such large datasets. The recent econometrics literature Lee and Ng (2020a, 2020b) suggests methods to deal with these circumstances that started to arise in today's data-rich environments. The *algorithmic subsampling* or *data sketching* explored by these authors paves

We benefited from very useful comments by Peter C. B. Phillips (the Editor), Matias Cattaneo (the Co-Editor), Sokbae (Simon) Lee, three anonymous referees, and participants in the 2021 North American Summer Meeting, International Association for Applied Econometrics, 2021 Asian Meeting, 2021 China Meeting of the Econometric Society, 26th International Panel Data Conference, 2021 Australasia Meeting of the Econometric Society, 2021 European Summer Meeting, and New York Camp Econometrics XVI. The usual disclaimer applies. We thank James M. Kilts Center, University of Chicago Booth School of Business for allowing us to use scanner data from the Dominicks Finer Foods (DFF) retail chain. H. Chiang is supported by the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin–Madison with funding from the Wisconsin Alumni Research Foundation. Address correspondence to Harold D. Chiang, Department of Economics, University of Wisconsin–Madison, Madison, WI 53706, USA; e-mail: hdchiang@wisc.edu.

the way for econometric and statistical analyses based on random subsampling of big data.

The existing study of algorithmic subsampling focuses on i.i.d. cases, and it has been "silent about how to deal with data that are dependent over time or across space" (Lee and Ng, 2020a, Sect. 8). On the other hand, some of big data may exhibit cross-sectional statistical dependence, such as multiway clustering. For instance, common scanner data (leading examples of big data) are clustered in two ways by markets and products. Common demand shocks within a market may induce statistical dependence among different products within that market. Similarly, common supply shocks by a producer may induce statistical dependence among different markets within the product produced by that producer. "A natural stochastic framework for the regression model with multiway clustered data is that of separately exchangeable random variables" (MacKinnon, Nielsen, and Webb, 2021).

In this paper, we propose a novel method of algorithmic subsampling for separately exchangeable random variables, which we will refer to as the *multiway algorithmic subsampling*, and develop asymptotic statistical properties of this method. We first establish basic theories for multiway algorithmic subsample means, namely their uniform weak law of large numbers and central limit theorem, which differ from the standard ones in a meaningful way. In particular, the form of the central limit theorem that is unique to multiway algorithmic subsampling entails a practically useful property of robustness against potential degeneracy. Researchers do not know ex ante how the data in use are affected by the cluster sampling. In cases where the cluster-specific shocks have no mean effect on the data, the standard multiway-cluster-robust asymptotic distribution *without* algorithmic subsampling would suffer from degeneracy, which can lead to either a Gaussian limiting distribution with a faster convergence rate or a non-Gaussian limiting distribution (cf. Menzel, 2021), and invalidates the statistical inference based on standard multiway cluster-robust standard errors. On the other hand, we show that multiway algorithmic subsampling allows for a nondegenerate asymptotic distribution regardless of whether the data are dependent or not. In other words, algorithmic subsampling ensures robustness against potential degeneracy, and even non-Gaussian degeneracy, of the asymptotic distribution, thereby allowing researchers to robustly enjoy valid statistical inference without knowing whether data are dependent or not. This finding about the added practical advantage of algorithmic subsampling is novel in the literature to the best of our knowledge. We emphasize that these advantages of robustness come at the cost of efficiency and power loss due to algorithmic subsampling.

Once these basic asymptotic statistical theories are established, we apply them to common econometric frameworks. Specifically, we propose a multiway algorithmic subsampling generalized method of moments (GMM) estimator, and derive asymptotic theories for it, including consistency, asymptotic normality, and consistent variance estimation. This multiway algorithmic subsampling GMM estimator also enjoys the aforementioned robustness property against potential degeneracy. Likewise, we also propose a multiway algorithmic subsampling M-estimator, and derive similar asymptotic theories for it, including the consistency, asymptotic normality, and consistent variance estimation.

1.1. Relation to the Literature

This paper intersects with two branches of the literature, namely, algorithmic subsampling and multiway clustering. In econometrics, algorithmic subsampling and its properties are first studied by Lee and Ng (2020a, 2020b). This literature has focused on random (i.i.d.) sampling as emphasized earlier. Robust variances under multiway clustering have been proposed by Cameron, Gelbach, and Miller (2011), Thompson (2011), and Cameron and Miller (2014). Asymptotic statistical properties under multiway clustering have been rigorously investigated by Chiang, Kato, and Sasaki (2021), Davezies, D'Haultfoeuille, and Guyonvarch (2021), MacKinnon et al. (2021), Menzel (2021), and Chiang et al. (2022) under various contexts. This literature has not considered algorithmic subsampling. To the best of our knowledge, this paper is the first to study the properties of algorithmic subsampling under multiway clustering, and therefore, is the first to propose the aforementioned advantage of algorithmic subsampling for robustness against potential degeneracy, including non-Gaussian degeneracy, in the asymptotic distribution under multiway clustering. We take advantage of the asymptotic distributional theory for incomplete one-sample U-statistics (Janson, 1984) to develop parts of our basic theoretical results. The method of algorithmic subsampling is also closely related to the general scheme of resampling methods for clustered data, which has been studied for one or multiway clustering by MacKinnon and Webb (2017), MacKinnon and Webb (2018), Djogbenou, MacKinnon, and Nielsen (2019), Chiang et al. (2021), Davezies et al. (2021), MacKinnon et al. (2021), and Menzel (2021), to name but a few.

1.2. Organization

The rest of this paper is organized as follows: Section 2 introduces multiway algorithmic subsampling and presents its asymptotic statistical theories. Sections 3 and 4 demonstrate applications to the GMM and M-estimation frameworks, respectively. Section 5 presents Monte Carlo simulation studies. Section 6 presents an empirical application to scanner data. Section 7 concludes. All mathematical proofs and details are collected in the Appendix.

2. MULTIWAY ALGORITHMIC SUBSAMPLING

Suppose a researcher observes data $\{W_{ij} : 1 \leq i \leq N, 1 \leq j \leq M\}$, where N and M are the sample sizes in the first and second cluster dimensions. For instance, N and M are the number of markets and the number of products, respectively, in scanner data. The data may be two-way clustered, in the sense that we allow for arbitrary statistical dependence of W_{ij} across $j \in \{1,\ldots,M\}$ within each market i (due to a common demand shock) and also allow for arbitrary statistical dependence of W_{ij} across $i \in \{1,\ldots,N\}$ within each product j (due to a common supply shock). A formal assumption of this sampling process will be stated as Assumption 1. Using such two-way cluster sampled data, we are interested in (uniformly) consistent estimation of and statistical inference about $E[f(W_{ij})]$ based on standard econometric techniques. Since scanner data are very big, however, computational bottlenecks in time or memory may limit or even prohibit implementation of standard econometric analysis. Lee and Ng (2020a, 2020b) therefore suggest algorithmic subsampling of big data to alleviate computational burdens.

Adapting the ideas of Lee and Ng (2020a, 2020b) to our framework of two-way clustered data, we propose the following multiway algorithmic subsampling procedure. (We remark that algorithmic subsampling is different from subsampling as a resampling method.) Let p_{NM} denote the probability of subsample selection that may depend on the current sample size (N,M). Generate i.i.d Bernoulli(p_{NM}) random variables $\{Z_{ij} : 1 \le i \le N, 1 \le j \le M\}$ independently from data. Let $\widehat{L} = \sum_{i=1}^{N}\sum_{j=1}^{M} Z_{ij}$ denote the number of nonzero elements. Note that $\widehat{L}$ follows Binomial(NM, p_{NM}), and thus $L \equiv E[\widehat{L}] = NMp_{NM}$ in particular. In fact, this formulation of algorithmic subsampling is called the Bernoulli subsampling, and is one of the alternative approaches to subsampling proposed by Lee and Ng (2020a). We focus on Bernoulli subsampling in this paper for simplicity as well as its desired property of the aforementioned robustness against degeneracy. That said, we remark that it is also feasible to use alternative subsampling methods (namely, the uniform subsampling with and without replacement) proposed by Lee and Ng (2020a).

We use $\widehat{L}^{-1}\sum_{i=1}^{N}\sum_{j=1}^{M} Z_{ij} f\left(W_{ij}\right)$ to estimate and make inference about $E[f(W_{ij})]$. To this end, we first develop the uniform weak law of large numbers and the central limit theorem under this setting of multiway algorithmic subsampling in Sections 2.1 and 2.2, respectively. We then apply these basic theories in turn to establish the consistency and the asymptotic normality for the GMM and M-estimation in Sections 3 and 4, respectively. Hereafter, for conciseness of notations, p_{NM} will be abbreviated as p. We let $[k]$ denote the set $\{1,\ldots,k\}$ for any $k \in \mathbb{N}$, and let $[k]^c = \mathbb{N}\backslash[k]$ for any $k \in \mathbb{N}$. We use the short-hand notation $\underline{C} = \min\{N,M\}$. Throughout the paper, the asymptotics is understood as $\underline{C} \to \infty$. For a vector $v \in \mathbb{R}^k$, let $\|v\|$ denote the Euclidean norm of v.

2.1. The Uniform Weak Law of Large Numbers

We first formally state the assumption of two-way cluster sampling.

Assumption 1 (Sampling). (i) $\left(W_{ij}\right)_{(i,j)\in\mathbb{N}^2}$ is an infinite sequence of separately exchangeable d-dimensional random vectors. That is, for any permutations π_1 and π_2 of $\mathbb{N}$, we have $\left(W_{ij}\right)_{(i,j)\in\mathbb{N}^2} \overset{\mathrm{d}}{=} \left(W_{\pi_1(i)\pi_2(j)}\right)_{(i,j)\in\mathbb{N}^2}$. (ii) $\left(W_{ij}\right)_{(i,j)\in\mathbb{N}^2}$ is dissociated. That is, for any $(c_1,c_2) \in \mathbb{N}^2$, $\left(W_{ij}\right)_{i\in[c_1],j\in[c_2]}$ is independent of $\left(W_{ij}\right)_{i\in[c_1]^c,j\in[c_2]^c}$.

Part (i) requires a form of the identical distribution condition in separate permutations of the i index and the j index. Although we relax the independent sampling, we maintain a form of identical distribution as such. Part (ii) requires that sets of observations are independent if they do not share the same i index or the same j index, i.e., $\left(W_{ij} : i \in \{1,\ldots,c_1\}, j \in \{1,\ldots,c_2\}\right)$ and $\left(W_{ij} : i \in \{c_1+1,\ldots\}, j \in \{c_2+1,\ldots\}\right)$ are assumed to be independent for any $(c_1,c_2) \in \mathbb{N}^2$. However, where any observations are sharing either the same i index or the same j index, then they are allowed to be arbitrarily dependent. For example, in the scanner data, two observations in the same market i may be dependent due to a common demand shock, and likewise two observations in the same product j may also be dependent due to a common supply shock.

Assumption 1 consists of a sufficient condition for what we actually need. These conditions can be relaxed to the assumption that the data $(W_{ij})_{(i,j)\in\mathbb{N}^2}$ are generated via the process $W_{ij} = f(\alpha_i, \beta_j, \varepsilon_{ij})$ for some Borel-measurable function f, where $(\alpha_i)_{i\in\mathbb{N}}$, $(\beta_j)_{j\in\mathbb{N}}$, and $(\varepsilon_{ij})_{(i,j)\in\mathbb{N}^2}$ are mutually independent, and each of $(\alpha_i)_{i\in\mathbb{N}}$, $(\beta_j)_{j\in\mathbb{N}}$, and $(\varepsilon_{ij})_{(i,j)\in\mathbb{N}^2}$ is i.i.d. This data generating process, or so-called Aldous–Hoover–Kallenberg representation, is implied by Assumption 1. We can interpret α_i and β_j as i- and j-specific effects, respectively, while ε_{ij} is an idiosyncratic effect. This representation is also consistent with the data generating processes considered in the simulation studies.

For convenience of stating the next assumption, we introduce additional notations and definitions. Let (T,d) be pseudometric space.[1] For $\varepsilon > 0$, an ε-net of T is a subset T_ε of T such that for every $t \in T$ there exists $t_\varepsilon \in T_\varepsilon$ with $d(t,t_\varepsilon) \leq \varepsilon$. We define the ε-covering number $N(T,d,\varepsilon)$ of T by

$$N(T,d,\varepsilon) = \inf\{\mathrm{Card}(T_\varepsilon) : T_\varepsilon \text{ is an } \varepsilon\text{-net of } T\}.$$

For any probability measure Q on a measurable space $(S,\mathcal{S})$ and any $q \geq 1$, define $\|f\|_{Q,q} = \left\{\int |f|^q dQ\right\}^{1/q}$, and let $L^q(S) = \{f : S \to \mathbb{R} : \|f\|_{Q,q} < \infty\}$. A function $G : S \to \mathbb{R}$ is an envelope of a class of functions $\mathcal{G} \ni g, g : S \to \mathbb{R}$, if $\sup_{g\in\mathcal{G}} |g(s)| \leq G(s)$ for all $s \in S$. With these notations and definitions, we state the following assumption regarding the function class where f resides.

Assumption 2 (Function class). The function class $\mathcal{F}$ satisfies (i) $E\left[f\left(W_{ij}\right)\right] = 0$ for all $f \in \mathcal{F}$. (ii) $\mathcal{F}$ admits an envelope F satisfying $E\left[F(W_{ij})\right] < \infty$ with $\sup_Q N(\mathcal{F}, \|\cdot\|_{Q,2}, \epsilon\|F\|_{Q,2}) < \infty$ for all $\epsilon > 0$, where Q is any finite discrete measure. (iii) $\mathcal{F}$ is pointwise measurable.[2]

Part (i) is a location normalization (centering) and is therefore without loss of generality. Although this part will not be needed in the short run (Lemma 1), we state it here as this Assumption 2 collects requirements about the function space where f resides. Part (ii) is a regularity condition imposed to establish a uniform weak law of large numbers. Part (iii) is a technical requirement that is used to avoid

[1] That is, $d(x,y) = 0$ does not imply $x = y$.

[2] For its definition, see van der Vaart and Wellner (1996, p. 110) for instance.

measurability issues. At this moment, we are stating these high-level assumptions for the sake of generality, but we will provide the standard lower-level primitive conditions in the context of the application to the GMM presented in Section 3 and the application to the M-estimation presented in Section 4.

Under these assumptions, we can establish the uniform weak law of large numbers for multiway algorithmic subsample means as formally stated in the lemma below.

LEMMA 1 (Uniform weak law of large numbers). *Suppose that Assumption 1 holds and that $\mathcal{F}$ satisfies Assumption 2 (ii), (iii). Then, for any $f \in \mathcal{F}$, we have*

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{\widehat{L}}\sum_{i=1}^{N}\sum_{j=1}^{M}Z_{ij}f\left(W_{ij}\right)-E\left[f(W_{11})\right]\right|\xrightarrow{P}0.$$

This result is not very surprising, but we state above as Lemma 1 and prove it (in Appendix C.1) for the following two reasons. First, this is the first time it is stated and proved in the literature to the best of our knowledge. Second, more importantly, this lemma serves as a useful auxiliary device for other results to be presented below that are practically more relevant.

2.2. The Central Limit Theorem

To establish the central limit theorem under multiway algorithmic subsampling, we augment our assumptions with the following additional condition. Recall the short-hand notation $\underline{C} = \min\{N,M\}$.

Assumption 3. There exists a constant $\Lambda \geq 0$ such that $(\underline{C}/NM)((1-p_{NM})/p_{NM}) \to \Lambda$.

It entails that there exist constants $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ such that $\underline{C}/N \to \lambda_1$, $\underline{C}/M \to \lambda_2$. To facilitate the subsequent discussions, we introduce a notion of degenerate asymptotic distribution. For any scalar-valued sequence of random variables $\left(X_{ij}\right)_{(i,j)\in\mathbb{N}^2}$ satisfying Assumption 1, we say the asymptotic distribution is degenerate if $\mathrm{Var}\left((\sqrt{\underline{C}}/NM)\sum_{i=1}^{N}\sum_{j=1}^{M}X_{ij}\right) \to 0$ as $\underline{C} \to \infty$. The following theorem establishes the central limit theorem under multiway algorithmic subsampling.

THEOREM 1 (Central limit theorem). *Suppose that Assumptions 1, 2(i), (iii), and 3 hold, and that the class $\mathcal{F} = \{f_1,\ldots,f_k\}$ is finite, independent of sample size, and admits an envelope F satisfying $E\left[F(W_{ij})^2\right] < \infty$. Let $f = (f_1,\ldots,f_k)^T$. Then,*

$$\sqrt{\underline{C}}\frac{1}{\widehat{L}}\sum_{i=1}^{N}\sum_{j=1}^{M}Z_{ij}f\left(W_{ij}\right)\xrightarrow{d}N\left(0,\Gamma\right),$$

where the variance is given by $\Gamma = \Gamma_A + \Lambda\Gamma_B$, $\Gamma_A = \lambda_1 E\left[f(W_{11})f^T(W_{12})\right] + \lambda_2 E\left[f(W_{11})f^T(W_{21})\right]$ *and* $\Gamma_B = E\left[f(W_{11})f^T(W_{11})\right]$.

A proof is provided in Appendix C.2. This central limit theorem entails a novel and useful feature of algorithmic subsampling for multiway clustered data in practice. Notably, the asymptotic variance Γ consists of a sum of two components, Γ_A and $\Lambda\Gamma_B$. This is in contrast with algorithmic subsampling for independent data, where the asymptotic variance consists of only one term. The first part, Γ_A, in fact coincides with the asymptotic variance that we would get without algorithmic subsampling. Specifically, $\Gamma = \Gamma_A$ if $p = 1$. More generally, if p is a constant, as the sample sizes (N, M) increase, then $\Lambda = 0$ so that $\Gamma = \Gamma_A$. On the other hand, if p is chosen so that $\Lambda = \lim_{N,M\to\infty}(\underline{C}/NM)((1-p)/p) > 0$, then the second part, $\Lambda\Gamma_B$, is also present. Furthermore, Γ_B is nonzero whenever the distribution of $f(W_{ij})$ is nondegenerate, and this feature ensures robustness in inference against possible events of no cross-sectional dependence.

In practice, a researcher may not ex ante know whether data exhibit cross-sectional dependence ($E[f(W_{11})f^T(W_{12})] \neq 0$ or $E[f(W_{11})f^T(W_{21})] \neq 0$) or not. If a researcher knew the true dependence structure, he or she could set the correct cluster dimension to conduct valid inference. However, this premise is implausible. In cases where there is no cross-sectional dependence, then $\Gamma_A = 0$ and the statistical inference based on the asymptotic normality *without* algorithmic subsampling would suffer from the degeneracy problem. Because of algorithmic subsampling, however, we can robustly safeguard against such degenerate asymptotic distributions without requiring a prior knowledge of the researcher about the presence/absence of cross-sectional dependence in data. This result is novel in the literature, and also uncovers an additional useful property of algorithmic subsampling in practice.[3] Simulation studies presented in Section 5 support this practically relevant property of multiway algorithmic subsampling.

Intuitively, algorithmic subsampling with smaller p makes it less likely that multiple observations from the same row i or the same column j are selected. Thus, it results in placing relatively more weights on the variance $E[f(W_{ij})f(W_{ij})]$ than on the covariances, $E[f(W_{ij})f(W_{ij'})]$ and $E[f(W_{ij})f(W_{i'j})]$, in the asymptotic distribution. Hence, the part $E[f(W_{ij})f(W_{ij})]$ of the asymptotic variance becomes dominant in the case of degenerate covariances, and this feature of algorithmic subsampling prevents the degeneracy problem.

We can apply these theoretical results to a number of common frameworks of econometric analysis. Two of the most frequently used classes of econometric methods are the GMM and the M-estimation. Therefore, we will demonstrate applications of these basic theories of the uniform weak law of large numbers

[3]Indeed, the method of inference by MacKinnon et al. (2021) as well as Cameron et al. (2011) adapts to specific classes of degenerate asymptotic distributions. However, these are restricted to the cases of Gaussian degeneracy, where the convergence rate is $\sqrt{NM}$, yet the asymptotic distribution is still Gaussian. On the other hand, these existing methods of inference by Cameron et al. (2011) and MacKinnon et al. (2021) do not adapt to the class of non-Gaussian degenerate asymptotic distributions.

(Lemma 1) and the central limit theorem (Theorem 1) to establish the consistency and the asymptotic normality of the GMM and M-estimators under multiway algorithmic subsampling in Sections 3 and 4.

We conclude this section with a remark on alternative subsampling methods. As mentioned earlier, our method is based on the Bernoulli subsampling, and is one of the alternative approaches to subsampling proposed by Lee and Ng (2020a). Besides the Bernoulli subsampling on which we focus in this paper, they propose uniform subsampling with replacement, uniform subsampling without replacement, and leverage score subsampling. Among these alternative methods, it is also feasible to use uniform subsampling with replacement and uniform subsampling without replacement. Similar asymptotic properties will follow through similar lines of the argument following Janson (1984) to those in the proof of Theorem 1. See Appendix H for details.

2.3. Application to the Ordinary Least Squares

This section demonstrates an application of the basic theories to the ordinary least squares (OLS) estimator. Consider the linear regression model

$$Y_{ij} = X_{ij}^T\beta + u_{ij} \qquad 1 \le i \le N, 1 \le j \le M, \tag{2.1}$$

where Y_{ij} is a response variable, X_{ij} is a vector of d covariates, and u_{ij} is an error satisfying $E\left[u_{ij}|X_{ij}\right] = 0$. Let $W_{ij} = \left(Y_{ij}, X_{ij}^T\right)^T$, and we apply the proposed multiway algorithmic subsampling to W_{ij}. The parameter of interest is the vector of linear projection coefficients

$$\beta = E\left[X_{11}X_{11}^T\right]^{-1} E\left[X_{11}Y_{11}\right], \tag{2.2}$$

and multiway algorithmic subsampling OLS estimator is

$$\widehat{\beta} = \left(\frac{1}{\widehat{L}}\sum_{i=1}^{N}\sum_{j=1}^{M} Z_{ij}X_{ij}X_{ij}^T\right)^{-1}\left(\frac{1}{\widehat{L}}\sum_{i=1}^{N}\sum_{j=1}^{M} Z_{ij}X_{ij}Y_{ij}\right). \tag{2.3}$$

Applications of Lemma 1 and Theorem 1 yield the following limit distribution property about $\widehat{\beta}$.

COROLLARY 1. *Suppose that Assumption 1 holds for* $W_{ij} = (Y_{ij}, X_{ij}^T)^T$. *Assume* $E\left[|Y_{11}|^4\right] < \infty$, $E\left[||X_{11}||^4\right] < \infty$, *and that* $E\left[X_{11}X_{11}^T\right]$ *is nonsingular. For* β *and* $\widehat{\beta}$ *defined in (2.2) and (2.3), we have*

$$\sqrt{\underline{C}}\left(\widehat{\beta} - \beta\right) \xrightarrow{d} N(0, V),$$

where $V = J^{-1}\Gamma_{OLS}J^{-1}$, $J = E\left[X_{11}X_{11}^T\right]$, $\Gamma_{OLS} = \Gamma_{OLS,1} + \Lambda\Gamma_{OLS,2}$, $\Gamma_{OLS,1} = \lambda_1 E[X_{11}u_{11}(X_{12}u_{12})^T] + \lambda_2 E[X_{11}u_{11}(X_{21}u_{21})^T]$, *and* $\Gamma_{OLS,2} = E[X_{11}u_{11}(X_{11}u_{11})^T]$.

See Appendix C.3 for proof of this corollary.

3. APPLICATION TO GENERALIZED METHOD OF MOMENTS

In this section, we apply the basic methods and theories presented in Section 2 to multiway algorithmic subsampling GMM. Suppose that an economic model implies moment restrictions $E\left[g\left(W_{ij},\theta^0\right)\right]=0$ for a true parameter vector $\theta^0=(\theta_1^0,\dots,\theta_k^0)^T\in\Theta$, $\Theta\subset\mathbb{R}^k$, where $g=(g_1,\dots,g_m)^T$, and $m\geq k$. With a Bernoulli sample $\{Z_{ij}:1\leq i\leq N,1\leq j\leq M\}$, algorithmic subsample moment evaluated at $\theta=(\theta_1,\dots,\theta_k)^T\in\Theta$ is given by $\widehat{g}_{NM}(\theta)=\widehat{L}^{-1}\sum_{i=1}^N\sum_{j=1}^M Z_{ij}g\left(W_{ij},\theta\right)$. Let $\widehat{V}$ be a positive semi-definite random matrix, which may depend on θ. We define multiway algorithmic subsampling GMM estimator $\widehat{\theta}$ as the solution to

$$\max_{\theta\in\Theta}\widehat{Q}_{NM}(\theta),$$

where $\widehat{Q}_{NM}(\theta)=-\widehat{g}_{NM}(\theta)^T\widehat{V}\widehat{g}_{NM}(\theta)$. The true parameter vector $\theta^0\in\Theta$ is assumed to uniquely solve the population problem $\max_{\theta\in\Theta}-E[g(W_{ij},\theta)]^T VE[g(W_{ij},\theta)]$, where V is positive semi-definite and $\widehat{V}\xrightarrow{P}V$.

3.1. Consistency and Asymptotic Normality

To establish the consistency and asymptotic normality for multiway algorithmic subsampling GMM estimator $\widehat{\theta}$, we make the following assumption. To concisely state the following assumption, we introduce one additional definition regarding Lipschitz continuity. A function $g:\mathbb{R}^k\to\mathbb{R}$, is Lipschitz with a universal Lipschitz constant, if there exists a positive constant M such that $\left|g(w,\theta)-g\left(w,\theta'\right)\right|\leq M\left\|\theta-\theta'\right\|$ for all $w\in\operatorname{supp}(W_{ij})$.

Assumption 4.

(i) V is positive semi-definite, and $VE[g(W_{ij},\theta)]=0$ only if $\theta=\theta^0$.
(ii) $\theta^0\in\operatorname{int}(\Theta)$, where Θ is a compact subset of $\mathbb{R}^k$.
(iii) (a) $\theta\mapsto g_r(w,\theta)$ is Lipschitz with a universal Lipschitz constant.
  (b) Each coordinate of $\theta\mapsto\nabla_\theta g_r(w,\theta)$ is Lipschitz with a universal Lipschitz constant.
(iv) $E\left[\sup_{\theta\in\Theta}\left\|g\left(W_{ij},\theta\right)\right\|\right]<\infty$.
(v) G^TVG is nonsingular, where $G=E\left[\nabla_\theta g\left(W_{ij},\theta^0\right)\right]$.
(vi) $E\left[\sup_{\theta\in\Theta}\left\|\nabla_\theta g\left(W_{ij},\theta\right)\right\|\right]<\infty$.
(vii) $g_{\sup}(\cdot)=\max_{r\in\{1,\dots,m\}}|g_r(\cdot,\theta)|$ satisfies $E[g_{\sup}(W_{ij})^2]<\infty$.

Assumption 4 is analogous to the conditions required for Theorems 2.6 and 3.4 in Newey and McFadden (1994), which state the consistency and asymptotic normality, respectively, of the GMM estimator under the conventional random sampling.

We first state the consistency of the multiway algorithmic subsampling GMM estimator.

LEMMA 2 (Consistency of multiway algorithmic subsampling GMM estimator). *If Assumptions 1 and 4(i)–(iv) hold and that* $\widehat{V} \xrightarrow{P} V$, *then* $\widehat{\theta} \xrightarrow{P} \theta^0$.

A proof is provided in Appendix E.1. It follows from combining the arguments in the proofs of Newey and McFadden (1994, Thm. 2.1) with our uniform weak law of large numbers for multiway algorithmic subsampling (Lemma 1) presented in Section 2.1.

We next state the asymptotic normality of the multiway algorithmic subsampling GMM estimator.

THEOREM 2 (Asymptotic normality of multiway algorithmic subsampling GMM estimator). *If Assumptions 1, 3, and 4 hold, and that* $\widehat{V} \xrightarrow{P} V$, *then*

$$\sqrt{\underline{C}}\left(\widehat{\theta}-\theta^0\right) \xrightarrow{d} N\left(0, \left(G^T VG\right)^{-1} G^T V\Omega VG\left(G^T VG\right)^{-1}\right),$$

where $G = E\left[\nabla_\theta g\left(W_{11},\theta^0\right)\right]$ *and* $\Omega = \Gamma_1 + \Lambda\Gamma_2$, *with* $\Gamma_1 = \lambda_1 E[g(W_{11},\theta^0)g^T(W_{12},\theta^0)] + \lambda_2 E\left[g\left(W_{11},\theta^0\right)g^T\left(W_{21},\theta^0\right)\right]$ *and* $\Gamma_2 = E\left[g\left(W_{11},\theta^0\right)g^T\left(W_{11},\theta^0\right)\right]$.

A proof is provided in Appendix E.2. It follows from combining the arguments in the proofs of Newey and McFadden (1994, Thm. 3.4) with our central limit theorem for multiway algorithmic subsampling (Theorem 1) presented in Section 2.2.

3.2. Algorithmic Subsampling Variance Estimation

The components, G and Ω in Theorem 2, of the asymptotic variance of multiway algorithmic subsampling GMM estimator can be estimated by

$$\widetilde{G} = \frac{1}{\widehat{L}}\sum_{i=1}^{N}\sum_{j=1}^{M} Z_{ij}\nabla_\theta g\left(W_{ij},\widehat{\theta}\right)$$

and

$$\widetilde{\Omega} = \widetilde{\Gamma}_1 + \Lambda\widetilde{\Gamma}_2,$$

respectively, where

$$\widetilde{\Gamma}_1 = \frac{\underline{C}}{\widehat{L}^2}\sum_{i=1}^{N}\sum_{1\le j,j'\le M} Z_{ij}Z_{ij'}\, g\left(W_{ij},\widehat{\theta}\right)g^T\left(W_{ij'},\widehat{\theta}\right)$$
$$+\frac{\underline{C}}{\widehat{L}^2}\sum_{1\le i,i'\le N}\sum_{j=1}^{M} Z_{ij}Z_{i'j}\, g\left(W_{ij},\widehat{\theta}\right)g^T\left(W_{i'j},\widehat{\theta}\right)$$

and

$$\widetilde{\Gamma}_2 = \frac{1}{\widehat{L}} \sum_{i=1}^{N} \sum_{j=1}^{M} Z_{ij} g\left(W_{ij}, \widehat{\theta}\right) g^T\left(W_{ij}, \widehat{\theta}\right).$$

We propose to estimate the asymptotic variance $\left(G^T VG\right)^{-1} G^T V\Omega VG\left(G^T VG\right)^{-1}$ by the sample counterpart $\left(\widetilde{G}^T\widehat{V}\widetilde{G}\right)^{-1}\widetilde{G}^T\widehat{V}\widetilde{\Omega}\widehat{V}\widetilde{G}\left(\widetilde{G}^T\widehat{V}\widetilde{G}\right)^{-1}$. To guarantee that this algorithmic subsampling variance estimator works asymptotically, we make the following assumption in addition.

Assumption 5.

(i) $\theta \mapsto E\left[\nabla_\theta g\left(W_{ij}, \theta\right)\right]$ is continuous at θ^0.
(ii) $\theta \mapsto \lambda_1 E\left[g\left(W_{ij}, \theta\right) g^T\left(W_{ij}, \theta\right)\right] + \lambda_2 E\left[g\left(W_{ij}, \theta\right) g^T\left(W_{ij}, \theta\right)\right]$ is continuous at θ^0.
(iii) $\theta \mapsto E\left[g\left(W_{ij}, \theta\right) g^T\left(W_{ij}, \theta\right)\right]$ is continuous at θ^0.

With this additional assumption, $\left(\widetilde{G}^T\widehat{V}\widetilde{G}\right)^{-1}\widetilde{G}^T\widehat{V}\widetilde{\Omega}\widehat{V}\widetilde{G}\left(\widetilde{G}^T\widehat{V}\widetilde{G}\right)^{-1}$ is consistent for the asymptotic variance $\left(G^T VG\right)^{-1} G^T V\Omega VG\left(G^T VG\right)^{-1}$, as formally stated in the following theorem.

THEOREM 3 (Consistent asymptotic variance estimation of multiway algorithmic subsampling GMM estimator). *If Assumptions 1 and 3–5 hold and that $\widehat{V} \xrightarrow{P} V$, then*

$$\left(\widetilde{G}^T\widehat{V}\widetilde{G}\right)^{-1}\widetilde{G}^T\widehat{V}\widetilde{\Omega}\widehat{V}\widetilde{G}\left(\widetilde{G}^T\widehat{V}\widetilde{G}\right)^{-1} \xrightarrow{P} \left(G^T VG\right)^{-1} G^T V\Omega VG\left(G^T VG\right)^{-1}.$$

A proof is provided in Appendix E.3. It follows by combining Lemma 1 and similar lines of arguments to those in the proofs of Lemma 2 and Theorem 2.

4. APPLICATION TO M-ESTIMATION

In this section, we apply the basic methods and theories presented in Section 2 to multiway algorithmic subsampling M-estimation. Let $\Theta \subset \mathbb{R}^k$ be a parameter space and define the class $\mathcal{Q} = \{q(\cdot, \theta) : \theta \in \Theta\}$ of functions $q(\cdot, \theta)$ indexed by θ. With a Bernoulli sample $\{Z_{ij}, 1 \le i \le N, 1 \le j \le M\}$, we define multiway algorithmic subsampling M-estimator $\widehat{\theta}$ as the solution to

$$\max_{\theta \in \Theta} -\frac{1}{\widehat{L}} \sum_{i=1}^{N} \sum_{j=1}^{M} Z_{ij} q\left(W_{ij}, \theta\right).$$

The true parameter vector $\theta^0 = (\theta_1^0, \ldots, \theta_k^0)^T \in \Theta$ is assumed to uniquely solve the population maximization problem $\max_{\theta \in \Theta} -E\left[q\left(W_{ij}, \theta\right)\right]$, in the sense that $E\left[q\left(W_{ij}, \theta^0\right)\right] < E\left[q\left(W_{ij}, \theta\right)\right]$ holds for all $\theta = (\theta_1, \ldots, \theta_k)^T \in \Theta$ and $\theta \neq \theta^0$. For each $\theta \in \Theta$, let $-\widehat{L}^{-1}\sum_{i=1}^{N}\sum_{j=1}^{M} Z_{ij} q\left(W_{ij},\theta\right)$ and $-E\left[q\left(W_{ij},\theta\right)\right]$ be denoted by $\widehat{Q}_{NM}(\theta)$ and $Q_0(\theta)$, respectively, for conciseness.

4.1. Consistency and Asymptotic Normality

To establish the consistency and asymptotic normality for multiway algorithmic subsampling M-estimator $\widehat{\theta}$, we make the following assumption.

Assumption 6.

(i) $\theta^0 \in \operatorname{int}(\Theta)$, where Θ is a compact subset of $\mathbb{R}^k$, and $E[q(W_{ij},\theta^0)] < E[q(W_{ij},\theta)]$ for all $\theta \in \Theta \backslash \{\theta_0\}$.

(ii) (a) $\theta \mapsto q(w,\theta)$ is Lipschitz with a universal Lipschitz constant.

(b) Each coordinate of $\theta \mapsto \nabla_\theta q(w,\theta)$ is Lipschitz with a universal Lipschitz constant.

(c) Each coordinate of $\theta \mapsto \nabla_{\theta\theta^T} q(w,\theta) = \partial^2 q(w,\theta)/\partial\theta\partial\theta^T$ is Lipschitz with a universal Lipschitz constant.

(iii) $E[\sup_{\theta\in\Theta} q\left(W_{ij},\theta\right)] < \infty$.

(iv) $E\left[\sup_{\theta\in\Theta}\left\|\nabla_{\theta\theta^T} q\left(W_{ij},\theta\right)\right\|\right] < \infty$.

(v) $H = H\left(\theta^0\right)$ is nonsingular, where $H(\theta) = -E\left[\nabla_{\theta\theta^T} q\left(W_{ij},\theta\right)\right]$.

(vi) $\dot{q}_{\sup}(\cdot) = \max_{r\in\{1,\ldots,k\}} |\partial q(\cdot,\theta)/\partial\theta_r|$ satisfies $E[\dot{q}_{\sup}(W_{ij})^2] < \infty$.

Assumption 6 is analogous to the conditions required for Theorems 2.1 and 3.1 in Newey and McFadden (1994), which state the consistency and asymptotic normality, respectively, of the M-estimator under the conventional random sampling.

We first state the consistency of multiway algorithmic subsampling M-estimator.

LEMMA 3 (Consistency of multiway algorithmic subsampling M-estimator). *If Assumptions 1 and 6(i)–(iii) hold, then* $\widehat{\theta} \xrightarrow{P} \theta^0$.

A proof is provided in Appendix F.1. It follows from combining the arguments in the proof of Newey and McFadden (1994, Thm. 2.1) with our uniform weak law of large numbers for multiway algorithmic subsampling (Lemma 1) presented in Section 2.1.

We next state the asymptotic normality of multiway algorithmic subsampling M-estimator.

THEOREM 4 (Asymptotic normality of multiway algorithmic subsampling M-estimator). *If Assumptions 1, 3, and 6 hold, then*

$$\sqrt{\underline{C}}\left(\widehat{\theta}-\theta^0\right) \xrightarrow{d} N\left(0, H^{-1}\Sigma H^{-1}\right),$$

where $H = -E[\nabla_{\theta\theta^T} q(W_{11},\theta^0)]$, $\Sigma = \Sigma_1 + \Lambda\Sigma_2$, $\Sigma_1 = \lambda_1 E[\nabla_\theta q(W_{11},\theta^0)\nabla_\theta q(W_{12},\theta^0)^T] + \lambda_2 E[\nabla_\theta q(W_{11},\theta^0)\nabla_\theta q(W_{21},\theta^0)^T]$ *and* $\Sigma_2 = E[\nabla_\theta q(W_{11},\theta^0)\nabla_\theta q(W_{11},\theta^0)^T]$.

A proof is provided in Appendix F.2. It follows from combining the arguments in the proof of Newey and McFadden (1994, Thm. 3.1) with our central limit theorem for multiway algorithmic subsampling (Theorem 1) presented in Section 2.2.

4.2. Algorithmic Subsampling Variance Estimation

The components, H and Σ in Theorem 4, of the asymptotic variance of multiway algorithmic subsampling M-estimator can be estimated by

$$\widetilde{H} = -\frac{1}{\widehat{L}} \sum_{i=1}^{N} \sum_{j=1}^{M} Z_{ij} \nabla_{\theta\theta^T} q\left(W_{ij}, \widehat{\theta}\right)$$

and $\widetilde{\Sigma} = \widetilde{\Sigma}_1 + \Lambda \widetilde{\Sigma}_2$, respectively, where

$$\begin{aligned}\widetilde{\Sigma}_1 =& \frac{C}{\widehat{L}^2} \sum_{i=1}^{N} \sum_{1 \leq j, j' \leq M} Z_{ij} Z_{ij'} \nabla_\theta q\left(W_{ij}, \widehat{\theta}\right) \nabla_\theta q\left(W_{ij'}, \widehat{\theta}\right)^T \\ &+ \frac{C}{\widehat{L}^2} \sum_{1 \leq i, i' \leq N} \sum_{j=1}^{M} Z_{ij} Z_{i'j} \nabla_\theta q\left(W_{ij}, \widehat{\theta}\right) \nabla_\theta q\left(W_{i'j}, \widehat{\theta}\right)^T\end{aligned}$$

and

$$\widetilde{\Sigma}_2 = \frac{1}{\widehat{L}} \sum_{i=1}^{N} \sum_{j=1}^{M} Z_{ij} \nabla_\theta q\left(W_{ij}, \widehat{\theta}\right) \nabla_\theta q\left(W_{ij}, \widehat{\theta}\right)^T.$$

Thus, we propose to estimate $H^{-1} \Sigma H^{-1}$ by the sample counterpart $\widetilde{H}^{-1} \widetilde{\Sigma} \widetilde{H}^{-1}$. To guarantee that this asymptotic variance estimator works, we use the following assumption in addition.

Assumption 7.

(i) $\theta \mapsto E\left[\nabla_{\theta\theta^T} q\left(W_{ij}, \theta\right)\right]$ is continuous at θ^0.

(ii) $\theta \mapsto \lambda_1 E\left[\nabla_\theta q\left(W_{ij}, \theta\right) \nabla_\theta q\left(W_{ij}, \theta\right)^T\right] + \lambda_2 E\left[\nabla_\theta q\left(W_{ij}, \theta\right) \nabla_\theta q\left(W_{ij}, \theta\right)^T\right]$ is continuous at θ^0.

(iii) $\theta \mapsto E\left[\nabla_\theta q\left(W_{ij}, \theta\right) \nabla_\theta q\left(W_{ij}, \theta\right)^T\right]$ is continuous at θ^0.

With this additional assumption, $\widetilde{H}^{-1} \widetilde{\Sigma} \widetilde{H}^{-1}$ is consistent for the asymptotic variance $H^{-1} \Sigma H^{-1}$, as formally stated in the following theorem.

THEOREM 5 (Consistency of the asymptotic variance of multiway algorithmic subsampling M-estimator). *If Assumptions 1, 3, 6, and 7 hold, then $\widetilde{H}^{-1} \widetilde{\Sigma} \widetilde{H}^{-1}$ is consistent for $H^{-1} \Sigma H^{-1}$.*

A proof is provided in Appendix F.3. It follows by combining Lemma 1 and similar lines of arguments to those in the proofs of Lemma 3 and Theorem 4.

5. SIMULATION STUDIES

As emphasized in Section 2, we discovered a new advantage of algorithmic subsampling that it allows for robustness in inference against potential degeneracy of the asymptotic distribution under multiway clustering. In this section, we use Monte Carlo simulations to demonstrate this robustness property. Following Menzel (2021), we consider two broad categories of designs, namely, additively separable designs (Section 5.1) and nonseparable designs (Section 5.2). For each of these two broad categories, we experiment with a design that leads to a non-degenerate asymptotic distribution and another design that leads to a degenerate asymptotic distribution if algorithmic subsampling were not to be employed. In total, we consider four designs. Multiway algorithmic subsampling will be shown to yield more accurate finite sample coverage results than conventional methods robustly across all four cases, thereby supporting the aforementioned theoretical discovery by this paper.

5.1. Additively Separable Designs

First, we generate the two-way clustered array $\{Y_{ij}\}_{i\in[N],j\in[M]}$ according to the additively separable model

$$Y_{ij} = \sigma_a\alpha_i + \sigma_b\beta_j + \sigma_e\varepsilon_{ij},$$

where β_j and ε_{ij} are i.i.d. standard normal, and $\alpha_i = (\zeta_i - \mu_\zeta)/\sigma_\zeta$ for $\log(\zeta_i) \overset{\text{i.i.d.}}{\sim} N(0,1)$, $\mu_\zeta = E[\zeta_i]$, and $\sigma_\zeta^2 = \mathrm{Var}(\zeta_i)$. With this basic setup, we consider two designs:

$$\text{Design 1}: \sigma_a^2 = 0.5, \sigma_b^2 = 0.1, \text{ and } \sigma_e^2 = 0.2;$$
$$\text{Design 2}: \sigma_a^2 = 0.0, \sigma_b^2 = 0.0, \text{ and } \sigma_e^2 = 0.2.$$

Note that Design 2, without i-specific randomness or j-specific randomness, would lead to a degenerate asymptotic distribution if algorithmic subsampling were not employed.

Table 1 reports simulation results for $N = M = 40$, 80, 160, 320, and 640. The top panel reports results for Design 1 (nondegenerate case), and the bottom panel reports results for Design 2 (degenerate case). Each panel contains results based on no algorithmic subsampling (i.e., $p = 1$)[4] and results based on algorithmic subsampling (with the subsampling probabilities of $p = 1\underline{C}/(NM)$ and $p = 2\underline{C}/(NM)$) for estimation of the mean. The asymptotic variance is estimated using a random subsample of 10% of the sample. The displayed statistics are the bias (Bias), the standard deviation (SD), the root mean square error (RMSE), and the 95% coverage (95% Cover).

Observe that the 95% coverage frequencies are closer to the nominal probability of 95% when using algorithmic subsampling compared to not using it. This

[4]The 95% coverage is computed based on our asymptotic variance formula as the special case with $p = 1$.

TABLE 1. Simulation results for the additively separable design with $N = M = 40$, 80, 160, 320, and 640 based on 2,500 Monte Carlo iterations

Design 1: Non-degenerate case

		No algorithmic subsampling ($p = 1$)				Algorithmic subsampling $p = 1\underline{C}/(NM)$				Algorithmic subsampling $p = 2\underline{C}/(NM)$			
N	M	Bias	SD	RMSE	95%	Bias	SD	RMSE	95%	Bias	SD	RMSE	95%
40	40	0.006	0.127	0.127	**0.885**	0.005	0.194	0.194	**0.926**	−0.002	0.155	0.155	**0.908**
80	80	−0.001	0.087	0.087	**0.902**	0.002	0.131	0.131	**0.925**	−0.002	0.110	0.110	**0.916**
160	160	−0.001	0.060	0.060	**0.918**	0.000	0.094	0.094	**0.925**	0.002	0.079	0.079	**0.923**
320	320	0.001	0.044	0.044	**0.916**	0.000	0.068	0.068	**0.932**	0.000	0.057	0.057	**0.927**
640	640	0.001	0.032	0.032	**0.922**	0.000	0.047	0.047	**0.948**	−0.001	0.040	0.040	**0.934**

Design 2: Degenerate case

		No algorithmic subsampling ($p = 1$)				Algorithmic subsampling $p = 1\underline{C}/(NM)$				Algorithmic subsampling $p = 2\underline{C}/(NM)$			
N	M	Bias	SD	RMSE	95%	Bias	SD	RMSE	95%	Bias	SD	RMSE	95%
40	40	0.000	0.011	0.011	**0.999**	−0.001	0.105	0.105	**0.981**	−0.001	0.049	0.049	**0.986**
80	80	0.000	0.006	0.006	**1.000**	−0.001	0.051	0.051	**0.963**	0.001	0.036	0.036	**0.970**
160	160	0.000	0.003	0.003	**1.000**	0.001	0.036	0.036	**0.959**	0.000	0.026	0.026	**0.961**
320	320	0.000	0.001	0.001	**1.000**	0.000	0.025	0.025	**0.959**	0.000	0.018	0.018	**0.960**
640	640	0.000	0.001	0.001	**1.000**	0.000	0.018	0.018	**0.944**	0.000	0.013	0.013	**0.950**

Note: The top panel reports results for Design 1 (nondegenerate case), whereas the bottom panel reports results for Design 2 (degenerate case). Each panel contains results based on no algorithmic subsampling ($p = 1$), results based on algorithmic subsampling with $p = 1\underline{C}/(NM)$, and results based on algorithmic subsampling with $p = 2\underline{C}/(NM)$ for estimation of the mean. The displayed statistics are the bias (Bias), the standard deviation (SD), the root mean square error (RMSE), and the 95% coverage (95%).

observation is robustly true in both Design 1 (nondegenerate case) and Design 2 (degenerate case). For Design 2 or the degenerate case, in particular, the coverage frequency moves away from the nominal probability as the sample size increases if algorithmic subsampling were not used. On the other hand, the coverage frequency approaches the nominal probability as the sample size increases if algorithmic subsampling is used. These results demonstrate the aforementioned robustness property of multiway algorithmic subsampling against the potential degeneracy of the asymptotic distribution. We also experimented with additional simulation settings with much larger N and M and other subsampling probabilities for algorithmic subsampling variance estimation, but we observe the same qualitative patterns in the results under these alternative settings.

On the one hand, $p = 1$ leads to more precision, as quantified by smaller RMSE. On the other hand, $p = 1$ leads to larger coverage as observed above. These two phenomena may appear contradictory at first glance. The relevant issues are with the variance estimation, and not with the point estimates. These results precisely highlight the cases of degeneracy. The asymptotic normality with the $\sqrt{\underline{C}}$-rate fails under the degeneracy if we do not use algorithmic subsampling, i.e., if $p = 1$. Therefore, the standard errors are misleadingly larger compared to the actual RMSE of the estimator and the simulated coverage rates exceed the nominal coverage probability in the degenerate case with $p = 1$. This is the main reason why we propose to use algorithmic subsampling (i.e., $p < 1$) to have inference with estimated variance robust against the degeneracy.

5.2. Nonseparable Designs

Second, we generate the two-way clustered array $\{Y_{ij}\}_{i\in[N],j\in[M]}$ according to the nonadditive model

$$Y_{ij} = (\alpha_i - \mu_a)(\beta_j - \mu_b) - \mu_a\mu_b + \varepsilon_{ij},$$

where α_i, β_j and ε_{ij} are i.i.d. standard normal. With this basic setup, we consider two designs:

$$\text{Design 3}: \mu_a = 1.0, \text{ and } \mu_b = 1.0; \qquad \text{and}$$
$$\text{Design 4}: \mu_a = 0.0, \text{ and } \mu_b = 0.0.$$

Note that Design 4 would lead to a degenerate asymptotic distribution that is Gaussian chaos, which is non-Gaussian (cf. Menzel, 2021), if algorithmic subsampling were not employed.

Table 2 reports simulation results for $N = M = 40$, 80, 160, 320, and 640. The top panel reports results for Design 3 (nondegenerate case), and the bottom panel reports results for Design 4 (degenerate case). Each panel contains results based on no algorithmic subsampling (i.e., $p = 1$) and results based on algorithmic subsampling (with the subsampling probabilities of $p = 1\underline{C}/(NM)$ and $p = 2\underline{C}/(NM)$) for estimation of the mean. The asymptotic variance is estimated using a random subsample of 10% of the sample. The displayed statistics are the bias (Bias), the standard deviation (SD), the root mean square error (RMSE), and the 95% coverage (95% Cover).

Similarly to the case with the additively separable design, observe that the 95% coverage frequencies are closer to the nominal probability of 95% with the use of algorithmic subsampling than without the use of it. This observation is robustly true in both Design 3 (nondegenerate case) and Design 4 (degenerate case). For Design 4 or the degenerate case, in particular, the coverage frequency moves away from the nominal probability as the sample size increases if algorithmic subsampling were not used. On the other hand, the coverage frequency approaches the nominal probability as the sample size increases if algorithmic subsampling is used. As before, these results demonstrate the aforementioned robustness property of multiway algorithmic subsampling against the potential degeneracy of the asymptotic distribution. We also experimented with additional simulation settings

TABLE 2. Simulation results for the nonseparable design with $N = M = 40$, 80, 160, 320, and 640 based on 2,500 Monte Carlo iterations

Design 3: Non-degenerate case													
		No algorithmic subsampling				Algorithmic subsampling							
		$(p = 1)$				$p = 1\underline{C}/(NM)$				$p = 2\underline{C}/(NM)$			
N	M	Bias	SD	RMSE	95%	Bias	SD	RMSE	95%	Bias	SD	RMSE	95%
40	40	−0.002	0.221	0.221	**0.943**	0.007	0.382	0.382	**0.955**	0.003	0.316	0.315	**0.944**
80	80	0.008	0.160	0.160	**0.943**	−0.004	0.272	0.272	**0.949**	0.003	0.221	0.221	**0.956**
160	160	−0.002	0.113	0.113	**0.942**	0.001	0.190	0.190	**0.952**	0.001	0.158	0.158	**0.945**
320	320	0.000	0.079	0.079	**0.944**	0.004	0.140	0.140	**0.940**	0.004	0.110	0.110	**0.953**
640	640	−0.002	0.055	0.055	**0.956**	−0.001	0.096	0.096	**0.954**	0.001	0.079	0.079	**0.949**
Design 4: Degenerate case													
		No algorithmic subsampling				Algorithmic subsampling							
		$(p = 1)$				$p = 1\underline{C}/(NM)$				$p = 2\underline{C}/(NM)$			
N	M	Bias	SD	RMSE	95%	Bias	SD	RMSE	95%	Bias	SD	RMSE	95%
40	40	0.000	0.036	0.036	**1.000**	−0.004	0.221	0.221	**0.980**	−0.004	0.160	0.160	**0.980**
80	80	0.000	0.018	0.018	**1.000**	−0.001	0.155	0.155	**0.971**	0.001	0.110	0.110	**0.975**
160	160	0.000	0.009	0.009	**1.000**	0.002	0.113	0.113	**0.956**	−0.001	0.078	0.078	**0.966**
320	320	0.000	0.005	0.005	**0.999**	0.000	0.078	0.078	**0.956**	0.000	0.057	0.057	**0.952**
640	640	0.000	0.002	0.002	**0.999**	0.001	0.057	0.057	**0.946**	−0.001	0.039	0.039	**0.952**

Note: The top panel reports results for Design 3 (nondegenerate case), whereas the bottom panel reports results for Design 4 (degenerate case). Each panel contains results based on no algorithmic subsampling ($p = 1$), results based on algorithmic subsampling with $p = 1\underline{C}/(NM)$, and results based on algorithmic subsampling with $p = 2\underline{C}/(NM)$ for estimation of the mean. The displayed statistics are the bias (Bias), the standard deviation (SD), the root mean square error (RMSE), and the 95% coverage (95%).

with much larger N and M and other subsampling probabilities for algorithmic subsampling variance estimation, but we observe the same qualitative patterns in the results under these alternative settings.

6. APPLICATION TO SCANNER DATA

In this section, we demonstrate an application of our proposed method to an analysis of demand for differentiated products using scanner data from the Dominick's Finer Foods (DFF) retail chain.[5] Scanner data may be subject to two-way cluster dependence, as mentioned in Section 1. Specifically, common demand shocks within a market may induce statistical dependence among different products within that market. Similarly, common supply shocks by a producer may induce statistical

[5] We thank James M. Kilts Center, University of Chicago Booth School of Business for allowing us to use this dataset. It is available at https://www.chicagobooth.edu/research/kilts/datasets/dominicks.

dependence among different markets within the product produced by that producer. In this light, a researcher would like to use a two-way cluster robust variance estimate for inference about the model parameters. However, the scanner data from the DFF retail chain are too large, and today's computational resources will not permit the two-way cluster robust variance estimation in reasonable lengths of time. A simple way to overcome this problem is to use the full sample for parameter estimation and to use a subsample for variance estimation, but this approach fails to deliver robustly valid inference. Hence, we use our proposed multiway algorithmic subsampling method for estimation and two-way cluster robust inference about the key demand model parameter.

Following the literature (for instance, see a survey by Nevo, 2000) on analysis of demand for differentiated products with an additive Type-I-Extreme-Value error, we use the GMM approach with the moment restriction

$$g(W_{ij},\theta) = \zeta_{ij}(\ln(S_{ij}) - \ln(S_{0j}) - \ln(P_{ij})\theta_1 - X_{ij}^T\theta_{-1}), \qquad \textbf{(6.1)}$$

where i indexes products (universal product code, hereafter referred to as UPC), j indexes markets (store $\times$ week), S_{ij} denotes the share of product i in market j, P_{ij} denotes the price, X_{ij} denotes a vector of controls (the UPC fixed effects and a time trend), ζ_{ij} denotes instruments, and $W_{ij} = (S_{ij}, P_{ij}, X_{ij}^T, \zeta_{ij}^T)^T$.[6] In addition to the elements in X_{ij}, the instrument vector includes ζ_{ij} as an excluded variable the wholesale costs, which are calculated by inverting the gross margin. We drop those observations for which $\ln(S_{ij}) - \ln(S_{0j})$ is not finite,[7] as well as those observations with missing values. The parameter vector in the model consists of $\theta = (\theta_1, \theta_{-1}^T)^T$, and we are in particular interested in the price coefficient θ_1.

We consider four product categories: beer, oats, snacks, and canned tuna. Table 3 summarizes the sizes of the original data in terms of various dimensions. It first shows the number of UPCs, the number of weeks, and the number of stores for each product category. As we define a product as that identified by the UPC, the number of products N coincides with the number of UPCs. We define a market as the unique combination of the week and the store. Therefore, the number of markets M is close to, but is generally smaller than, the product of the number of weeks and the number of stores. It is smaller than the naïve product because of the unbalancedness of data. Finally, the bottom row shows the total number of observations, which is again smaller than the naïve product NM because of the unbalancedness in data.

[6] In cases where the model involves product fixed effects, algorithmic subsampling can be applied to within-transformation. This operation incurs additional computational costs, although this is a common issue in fixed-effect methods in general. In cases where a model involves two-way fixed effects, two-way differencing may induce a more complicated dependence structure especially under unbalanced panels. An alternative approach may be to use instrumental variables. We leave rigorous treatments of such a variety of extensions to fixed-effect models for future research.

[7] In other words, we drop observations with the zero market share. Dropping these observations may generally incur a trimming bias. We adopt this trimming as it is a standard practice in the literature of demand analysis for differentiated products markets, and we consider the possibly biased estimand as our pseudo-true value.

TABLE 3. Data sizes of the four product categories: beer, oats, snacks, and canned tuna

	Beer	Oats	Snacks	Tuna
No. of UPCs	788	96	425	94
No. of weeks	303	306	386	375
No. of stores	89	93	94	93
No. of products N	788	96	425	94
No. of markets M	22,299	26,210	32,708	31,853
No. of observations	3,990,672	1,333,465	5,427,491	1,048,575

TABLE 4. Results of the estimation of the price coefficient

	Beer	Beer	Oats	Oats	Snacks	Snacks	Tuna	Tuna
p	$\frac{100\underline{C}}{NM}$	$\frac{200\underline{C}}{NM}$	$\frac{100\underline{C}}{NM}$	$\frac{200\underline{C}}{NM}$	$\frac{100\underline{C}}{NM}$	$\frac{200\underline{C}}{NM}$	$\frac{100\underline{C}}{NM}$	$\frac{200\underline{C}}{NM}$
	0.004	0.009	0.004	0.008	0.003	0.006	0.003	0.006
Price coefficient †	−0.223**	−0.334***	−1.186***	−1.273***	−1.155***	−1.105***	−1.605*	−0.936*
	(0.102)	(0.066)	(0.173)	(0.103)	(0.159)	(0.151)	(0.985)	(0.500)
Computational time‡								
Parameter estimation	7.313	13.952	0.081	0.129	2.525	5.582	0.063	0.092
Variance estimation	1,223	4,458	34	189	676	2,901	10	46

† The standard errors are shown in parentheses under the estimates. ‡ Computational time is expressed in seconds based on a single processor of Intel Xeon Processor E5-2687W V4. ***$p < 0.01$, **$p < 0.05$, *$p < 0.10$.

We now apply our multiway algorithmic subsampling GMM with the moment function defined in (6.1) for each of the four product categories. Table 4 summarizes the estimation results. The table displays the probability p of algorithmic subsampling, the corresponding estimates and their standard errors for the price coefficient, and computational time in seconds for each of parameter estimation and asymptotic variance estimation.

First, observe that the estimates of the price coefficient are negative, as expected, and are statistically significant at the level of 95% for each column except for tuna despite efficiency loss due to algorithmic subsampling and despite the two-way cluster robustness in the asymptotic variance. As emphasized in Sections 2 and 5, algorithmic subsampling with $p \propto \underline{C}/(NM)$ allows these standard errors to have asymptotically accurate coverage robustly against potential degeneracy, unlike the conventional two-way cluster robust standard errors without algorithmic subsampling.

Second, the computational time for parameter estimation is within about a dozen seconds for each column, given that algorithmic subsampling extracts only the proportions, $p \approx 0.003$–0.009, of the original sample sizes. However, it is the asymptotic variance estimation that costs more computational time under multiway cluster dependence. Focusing on the beer product category, for instance, even the algorithmic subsampling that extracts only the $p \approx 0.004$ portion of the original sample size requires 1,223 seconds of computation for variance estimation. When the proportion doubles to $p \approx 0.009$, then the computational time nearly quadruples to 4,458 seconds. A naïve calculation implies that the use of the full sample without algorithmic subsampling would require about 3 years.

7. CONCLUSION

In this paper, we propose a novel method of algorithmic subsampling for multiway cluster-dependent data. We develop asymptotic statistical properties of this proposed method. Specifically, we develop a new uniform weak law of large numbers and a new central limit theorem for multiway algorithmic subsample means. As a consequence of the new central limit theorem, we show that algorithmic subsampling allows for robustness against the potential degeneracy of the asymptotic distribution under multiway clustering at the cost of efficiency and power loss due to the subsampling. Applying these basic asymptotic statistical theories, we derive the consistency and the asymptotic normality for the multiway algorithmic subsampling GMM estimator and the multiway algorithmic subsampling M-estimator.

Our main finding that algorithmic subsampling allows for the robustness against degeneracy in the asymptotic distribution is novel in the literature on multiway clustering. Indeed, the method of inference by MacKinnon et al. (2021) as well as Cameron et al. (2011) adapts to the Gaussian degeneracy. However, these existing methods do not adapt to the class of non-Gaussian degenerate asymptotic distributions. In contrast, the asymptotic distribution under algorithmic subsampling adapts even to the non-Gaussian degeneracy as well. The bootstrap method of Menzel (2021) is robust against non-Gaussian degeneracy. Our proposed method via algorithmic subsampling leads to the exact limit distribution, and thus nonconservative inference, unlike the method of Menzel (2021). This said, we again emphasize that these merits come at the cost of efficiency and power loss by disposing parts of big data.

Finally, we shed some light on possible future directions. This paper considers non-nested multiway clustering (as in Cameron et al., 2011). In practice, researchers may be interested in applications with nested clustering in one or more cluster dimensions. Under the current framework, one could take the coarsest levels of clustering. Handling it in a more efficient way is a useful topic but is beyond the scope of this paper. In addition, in MacKinnon, Nielsen, and Webb (2023), formal theory is developed for testing the correct level of (one-way) clustering. One could consider generalizing such a test for multiway nested clustering, which is also left for future research.

APPENDIX

Throughout this appendix, for any arrays (a_{NM}) and (b_{NM}), denote $a_{NM} \lesssim b_{NM}$ for $a_{NM} \leq Cb_{NM}$ for some positive constant C independent of sample size.

A. Choice of the Subsample Size

Theorem 1 provides a guidance on rates at which p should converge in order to guarantee robustness against degeneracy under multiway cluster sampling. Specifically, $p = p_{\mathrm{NM}}$ should be chosen so that $\Lambda = \lim_{N,M\to\infty}(\underline{C}/(NM))((1-p)/p) > 0$ holds. To this goal, it is in particular sufficient to choose

$$p = c\frac{\underline{C}}{NM} \text{ for some } c > 0.$$

For our asymptotic properties with robustness, any choice of a positive constant c works in theory. Simulation studies presented in Section 5 demonstrate that even the naïve choices, such as $c = 1$ and $c = 2$, result in excellent finite-sample performances across various alternative data generating designs.

That said, it is also useful as well to provide a data-driven method to choose c based on a well-defined criterion. In this section, we propose a method to this end following the idea of power analysis which is often employed to determine experimental sample size. Suppose that a researcher has in mind a maximum tolerable level $V_{\max}$ of the approximate variance $\Gamma/\underline{C}$ of $\widehat{L}^{-1}\sum_{i=1}^{N}\sum_{j=1}^{M} Z_{ij}f\left(W_{ij}\right)$ in the asymptotic normal approximation by Theorem 1.

First, choose a preliminary positive value of c^{pre}, set $p^{\mathrm{pre}} = c^{\mathrm{pre}}\underline{C}/(NM)$, generate i.i.d Bernoulli$\left(p^{\mathrm{pre}}\right)$ random variables $\{Z_{ij}^{\mathrm{pre}} : 1 \leq i \leq N, 1 \leq j \leq M\}$ independently from data, and set $\widehat{L}^{\mathrm{pre}} = \sum_{i=1}^{N}\sum_{j=1}^{M} Z_{ij}^{\mathrm{pre}}$. Then, estimate $\Gamma_A = \lambda_1 E\left[f(W_{11})f^T(W_{12})\right] + \lambda_2 E\left[f(W_{11})f^T(W_{21})\right]$ and $\Gamma_B = E\left[f(W_{11})f^T(W_{11})\right]$ by

$$\widehat{\Gamma}_A^{\mathrm{pre}} = \frac{\underline{C}}{(\widehat{L}^{\mathrm{pre}})^2}\sum_{i=1}^{N}\sum_{1\leq j,j'\leq M} Z_{ij}^{\mathrm{pre}}Z_{ij'}^{\mathrm{pre}} f\left(W_{ij}\right)f\left(W_{ij'}\right)$$
$$+ \frac{\underline{C}}{(\widehat{L}^{\mathrm{pre}})^2}\sum_{1\leq i,i'\leq N}\sum_{j=1}^{M} Z_{ij}^{\mathrm{pre}}Z_{i'j}^{\mathrm{pre}} f\left(W_{ij}\right)f\left(W_{i'j}\right)$$

and

$$\widehat{\Gamma}_B^{\mathrm{pre}} = \frac{1}{\widehat{L}^{\mathrm{pre}}}\sum_{i=1}^{N}\sum_{j=1}^{M} Z_{ij}^{\mathrm{pre}} f\left(W_{ij}\right)f\left(W_{ij}\right),$$

respectively. Finally, solve

$$\underline{C}V_{\max} = \widehat{\Gamma}_A^{\mathrm{pre}} + \frac{\underline{C}}{NM}\frac{NM - c\underline{C}}{c\underline{C}}\widehat{\Gamma}_B^{\mathrm{pre}}$$

for c to find the value c^* of c. This plug-in procedure yields the subsample size rate $p^{\mathrm{pre}} = c^{\mathrm{pre}}\underline{C}/(NM)$, under which the approximate variance $\Gamma/\underline{C}$ of $\widehat{L}^{-1}\sum_{i=1}^{N}\sum_{j=1}^{M} Z_{ij}f\left(W_{ij}\right)$ is

TABLE A.1. Simulation results for the additively separable design with $N = M =$ 40, 80, 160, 320 and 640 based on 2,500 Monte Carlo iterations

Design 2: Degenerate case									
		No algorithmic subsampling ($p = 1$)				Algorithmic subsampling $p = c^{*}\underline{C}/(NM)$			
N	M	Bias	SD	RMSE	95%	Bias	SD	RMSE	95%
40	40	0.000	0.011	0.011	**0.999**	0.001	0.071	0.071	**0.987**
80	80	0.000	0.006	0.006	**1.000**	0.001	0.050	0.050	**0.972**
160	160	0.000	0.003	0.003	**1.000**	−0.001	0.036	0.036	**0.964**
320	320	0.000	0.001	0.001	**1.000**	0.000	0.025	0.025	**0.953**
640	640	0.000	0.001	0.001	**1.000**	0.001	0.017	0.017	**0.955**

Note: Each panel contains results based on no algorithmic subsampling ($p = 1$) and results based on algorithmic subsampling with $p = c^{*}\underline{C}/(NM)$, for estimation of the mean. The displayed statistics are the bias (Bias), the standard deviation (SD), the root mean square error (RMSE), and the 95% coverage (95%).

close to the target level $V_{\max}$. Note the similarity of this procedure to the power analysis for sample size calculation, which is often employed by experimental researchers.

We remark that this proposed procedure of choosing the subsample size differs from that proposed by Lee and Ng (2020a, Sect. 7.2). This difference in the approaches taken is due to the different goals under different dependence structures. Lee and Ng (2020a) base their requirement for the subsample size on a condition that guarantees the subspace embedding (Lee and Ng, 2020a, Def. 1). On the other hand, we base our requirement for the subsample size on attaining robustness against degeneracy under multiway cluster sampling.

While the simulation studies presented in Section 5 are based fixed $c \in \{1,2\}$, we now present simulation results under the above choice rule of the subsample size. We set $V_{\max} = 0.5/\underline{C}$ throughout, and start with $c = 1$ for preliminary estimation of $\widehat{\Gamma}_A^{\text{pre}}$ and $\widehat{\Gamma}_B^{\text{pre}}$. Focusing on the degenerate case, Table A.1 summarizes simulation results under the additively separable designs, as the counterpart of Table 1 in the main text. Similarly, focusing on the degenerate case, Table A.2 summarizes simulation results under the nonseparable designs, as the counterpart of Table 2 in the main text. Overall, we observe qualitatively similar patterns here to those presented in the main text.

B. Additional Simulation Results

The simulation studies presented in Section 5 in the main text and Appendix A compare inference results only across those methods that assume two-way clustering. This section extends these simulation analyses by comparing the finite-sample performance of our proposed method with more conventional methods that assume i.i.d. sampling and one-way clustering as well as two-way clustering.

We continue to use the same simulation designs from Section 5. Namely, data are generated according to Designs 1–4. Sample sizes are varied as $N = M = 40$, 80, 160, 320, and 640. Each set of simulations consists of 2,500 Monte Carlo iterations. Unlike Section 5,

TABLE A.2. Simulation results for the nonseparable design with $N = M = 40$, 80, 160, 320, and 640 based on 2,500 Monte Carlo iterations

Design 4: Degenerate case									
		No algorithmic subsampling ($p = 1$)				Algorithmic subsampling $p = c^* \underline{C}/(NM)$			
N	M	Bias	SD	RMSE	95%	Bias	SD	RMSE	95%
40	40	0.000	0.036	0.036	**1.000**	0.000	0.084	0.084	**0.990**
80	80	0.000	0.018	0.018	**1.000**	0.000	0.067	0.067	**0.984**
160	160	0.000	0.009	0.009	**1.000**	−0.001	0.051	0.051	**0.966**
320	320	0.000	0.005	0.005	**0.999**	0.001	0.038	0.038	**0.957**
640	640	0.000	0.002	0.002	**0.999**	−0.000	0.027	0.027	**0.958**

Note: Each panel contains results based on no algorithmic subsampling ($p = 1$) a nd results based on algorithmic subsampling with $p = c^* \underline{C}/(NM)$, for estimation of the mean. The displayed statistics are the bias (Bias), the standard deviation (SD), the root mean square error (RMSE), and the 95% coverage (95%).

however, we also compute 95% coverage frequencies with the Eicker–Huber–White robust variance estimator (0-Way Cluster) and the conventional one-way cluster-robust variance estimator (1-Way Cluster) in addition to the two-way cluster-robust variance estimator (2-Way Cluster). Tables B.1 and B.2 summarize the results for Designs 1 and 2 and Designs 3 and 4, respectively.

In each of these two tables, we make the following observations. First, the 0-Way Cluster method suffers from severe under-coverage across all the sample sizes under the nondegenerate designs. Second, the 1-Way Cluster method suffers from under-coverage across all the sample sizes under the nondegenerate designs, while it in contrast suffers from over-coverage across all the sample sizes under the degenerate designs.

Third, comparisons between the 2-Way Cluster method without algorithmic subsampling and the 2-Way Cluster method with algorithmic subsampling remain the same as those presented in Section 5 in the main text. In particular, we conclude that the 2-Way Cluster with algorithmic subsampling is the only approach that delivers correct coverage across all the designs.

C. Proofs of the Main Results

C.1. Proof of Lemma 1

Proof. By the definition of Z_{ij}, it can be written as $Z_{ij} = \mathbb{1}\{U_{ij} \le p_{NM}\}$ for some i.i.d. $U_{ij} \sim \text{Unif}(0,1)$ independent from the data. Define $\widetilde{\mathcal{F}} = \{(u,w) \mapsto f(w) : f \in \mathcal{F}\}$ and $\widetilde{\mathcal{G}}_{NM} = \{(u,w) \overset{\widetilde{g}_{NM}}{\mapsto} \mathbb{1}(u \le p_{NM})/p_{NM}\}$. Note that Assumption 2(ii), (iii) for $\mathcal{F}$ implies that the same conditions hold with $\widetilde{\mathcal{F}}$ in place of $\mathcal{F}$. Also, note that for each (N,M), $\widetilde{\mathcal{G}}_{NM}$ consists of a single function with itself as an envelope. Therefore, by Theorem 9.15 in Kosorok (2008), for $\widetilde{g}_{NM}\widetilde{\mathcal{F}} = \{\widetilde{g}_{NM} f : f \in \widetilde{\mathcal{F}}\}$, we have that

$$\sup_Q N\left(\widetilde{g}_{NM}\widetilde{\mathcal{F}}, \|\cdot\|_{Q,2}, \sqrt{2}\epsilon \, \|\widetilde{g}_{NM}F\|_{Q,2}\right) \le \sup_Q N\left(\widetilde{\mathcal{F}}, \|\cdot\|_{Q,2}, \epsilon \, \|F\|_{Q,2}\right) 1 < \infty$$

TABLE B.1. 95% coverage frequencies of various inference methods in the additively separable designs with $N = M = 40$, 80, 160, 320, and 640 based on 2,500 Monte Carlo iterations

Design 1: Non-Degenerate case						
		No algorithmic subsampling		Algorithmic subsampling		
		0-Way	1-Way	2-Way Cluster		
N	M	Cluster	Cluster	$p = 1$	$p = 1\underline{C}/(NM)$	$p = 2\underline{C}/(NM)$
40	40	0.270	0.834	0.885	0.926	0.908
80	80	0.192	0.837	0.902	0.925	0.916
160	160	0.140	0.854	0.918	0.925	0.923
320	320	0.097	0.846	0.916	0.932	0.927
640	640	0.074	0.845	0.922	0.948	0.934

Design 2: Degenerate case						
		No algorithmic subsampling		Algorithmic subsampling		
		0-Way	1-Way	2-Way Cluster		
N	M	Cluster	Cluster	$p = 1$	$p = 1\underline{C}/(NM)$	$p = 2\underline{C}/(NM)$
40	40	0.952	1.000	0.999	0.981	0.986
80	80	0.960	1.000	1.000	0.963	0.970
160	160	0.945	1.000	1.000	0.959	0.961
320	320	0.948	1.000	1.000	0.959	0.960
640	640	0.951	1.000	1.000	0.944	0.950

uniformly over (N,M) for any finite discrete measure Q and $\epsilon \in (0,1]$. Note that $\widetilde{g}_{NM}\widetilde{\mathcal{F}}$ satisfies Assumption 2(ii), (iii). Under Assumptions 1 and 2(ii), (iii), therefore, we can apply Lemma D.3 (Appendix D) to $\widetilde{g}_{NM}\widetilde{\mathcal{F}}$, and then apply the Markov inequality to get

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{Z_{ij}}{p_{NM}}f\left(W_{ij}\right)-\frac{1}{p_{NM}}E\left[Z_{11}f\left(W_{11}\right)\right]\right|\xrightarrow{P}0.$$

Since $E[Z_{11}f(W_{11})] = E[Z_{11}]E[f(W_{11})] = p_{NM}E[f(W_{11})]$, we in turn obtain

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{L}\sum_{i=1}^{N}\sum_{j=1}^{M}Z_{ij}f\left(W_{ij}\right)-E\left[f\left(W_{11}\right)\right]\right|\xrightarrow{P}0.$$

Finally, Lemma D.2 (Appendix D) implies that $\widehat{L}/L \xrightarrow{P} 1$, and thus

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{\widehat{L}}\sum_{i=1}^{N}\sum_{j=1}^{M}Z_{ij}f\left(W_{ij}\right)-E\left[f\left(W_{11}\right)\right]\right|\xrightarrow{P}0.$$

This completes the proof. □

TABLE B.2. 95% coverage frequencies of various inference methods in the nonseparable designs with $N = M = 40$, 80, 160, 320, and 640 based on 2,500 Monte Carlo iterations

Design 3: Non-degenerate case						
		No algorithmic subsampling			Algorithmic subsampling	
		0-Way	1-Way		2-Way Cluster	
N	M	Cluster	Cluster	$p=1$	$p=1\underline{C}/(NM)$	$p=2\underline{C}/(NM)$
40	40	0.341	0.914	0.943	0.955	0.944
80	80	0.240	0.927	0.943	0.949	0.956
160	160	0.170	0.924	0.942	0.952	0.945
320	320	0.122	0.916	0.944	0.940	0.953
640	640	0.089	0.923	0.956	0.954	0.949
Design 4: Degenerate case						
		No algorithmic subsampling			Algorithmic subsampling	
		0-Way	1-Way		2-Way Cluster	
N	M	Cluster	Cluster	$p=1$	$p=1\underline{C}/(NM)$	$p=2\underline{C}/(NM)$
40	40	0.950	1.000	1.000	0.980	0.980
80	80	0.940	1.000	1.000	0.971	0.975
160	160	0.946	1.000	1.000	0.956	0.966
320	320	0.946	1.000	0.999	0.956	0.952
640	640	0.953	1.000	0.999	0.946	0.952

C.2. Proof of Theorem 1

Proof. Consider the decomposition of $\widehat{L}^{-1}\sum_{i=1}^{N}\sum_{j=1}^{M} Z_{ij} f\left(W_{ij}\right)$ into two terms as

$$\frac{1}{\widehat{L}}\sum_{i=1}^{N}\sum_{j=1}^{M} Z_{ij} f\left(W_{ij}\right) = \frac{L}{\widehat{L}}\frac{1}{L}\sum_{i=1}^{N}\sum_{j=1}^{M} Z_{ij} f\left(W_{ij}\right) = \frac{L}{\widehat{L}}\left(A_{NM} + \sqrt{1-p_{NM}}B_{NM}\right), \tag{C.1}$$

where A_{NM} and B_{NM} are defined by

$$A_{NM} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} f\left(W_{ij}\right) \quad \text{and} \quad B_{NM} = \frac{1}{L}\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{Z_{ij}-p_{NM}}{\sqrt{1-p_{NM}}} f\left(W_{ij}\right),$$

respectively.

The first step is to get the asymptotic normality for A_{NM}. Our setup satisfies the first part of Assumption 3 in Davezies, D'Haultfoeuille, and Guyonvarch (2018), since class $\mathcal{F}$ is finite and $E[F^2] < \infty$. Under our Assumptions 1, 2(i), (iii), and 3, applying Theorem 3.1 of Davezies et al. (2018) yields

$$\sqrt{\underline{C}}A_{NM} \xrightarrow{d} N\left(0, \lambda_1 E\left[f\left(W_{11}\right) f^T\left(W_{12}\right)\right] + \lambda_2 E\left[f\left(W_{11}\right) f^T\left(W_{21}\right)\right]\right). \tag{C.2}$$

Next, we will obtain the variance–covariance matrix of B_{NM}. By the law of total covariance,

$$\begin{aligned}\mathrm{Cov}\left(B_{NM}, B_{NM}\right) = {} & E\left[\mathrm{Cov}\left(B_{NM}, B_{NM}\middle|\left\{W_{ij}\right\}_{i\in[N], j\in[M]}\right)\right] \\ & + \mathrm{Cov}\left(E\left[B_{NM}|\{W_{ij}\}_{i\in[N], j\in[M]}\right], E\left[B_{NM}\middle|\{W_{ij}\}_{i\in[N], j\in[M]}\right]\right).\end{aligned}$$

For the first term, we can write

$$\begin{aligned}& E\left[\mathrm{Cov}\left(B_{NM}, B_{NM}\middle|\left\{W_{ij}\right\}_{i\in[N], j\in[M]}\right)\right] \\ & = E\left[\frac{1}{L^2(1-p_{NM})}\sum_{i=1}^{N}\sum_{j=1}^{M} f\left(W_{ij}\right) f^T\left(W_{ij}\right)\mathrm{Var}\left(Z_{ij}|\{W_{ij}\}_{i\in[N], j\in[M]}\right)\right] \\ & = E\left[\frac{1}{L^2(1-p_{NM})}\sum_{i=1}^{N}\sum_{j=1}^{M} f\left(W_{ij}\right) f^T\left(W_{ij}\right) p_{NM}(1-p_{NM})\right] \\ & = \frac{p_{NM}}{L^2}\sum_{i=1}^{N}\sum_{j=1}^{M} E\left[f\left(W_{ij}\right) f^T\left(W_{ij}\right)\right] \\ & = \frac{1}{L} E\left[f\left(W_{ij}\right) f^T\left(W_{ij}\right)\right].\end{aligned}$$

For the last term, note that

$$\begin{aligned}E\left[B_{NM}\middle|\left\{W_{ij}\right\}_{i\in[N], j\in[M]}\right] & = \frac{1}{L\sqrt{1-p_{NM}}}\sum_{i=1}^{N}\sum_{j=1}^{M} f\left(W_{ij}\right) E\left[(Z_{ij}-p_{NM})|\{W_{ij}\}_{i\in[N], j\in[M]}\right] \\ & = \frac{1}{L\sqrt{1-p_{NM}}}\sum_{i=1}^{N}\sum_{j=1}^{M} f\left(W_{ij}\right) E\left[(Z_{ij}|\{W_{ij}\}_{i\in[N], j\in[M]}) - p_{NM}\right] \\ & = \frac{1}{L\sqrt{1-p_{NM}}}\sum_{i=1}^{N}\sum_{j=1}^{M} f\left(W_{ij}\right)\left(p_{NM}-p_{NM}\right) = 0.\end{aligned}$$

Therefore,

$$\mathrm{Cov}\left(E\left[B_{NM}|\{W_{ij}\}_{i\in[N], j\in[M]}\right], E\left[B_{NM}\middle|\{W_{ij}\}_{i\in[N], j\in[M]}\right]\right) = 0.$$

It thus follows that

$$\mathrm{Cov}\left(B_{NM}, B_{NM}\right) = \frac{1}{L} E\left[f\left(W_{ij}\right) f^T\left(W_{ij}\right)\right] = \frac{1}{L} E\left[f\left(W_{11}\right) f^T\left(W_{11}\right)\right], \tag{C.3}$$

where the second equality holds by Assumption 1(i).

We now show that the term $(A_{NM} + \sqrt{1-p_{NM}}B_{NM})$ is asymptotically normal. Pick any $q = (q_1, \ldots, q_k)^T \in \mathbb{R}^k$. For a given bounded sequence $\left\{a_{ij}\right\}$, define

$$Y_{NM,L} = \frac{1}{\sqrt{L}}\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{\left(Z_{ij}-p_{NM}\right) a_{ij}}{\sqrt{1-p_{NM}}} \quad \text{and} \quad \alpha_{NM}^2 = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} a_{ij}^2.$$

And, suppose $f(W_{ij})$ is bounded, by applying Lemma 2 of Janson (1984) with $a_{ij} = q^T f(W_{ij})$, conditionally on $\{W_{ij}\}_{i\in[N],j\in[M]}$, we obtain

$$E\left(e^{itY_{NM,L}}\middle|\{W_{ij}\}_{i\in[N],j\in[M]}\right) - e^{-t^2\alpha_{NM}^2/2} \to 0.$$

Meanwhile, $(NM)^{-1}\sum_{i=1}^N\sum_{j=1}^M f^T(W_{ij})qq^Tf(W_{ij}) \xrightarrow{P} E[f^T(W_{11})qq^Tf(W_{11})]$, and thus $\alpha_{NM}^2 \xrightarrow{P} \alpha^2$, where $\alpha^2 = E[f^T(W_{11})qq^Tf(W_{11})]$, so that the above conditional characteristic function converges to $e^{-t^2\alpha^2/2}$. Thus, conditionally on $\{W_{ij}\}_{i\in[N],j\in[M]}$, we have $\sqrt{L}q^TB_{NM}/\alpha \xrightarrow{d} N(0,1)$. Also note that conditional on $\{W_{ij}\}_{i\in[N],j\in[M]}$, A_{NM} is deterministic. In addition, we have already shown that $\sqrt{\underline{C}}A_{NM}$ is (unconditionally) asymptotically normal as in (C.2). Therefore, an application[8] of Theorem 2 in Chen and Rao (2007) yields that

$$\begin{aligned}\sqrt{\underline{C}}\frac{1}{L}\sum_{i=1}^N\sum_{j=1}^M Z_{ij}q^Tf\left(W_{ij}\right) &= \sqrt{\underline{C}}\frac{1}{NM}\sum_{i=1}^N\sum_{j=1}^M q^Tf\left(W_{ij}\right)\\ &\quad + \frac{\sqrt{\underline{C}}}{\sqrt{L}}\sqrt{1-p_{NM}}\frac{1}{\sqrt{L}}\sum_{i=1}^N\sum_{j=1}^M\frac{Z_{ij}-p_{NM}}{\sqrt{1-p_{NM}}}q^Tf\left(W_{ij}\right)\\ &\xrightarrow{d} N\left(0, q^T\Gamma q\right),\end{aligned}$$

recall that $\Gamma = \Gamma_A + \Lambda\Gamma_B$ with $\Gamma_A = \lambda_1E[f(W_{11})f^T(W_{12})] + \lambda_2E[f(W_{11})f^T(W_{21})]$ and $\Gamma_B = E[f(W_{11})f^T(W_{11})]$. The Cramér–Wold device now implies

$$\sqrt{\underline{C}}(A_{NM} + \sqrt{1-p_{NM}}B_{NM}) = \sqrt{\underline{C}}\frac{1}{L}\sum_{i=1}^N\sum_{j=1}^M Z_{ij}f\left(W_{ij}\right) \xrightarrow{d} N\left(0,\Gamma\right). \tag{C.5}$$

In cases where f is unbounded, one can approximate f in L^2 using a bounded function f' following the argument in Theorem 1 of Janson (1984, p. 499) under Lemma D.2 and the condition $E[F^2] < \infty$ in the statement of the theorem. The resulting errors in $\sqrt{\underline{C}}A_{NM}$ and $\sqrt{L}B_{NM}$ have variances bounded by $T_1E[(f(W_{11})-f'(W_{11}))(f(W_{12})-f'(W_{12}))^T] + T_2E[(f(W_{11})-f'(W_{11}))(f(W_{21})-f'(W_{21}))^T]$ and $T_3E[(f(W_{11})-f'(W_{11}))(f(W_{11})-f'(W_{11}))^T]$ from (C.2) and (C.3), where T_1, T_2, and T_3 are constants. The result then follows by letting $f' \to f$ with an application of the dominated convergence theorem.

Finally, $\sqrt{\underline{C}}L^{-1}\sum_{i=1}^N\sum_{j=1}^M Z_{ij}f(W_{ij})$ can be replaced by $\sqrt{\widehat{\underline{C}}}\widehat{L}^{-1}\sum_{i=1}^N\sum_{j=1}^M Z_{ij}f(W_{ij})$ by virtue of Lemma D.2 (Appendix D). □

C.3. Proof of Corollary 1

Proof. Since $E[||X_{11}||^4] < \infty$, we have $E[X_{r,11}^4] < \infty$, for any coordinate $X_{r,11}$ of X_{11}. The condition $E[|Y_{11}|^4] < \infty$ implies $E[u_{11}^4] < \infty$. By the Cauchy–Schwarz inequality, we have $E[(X_{r,11}u_{11})^2] < \infty$ and $E[(X_{r,11}X_{r',11})^2] < \infty$ for any r and r', and also

[8]We thank a reviewer for suggesting this proof strategy, which simplifies the proof.

$E[(||X_{11}u_{11}||)^2] < \infty$. Applying Lemma 1 to the function class $\mathcal{F}_{OLS,1} = \{f(W_{ij}) = X_{r,11}X_{r',11}, \text{ for all } r, r'\}$, we have

$$\frac{1}{\widehat{L}}\sum_{i=1}^{N}\sum_{j=1}^{M} Z_{ij}X_{ij}X_{ij}^T \xrightarrow{P} E[X_{11}X_{11}^T].$$

Now, let a vector μ have the same dimension as β, and denote $f_\mu(W_{ij}) = \mu^T X_{ij}u_{ij}$. Applying Theorem 1 to $\mathcal{F}_{OLS,2} = \{f_\mu(W_{ij}), i \in \{1,\ldots,N\}, j \in \{1,\ldots,M\}\}$, we obtain

$$\sqrt{\underline{C}}\frac{1}{\widehat{L}}\sum_{i=1}^{N}\sum_{j=1}^{M} Z_{ij}f_\mu(W_{ij}) \xrightarrow{d} N(0, \mu^T\Gamma_{OLS}\mu),$$

where $\Gamma_{OLS} = \Gamma_{OLS,1} + \Lambda\Gamma_{OLS,2}$, $\Gamma_{OLS,1} = \lambda_1 E[X_{11}u_{11}(X_{12}u_{12})^T] + \lambda_2 E[X_{11}u_{11}(X_{21}u_{21})^T]$, and $\Gamma_{OLS,2} = E[X_{11}u_{11}(X_{11}u_{11})^T]$. Cramér–Wold device thus yields

$$\sqrt{\underline{C}}\frac{1}{\widehat{L}}\sum_{i=1}^{N}\sum_{j=1}^{M} Z_{ij}X_{ij}u_{ij} \xrightarrow{d} N(0, \Gamma_{OLS}).$$

Finally, applying Slutsky's lemma yields

$$\sqrt{\underline{C}}(\widehat{\beta} - \beta) \xrightarrow{d} N(0, V),$$

where $V = J^{-1}\Gamma_{OLS}J^{-1}$ and $J = E\left[X_{11}X_{11}^T\right]$. □

D. Useful Lemmas

In this appendix section, we state auxiliary lemmas that are used to prove our main results. Each of these results is either coming directly from the existing literature or is the existing result with minor modifications. For the latter case, we provide a proof.

LEMMA D.1. *Let Θ be a compact subset of $\mathbb{R}^k$, and let $\mathcal{F} = \{f(\cdot,\theta) : \theta \in \Theta\}$ be a class of real-valued functions indexed by θ such that $f(w,\cdot)$ is continuous for all $w \in \operatorname{supp}(W_{ij})$. Then, $\mathcal{F}$ is a pointwise measurable class of functions.*

Proof. The proof is immediate and well known. We provide proof for completeness. Let $\mathcal{S} = \{f(\cdot,\theta) : \theta \in \Theta \cap \mathbb{Q}^k\}$, where $\mathbb{Q}$ is the rationals. Therefore, by the denseness of $\mathbb{Q}^k$, for each $w \in \operatorname{supp}(W_{ij})$, we can find $(\theta_m) \subset \Theta \cap \mathbb{Q}^k$, $\theta_m \to \theta$ as $m \to \infty$ and then the continuity implies $f(w,\theta_m) \to f(w,\theta)$, which coincides with the definition of pointwise measurability. □

The next lemma follows immediately from van der Vaart and Wellner (1996, Lem. 2.2.9).

LEMMA D.2 (Bernstein's inequality for Bernoulli r.v.'s). *For each $p \in (0,1]$, it holds that*

$$P(|\widehat{L}/L - 1| > \sqrt{2t/L + 2t/(3L)}) \leq 2e^{-t},$$

for every $t > 0$.

Lemmas D.3 and D.4 follow closely from Theorem 3.4(i) in Davezies et al. (2021) and Lemma D.12 in Davezies et al. (2018), respectively.

LEMMA D.3 (Glivenko–Cantelli for two-way clustered random variables). *Let $(\mathcal{F}_{NM})$ be a sequence of classes of functions that satisfies Assumption 2(iii) and such that each $\mathcal{F}_{NM}$ admits an envelope F_{NM} with $E[F_{NM}(W_{11})] \le \overline{M} < \infty$, $\sup_Q \log N\left(\mathcal{F}_{NM}, \|\cdot\|_{Q,2}, \epsilon\|F_{NM}\|_{Q,2}\right) < \infty$ for any finite discrete measure Q, $\epsilon \in (0,1]$, then under Assumption 1, we have*

$$E\left[\sup_{f\in\mathcal{F}_{NM}}\left|\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} f\left(W_{ij}\right) - E[f(W_{11})]\right|\right] = o(1).$$

Proof. The result is a minor modification of the proof of Theorem 3.4(i) in Davezies et al. (2021) with the standard Glivenko–Cantelli theorem modified for function classes changing with the sample size. Denote $\mathbb{P}_{NM} = (NM)^{-1}\sum_{i=1}^{N}\sum_{j=1}^{M}\delta_{X_{ij}}$, where δ_x is the Dirac measure at x. Following their symmetrization argument (which is nonasymptotic and independent of the function class) in the proof of Theorem 3.4(i) in Davezies et al. (2021), for each $K > 0$ and $\epsilon > 0$, denote $\mathcal{F}_{NM,K} = \mathcal{F}_{NM}\mathbb{1}\{F_{NM} > K\}$, then one has

$$\begin{aligned}
&E\left[\sup_{f\in\mathcal{F}_{NM}}\left|\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} f\left(W_{ij}\right) - E[f(W_{11})]\right|\right]\\
&\quad\lesssim E[F_{NM}\mathbb{1}\{F_{NM} > K\}] + E\left[\epsilon + \frac{K}{\sqrt{NM}}\sqrt{\log N(\mathcal{F}_{NM,K}, \|\cdot\|_{\mathbb{P}_{NM},1}, \epsilon)}\right].
\end{aligned}$$

The first term on the right-hand side is bounded by $\overline{M}$. To deal with the second term, by Jensen's inequality, it holds that $\|f - f'\|_{\mathbb{P}_{NM},1} \le \|f - f'\|_{\mathbb{P}_{NM},2}$. Thus, the smallest ϵ-net for $(\mathcal{F}_{NM}, \|\cdot\|_{\mathbb{P}_{NM},2})$ is an ϵ-net for $(\mathcal{F}_{NM}, \|\cdot\|_{\mathbb{P}_{NM},1})$. Thus, we have $N(\mathcal{F}_{NM}, \|\cdot\|_{\mathbb{P}_{NM},1}, \epsilon) \le N(\mathcal{F}_{NM}, \|\cdot\|_{\mathbb{P}_{NM},2}, \epsilon)$. The condition $\sup_Q \log N\left(\mathcal{F}_{NM}, \|\cdot\|_{Q,2}, \epsilon\|F_{NM}\|_{Q,2}\right) < \infty$ for all $\epsilon \in (0,1]$ implies $N(\mathcal{F}_{NM}, \|\cdot\|_{\mathbb{P}_{NM},2}, \epsilon) < \infty$ for all $\epsilon > 0$. Finally, observe that $E[\|F\|_{\mathbb{P}_{NM},1}] = E[F_{NM}] < M$. This concludes the proof. □

LEMMA D.4 (Lemma D.11 in Davezies et al. (2018) for sequences). *Let $(\mathcal{F}_{NM})$ and $(\mathcal{G}_{NM})$ be two pointwise measurable classes of functions. Suppose that each $\mathcal{F}_{NM}$ admits an envelope F_{NM} with $E\left[F_{NM}(W_{11})^2\right] < \infty$ and*

$$\int_0^1 \sup_Q \sqrt{\log N\left(\mathcal{F}_{NM}, \|\cdot\|_{Q,2}, \epsilon\|F_{NM}\|_{Q,2}\right)}d\epsilon \le \overline{M} < \infty,$$

where Q is taken over the set of all finite discrete measures and $\epsilon \in (0,1]$. Similarly, $(\mathcal{G}_{NM})$ admits a sequence of envelope functions (G_{NM}) with $E\left[G_{NM}(W_{11})^2\right] < \infty$ and

$$\int_0^1 \sup_Q \sqrt{\log N\left(\mathcal{G}_{NM}, \|\cdot\|_{Q,2}, \epsilon\|G_{NM}\|_{Q,2}\right)}d\epsilon \le \overline{M} < \infty.$$

Then, under Assumptions 1 and 3,

$$\lim_{\underline{C}\to\infty} E\left[\sup_{\mathcal{F}_{NM}\times\mathcal{G}_{NM}}\left|\frac{\underline{C}}{(NM)^2}\sum_{i=1}^{N}\sum_{1\le j,j'\le M} f\left(W_{ij}\right)g\left(W_{ij'}\right)-\lambda_1 E[f(W_{11})g(W_{12})]\right|\right]=0.$$

Proof. The proof follows the same steps of the proof of Lemma D.11 of Davezies et al. (2018) with the modification of $\mathcal{F}_{NM}$, F_{NM}, $\mathcal{G}_{NM}$, and G_{NM} in place of $\mathcal{F}$, F, $\mathcal{G}$, and G, respectively. Notice that their symmetrization arguments and Lemma D.4 are nonasymptotic and thus are not affected by such modification. The detail is omitted. □

E. Proofs for the Application to GMM

Define the class $\mathcal{G}=\{g_r(\cdot,\theta):\theta\in\Theta \text{ and } r\in\{1,\dots,m\}\}$ of functions indexed by r and θ and the class $\mathcal{G}'=\left\{\partial g_r(\cdot,\theta)/\partial\theta_l:\theta\in\Theta, r\in\{1,\dots,m\}, l\in\{1,\dots,k\}\right\}$ of functions indexed by r, l, and θ.

E.1. Proof of Lemma 2

Proof. We verify the conditions of Theorem 2.1 in Newey and McFadden (1994), where the population criterion is $Q_0(\theta)=-E\left[g\left(W_{ij},\theta\right)\right]^T VE\left[g\left(W_{ij},\theta\right)\right]$. Their condition 2.1(i), $Q_0(\theta)$ is uniquely maximized at θ^0, holds by Lemma 2.3 in Newey and McFadden (1994) under Assumption 4(i). Condition 2.1(ii) holds by Assumption 4(ii). Condition 2.1(iii) that Q_0 is continuous at θ follows from Assumption 4(iii)(a). Under Assumption 4(ii), (iii)(a), (iv), by Example 19.7 in Van der Vaart (2000) and Lemma 9.18 in Kosorok (2008), we know that the class $\mathcal{G}_r=\{g_r(\cdot,\theta),\theta\in\Theta\}$, for $r\in\{1,\dots,m\}$, has an envelope $\left|g_r(W_{ij},\theta^0)\right|+DM<\infty$, where D is a diameter of a set containing Θ. Thus, for any finite discrete measure Q and $\epsilon\in(0,1]$, $N\left(\mathcal{G}_r,\|\cdot\|_{Q,2},\epsilon\right)\le(1+4DM/\epsilon)^k$. Since $\mathcal{G}=\bigcup_{r=1}^m\mathcal{G}_r$, we obtain $N\left(\mathcal{G},\|\cdot\|_{Q,2},\epsilon\right)\le m(1+4DM/\epsilon)^k<\infty$, which implies that the class $\mathcal{G}$ satisfies Assumption 2(ii). $\mathcal{G}$ is a pointwise measurable class of functions since $\mathcal{G}_r$, for $r\in\{1,\dots,m\}$, is a pointwise measurable class of functions by Lemma D.1 under Assumption 4(ii), (iii)(a) and $\mathcal{G}=\bigcup_{r=1}^m\mathcal{G}_r$. Thus, with Assumption 1, by applying Lemma 1, we have $\sup_{\theta\in\Theta}\left\|\widehat{g}_{NM}(\theta)-E\left[g\left(W_{ij},\theta\right)\right]\right\|\xrightarrow{P}0$. By the triangle and Cauchy–Schwartz inequalities, we obtain

$$\begin{aligned}
&\left|\widehat{Q}_{NM}(\theta)-Q_0(\theta)\right|\\
&\quad\le\left|\left(\widehat{g}_{NM}(\theta)-E\left[g\left(W_{ij},\theta\right)\right]\right)^T\widehat{V}\left(\widehat{g}_{NM}(\theta)-E\left[g\left(W_{ij},\theta\right)\right]\right)\right|\\
&\qquad+\left|E\left[g\left(W_{ij},\theta\right)\right]^T\left(\widehat{V}+\widehat{V}^T\right)\left(\widehat{g}_{NM}(\theta)-E\left[g\left(W_{ij},\theta\right)\right]\right)\right|\\
&\qquad+\left|E\left[g\left(W_{ij},\theta\right)\right]^T\left(\widehat{V}-V\right)E\left[g\left(W_{ij},\theta\right)\right]\right|\\
&\quad\le\left\|\widehat{g}_{NM}(\theta)-E\left[g\left(W_{ij},\theta\right)\right]\right\|^2\left\|\widehat{V}\right\|+2\left\|E\left[g\left(W_{ij},\theta\right)\right]\right\|\left\|\widehat{g}_{NM}(\theta)-E\left[g\left(W_{ij},\theta\right)\right]\right\|\left\|\widehat{V}\right\|\\
&\qquad+\left\|E\left[g\left(W_{ij},\theta\right)\right]\right\|^2\left\|\widehat{V}-V\right\|.
\end{aligned}$$

Thus, $\sup_{\theta\in\Theta}\left|\widehat{Q}_{NM}(\theta)-Q_0(\theta)\right| \xrightarrow{P} 0$ so that condition 2.1(iv) is satisfied. Applying Theorem 2.1 in Newey and McFadden (1994), we therefore obtain $\widehat{\theta} \xrightarrow{P} \theta^0$. □

E.2. Proof of Theorem 2

Proof. Under Assumption 4(ii), (iii)(a), the first-order condition requires that $2\widehat{G}_{NM}\left(\widehat{\theta}\right)^T \widehat{V}\widehat{g}_{NM}\left(\widehat{\theta}\right) = 0$ holds with probability approaching one, where $\widehat{G}_{NM}(\theta) = \nabla_\theta \widehat{g}_{NM}(\theta)$. Expanding $\widehat{g}_{NM}\left(\widehat{\theta}\right)$ around θ^0 and multiplying by $\sqrt{\underline{C}}$, we have

$$\sqrt{\underline{C}}\left(\widehat{\theta}-\theta^0\right) = -\left[\widehat{G}_{NM}\left(\widehat{\theta}\right)^T\widehat{V}\widehat{G}_{NM}\left(\bar{\theta}\right)\right]^{-1}\widehat{G}_{NM}\left(\widehat{\theta}\right)^T\widehat{V}\sqrt{\underline{C}}\widehat{g}_{NM}\left(\theta^0\right),$$

where $\bar{\theta}$ is the mean value implied by the mean value theorem for each coordinate. Under Assumption 4(ii), (iii)(b), similar lines of argument to those in the proof of Lemma 2 yield $N\left(\mathcal{G}', \|\cdot\|_{Q,2}, \epsilon\right) < \infty$ for any finite discrete measure Q and $\epsilon \in (0,1]$. $\mathcal{G}'$ is a pointwise measurable class of functions by Lemma D.1 under Assumption 4(ii), (iii)(b). With Assumptions 1 and 4(vi), Lemma 1 thus yields

$$\sup_{\theta\in\Theta}\left|\frac{1}{\widehat{L}}\sum_{i=1}^{N}\sum_{j=1}^{M} Z_{ij}\frac{\partial g_r\left(W_{ij},\theta\right)}{\partial\theta_l} - E\left[\frac{\partial g_r\left(W_{11},\theta\right)}{\partial\theta_l}\right]\right| \xrightarrow{P} 0$$

for each r and l. Since there are only finite numbers of r and l, it follows that $\widehat{G}_{NM}(\widehat{\theta}) - G(\widehat{\theta}) \xrightarrow{P} 0$ and $\widehat{G}_{NM}(\bar{\theta}) - G(\bar{\theta}) \xrightarrow{P} 0$, where $G(\theta) = E[\nabla_\theta g(W_{11},\theta)]$. Also, since the conditions of Lemma 2 are satisfied, we have $\bar{\theta} \xrightarrow{P} \theta^0$ and $\widehat{\theta} \xrightarrow{P} \theta^0$. We thus obtain $G(\widehat{\theta}) - G \xrightarrow{P} 0$ and $G(\bar{\theta}) - G \xrightarrow{P} 0$ by the continuous mapping theorem under Assumption 4(iii)(b). Combining the above results yields $\widehat{G}_{NM}(\widehat{\theta}) \xrightarrow{P} G$ and $\widehat{G}_{NM}(\bar{\theta}) \xrightarrow{P} G$. Therefore, $[\widehat{G}_{NM}(\widehat{\theta})^T\widehat{V}\widehat{G}_{NM}(\bar{\theta})]^{-1}\widehat{G}_{NM}(\widehat{\theta})^T\widehat{V} \xrightarrow{P} (G^TVG)^{-1}G^TV$ follows by an application of the continuous mapping theorem under Assumption 4(v). Now, notice that finite function class $\{g_1(\cdot,\theta^0),\ldots,g_m(\cdot,\theta^0)\}$ is pointwise measurable since $\mathcal{G}$ is a pointwise measurable class of functions following Lemma 2 and $E[g(W_{ij},\theta^0)] = 0$. With $E[g_{\sup}(W_{ij})^2] < \infty$ under Assumption 4(vii), by applying Theorem 1 under Assumptions 1 and 3, we obtain $\sqrt{\underline{C}}\widehat{g}_{NM}(\theta^0) \xrightarrow{d} N(0,\Omega)$, where $\Omega = \Gamma_1 + \Lambda\Gamma_2$, with $\Gamma_1 = \lambda_1 E[g(W_{11},\theta^0)g^T(W_{12},\theta^0)] + \lambda_2 E[g(W_{11},\theta^0)g^T(W_{21},\theta^0)]$ and $\Gamma_2 = E[g(W_{11},\theta^0)g^T(W_{11},\theta^0)]$. Slutsky's theorem then implies

$$\sqrt{\underline{C}}\left(\widehat{\theta}-\theta^0\right) \xrightarrow{d} N\left(0, \left(G^TVG\right)^{-1}G^TV\Omega VG\left(G^TVG\right)^{-1}\right),$$

which concludes the proof. □

E.3. Proof of Theorem 3

Proof. First, we want to establish $\widetilde{G} \xrightarrow{P} G$ via $\|\widetilde{G}-G\| \le \|\widetilde{G}-G(\widehat{\theta})\| + \|G(\widehat{\theta})-G\|$, where $G(\theta) = E[\nabla_\theta g(W_{11},\theta)]$. Since the conditions of Lemma 2 are satisfied, it holds that $\widehat{\theta} \xrightarrow{P} \theta^0$. Under Assumption 5(i), we obtain $\left\|G\left(\widehat{\theta}\right)-G\right\| \xrightarrow{P} 0$ by the continuous mapping theorem. Note that $\mathcal{G}'$ is pointwise measurable and $N\left(\mathcal{G}', \|\cdot\|_{Q,2}, \epsilon\right) < \infty$ for any finite discrete measure Q, $\epsilon \in (0,1]$ by the proof of Theorem 2. Under Assumptions 1 and 4(vi), by applying Lemma 1, we thus obtain

$$\sup_{\theta\in\Theta}\left|\frac{1}{\widehat{L}}\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{\partial g_r\left(W_{ij},\theta\right)Z_{ij}}{\partial\theta_l}-E\left[\frac{\partial g_r\left(W_{11},\theta\right)}{\partial\theta_l}\right]\right|\xrightarrow{P}0$$

for each r and l. Since there are only finite numbers of l and r, we get

$$\sup_{\theta\in\Theta}\left\|\frac{1}{\widehat{L}}\sum_{i=1}^{N}\sum_{j=1}^{M}Z_{ij}\nabla_\theta g\left(W_{ij},\theta\right)-E\left[\nabla_\theta g\left(W_{11},\theta\right)\right]\right\|\xrightarrow{P}0.$$

It then follows that $\widetilde{G}\xrightarrow{P}G(\widehat{\theta})$. Combining the above arguments, we establish $\widetilde{G}\xrightarrow{P}G$.

We will next verify $\widetilde{\Gamma}_2\xrightarrow{P}\Gamma_2$. Define a new class $\mathcal{G}_{\text{sub}}=\{(w,z)\mapsto zg_r(w,\theta),$ $\theta\in\Theta, r\in\{1,\ldots,m\}\}$. For any finite discrete measure Q and $\epsilon\in(0,1]$, we have $\sup_Q N\left(\mathcal{G},\|\cdot\|_{Q,2},\epsilon\left\|g_{\text{sup}}\right\|_{Q,2}\right)<\infty$ by the proof of Lemma 2. By Theorem 9.15 in Kosorok (2008), therefore,

$$\begin{aligned}
&\sup_Q N\left(\mathcal{G}_{\text{sub}}\mathcal{G}_{\text{sub}},\|\cdot\|_{Q,2},\sqrt{2}\epsilon\left\|g_{\text{sup}}^2\right\|_{Q,2}\right)\\
&\leq\sup_Q N\left(\mathcal{G}_{\text{sub}},\|\cdot\|_{Q,2},\epsilon\left\|g_{\text{sup}}\right\|_{Q,2}\right)\sup_Q N\left(\mathcal{G}_{\text{sub}},\|\cdot\|_{Q,2},\epsilon\left\|g_{\text{sup}}\right\|_{Q,2}\right)<\infty
\end{aligned}$$

for any finite discrete measure Q and $\epsilon\in(0,1]$, where $\mathcal{G}_{\text{sub}}\mathcal{G}_{\text{sub}}$ is defined as the pointwise product. Note that $\mathcal{G}_{\text{sub}}\mathcal{G}_{\text{sub}}$ is a pointwise measurable class of functions since $\mathcal{G}_{\text{sub}}$ is a pointwise measurable class of functions by the arguments in the proof of Lemma 2. This implies that with $E[g_{\text{sup}}(W_{ij})^2]<\infty$ and Assumption 1, by applying Lemma D.3, we thus obtain

$$\begin{aligned}
&E\left[\sup_{\theta\in\Theta}\left\|\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}g\left(W_{ij},\theta\right)Z_{ij}\left(g\left(W_{ij},\theta\right)Z_{ij}\right)^T-E\left[g\left(W_{11},\theta\right)Z_{11}\left(g\left(W_{11},\theta\right)Z_{11}\right)^T\right]\right\|\right]\\
&\quad=o(1).
\end{aligned}$$

As we can write

$$\begin{aligned}
\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}g\left(W_{ij},\theta\right)Z_{ij}\left(g\left(W_{ij},\theta\right)Z_{ij}\right)^T&=\frac{L}{NM}\frac{1}{L}\sum_{i=1}^{N}\sum_{j=1}^{M}g\left(W_{ij},\theta\right)Z_{ij}\left(g\left(W_{ij},\theta\right)Z_{ij}\right)^T\\
&=p\frac{1}{L}\sum_{i=1}^{N}\sum_{j=1}^{M}g\left(W_{ij},\theta\right)Z_{ij}\left(g\left(W_{ij},\theta\right)Z_{ij}\right)^T
\end{aligned}$$

and

$$\begin{aligned}
E\left[g\left(W_{11},\theta\right)Z_{11}\left(g\left(W_{11},\theta\right)Z_{11}\right)^T\right]&=E\left[Z_{11}^2\right]E\left[g\left(W_{11},\theta\right)g^T\left(W_{11},\theta\right)\right]\\
&=pE\left[g\left(W_{11},\theta\right)g^T\left(W_{11},\theta\right)\right].
\end{aligned}$$

Therefore, by Markov's inequality, it follows that

$$\sup_{\theta\in\Theta}\left\|\frac{1}{L}\sum_{i=1}^{N}\sum_{j=1}^{M}g\left(W_{ij},\theta\right)Z_{ij}\left(g\left(W_{ij},\theta\right)Z_{ij}\right)^{T}-E\left[g\left(W_{11},\theta\right)g^{T}\left(W_{11},\theta\right)\right]\right\|\xrightarrow{P}0.$$

In addition,

$$\frac{1}{L}\sum_{i=1}^{N}\sum_{j=1}^{M}Z_{ij}g\left(W_{ij},\theta\right)gT\left(W_{ij},\theta\right)=\frac{1}{L}\sum_{i=1}^{N}\sum_{j=1}^{M}g\left(W_{ij},\theta\right)Z_{ij}\left(g\left(W_{ij},\theta\right)Z_{ij}\right)^{T}.$$

Thus, Lemma D.2 yields $\widetilde{\Gamma}_2\xrightarrow{P}\Gamma_2(\widehat{\theta})$, where $\Gamma_2(\theta)=E\left[g\left(W_{11},\theta\right)g^{T}\left(W_{11},\theta\right)\right]$. Meanwhile, $\widehat{\theta}\xrightarrow{P}\theta^0$ and we have $\Gamma_2(\widehat{\theta})\xrightarrow{P}\Gamma_2$ by Assumption 5(iii). Therefore, $\widetilde{\Gamma}_2\xrightarrow{P}\Gamma_2$ follows.

Finally, we establish $\widetilde{\Gamma}_1\xrightarrow{P}\Gamma_1$. Note that $\mathcal{G}_{\text{sub}}$ is a pointwise measurable class of functions and that $\sup_Q N\left(\mathcal{G}_{\text{sub}},\|\cdot\|_{Q,2},\epsilon\left\|g_{\sup}\right\|_{Q,2}\right)<\infty$ for any finite discrete measure Q and $\epsilon\in(0,1]$. With $E[g_{\sup}(W_{ij})^2]<\infty$ and Assumptions 1 and 3, and Lemma D.4 yields

$$\begin{aligned}\lim_{\underline{C}\to\infty}E\Bigg[\sup_{\theta\in\Theta}\Bigg\|&\frac{C}{(NM)^2}\sum_{i=1}^{N}\sum_{1\le j,j'\le M}g\left(W_{ij},\theta\right)Z_{ij}\left(g\left(W_{ij'},\theta\right)Z_{ij'}\right)^{T}\\&-E\left[\lambda_1 g(W_{11},\theta)Z_{11}(g\left(W_{12},\theta\right)Z_{12})^{T}\right]\Bigg\|\Bigg]=0.\end{aligned}$$

As we can write

$$\begin{aligned}&\frac{C}{(NM)^2}\sum_{i=1}^{N}\sum_{1\le j,j'\le M}g\left(W_{ij},\theta\right)Z_{ij}\left(g\left(W_{ij'},\theta\right)Z_{ij'}\right)^{T}\\&=\frac{L^2}{(NM)^2}\frac{C}{L^2}\sum_{i=1}^{N}\sum_{1\le j,j'\le M}g\left(W_{ij},\theta\right)Z_{ij}\left(g\left(W_{ij'},\theta\right)Z_{ij'}\right)^{T}\\&=p^2\frac{C}{L^2}\sum_{i=1}^{N}\sum_{1\le j,j'\le M}g\left(W_{ij},\theta\right)Z_{ij}\left(g\left(W_{ij'},\theta\right)Z_{ij'}\right)^{T}\end{aligned}$$

and

$$\begin{aligned}E\left[\lambda_1 g\left(W_{11},\theta\right)Z_{11}\left(g\left(W_{12},\theta\right)Z_{12}\right)^{T}\right]&=E[Z_{11}Z_{12}]E\left[\lambda_1 g\left(W_{11},\theta\right)g^{T}\left(W_{12},\theta\right)\right]\\&=p^2E\left[\lambda_1 g\left(W_{11},\theta\right)g^{T}\left(W_{12},\theta\right)\right].\end{aligned}$$

Therefore, by the Markov inequality, it follows that

$$\sup_{\theta\in\Theta}\left\|\frac{C}{L^2}\sum_{i=1}^{N}\sum_{1\le j,j'\le M}g\left(W_{ij},\theta\right)Z_{ij}\left(g\left(W_{ij'},\theta\right)Z_{ij'}\right)^{T}-\lambda_1 E\left[g\left(W_{11},\theta\right)g^{T}\left(W_{12},\theta\right)\right]\right\|\xrightarrow{P}0,$$

as $\underline{C} \to \infty$. In addition, a symmetric argument also shows that

$$\sup_{\theta\in\Theta}\left\|\frac{C}{L^2}\sum_{1\leq i,i'\leq N}\sum_{j=1}^{M} g\left(W_{ij},\theta\right)Z_{ij}\left(g\left(W_{i'j},\theta\right)Z_{i'j}\right)^T-\lambda_2 E\left[g\left(W_{11},\theta\right)g^T\left(W_{21},\theta\right)\right]\right\| \xrightarrow{P} 0,$$

as $\underline{C} \to \infty$. Also, note that

$$\frac{C}{L^2}\sum_{i=1}^{N}\sum_{1\leq j,j'\leq M} Z_{ij}Z_{ij'}g(W_{ij},\theta)g^T(W_{ij'},\theta)=\frac{C}{L^2}\sum_{i=1}^{N}\sum_{1\leq j,j'\leq M} g(W_{ij},\theta)Z_{ij}(g(W_{ij'},\theta)Z_{ij'})^T$$

and

$$\frac{C}{L^2}\sum_{1\leq i,i'\leq N}\sum_{j=1}^{M} Z_{ij}Z_{i'j}g(W_{ij},\theta)g^T(W_{i'j},\theta)=\frac{C}{L^2}\sum_{1\leq i,i'\leq N}\sum_{j=1}^{M} g(W_{ij},\theta)Z_{ij}(g(W_{i'j},\theta)Z_{i'j})^T$$

hold. Therefore, Lemma D.2 yields $\widetilde{\Gamma}_1 \xrightarrow{P} \Gamma_1(\widehat{\theta})$, where

$$\Gamma_1(\theta)=\lambda_1 E\left[g\left(W_{11},\theta\right)g^T\left(W_{12},\theta\right)\right]+\lambda_2 E\left[g\left(W_{11},\theta\right)g^T\left(W_{21},\theta\right)\right].$$

Meanwhile, since $\widehat{\theta} \xrightarrow{P} \theta^0$, we get $\Gamma_1(\widehat{\theta}) \xrightarrow{P} \Gamma_1$ by Assumption 5(ii). We thus obtain $\widetilde{\Gamma}_1 \xrightarrow{P} \Gamma_1$.

Combining the above results yields $\widetilde{\Omega} \xrightarrow{P} \Omega$ by continuous mapping theorem. By another application of the continuous mapping theorem and the continuity of matrix inversion under Assumption 4(v), we get $(\widetilde{G}^T\widehat{V}\widetilde{G})^{-1} \xrightarrow{P} (G^TVG)^{-1}$. Therefore, $(\widetilde{G}^T\widehat{V}\widetilde{G})^{-1}\widetilde{G}^T\widehat{V}\widetilde{\Omega}\widehat{V}\widetilde{G}(\widetilde{G}^T\widehat{V}\widetilde{G})^{-1}$ is consistent for $(G^TVG)^{-1}G^TV\Omega VG(G^TVG)^{-1}$ by the continuous mapping theorem. □

F. Proofs for the Application to M-Estimation

For convenience, define the class $\mathcal{Q}''=\left\{\partial^2 q(\cdot,\theta)/\partial\theta_r\partial\theta_l : \theta\in\Theta \text{ and } r,l\in\{1,\ldots,k\}\right\}$ of functions indexed by θ, r, and l.

F.1. Proof of Lemma 3

Proof. We verify the conditions of Theorem 2.1 in Newey and McFadden (1994). Condition 2.1(i) that $Q_0(\theta)$ is uniquely maximized at θ^0 follows from the second part in Assumption 6(i). Condition 2.1(ii) holds by the first part in Assumption 6(i). Condition 2.1(iii) that $Q_0(\theta)$ is continuous at θ follows from Assumption 6(ii)(a). Under Assumption 6(i), (ii)(a), (iii), by Example 19.7 in Van der Vaart (2000) and Lemma 9.18 in Kosorok (2008), the class $\mathcal{Q}$ has an envelope $\left|q\left(W_{ij},\theta^0\right)\right|+DM<\infty$, where D is the diameter of a set containing Θ. Thus, for any finite discrete measure Q and $\epsilon\in(0,1]$, $N\left(\mathcal{Q},\|\cdot\|_{Q,2},\epsilon\right)\leq(1+4DM/\epsilon)^k$, which implies that the class $\mathcal{Q}$ satisfies Assumption 2(ii). $\mathcal{Q}$ is a pointwise measurable class of functions by Lemma D.1 under Assumption 6(i), (ii)(a). Thus, with Assumption 1, by applying Lemma 1, we have $\sup_{\theta\in\Theta}\left|\widehat{Q}_{NM}(\theta)-Q_0(\theta)\right| \xrightarrow{P} 0$, so that

 https://doi.org/10.1017/S0266466623000166

condition 2.1(iv) is satisfied. Applying Theorem 2.1 in Newey and McFadden (1994), we obtain $\widehat{\theta} \xrightarrow{P} \theta^0$. □

F.2. Proof of Theorem 4

Proof. We verify the conditions of Theorem 3.1 in Newey and McFadden (1994). Conditions 3.1(i), (ii), and (v) follow from the Assumption 6(i), (ii)(c), (v). Note that the finite function class $\left\{\partial q(\cdot,\theta^0)/\partial\theta_1,\ldots,\partial q(\cdot,\theta^0)/\partial\theta_k\right\}$ is pointwise measurable since $\mathcal{Q}'$ is by Lemma D.1 under Assumption 6(i), (ii)(b). With $E[\dot{q}_{\sup}(W_{ij})^2]<\infty$ under Assumption 6(vi), by applying Theorem 1 under Assumptions 1 and 3, we obtain $\sqrt{\underline{C}}\nabla_\theta \widehat{Q}_{NM}\left(\theta^0\right) \xrightarrow{d} N(0,\Sigma)$, where $\Sigma=\Sigma_1+\Lambda\Sigma_2$,

$$\Sigma_1=\lambda_1 E\left[\nabla_\theta q\left(W_{11},\theta^0\right)\nabla_\theta q\left(W_{12},\theta^0\right)^T\right]+\lambda_2 E\left[\nabla_\theta q\left(W_{11},\theta^0\right)\nabla_\theta q\left(W_{21},\theta^0\right)^T\right],$$
$$\Sigma_2=E\left[\nabla_\theta q\left(W_{11},\theta^0\right)\nabla_\theta q\left(W_{11},\theta^0\right)^T\right].$$

This implies that condition 3.1(iii) is satisfied. Under Assumption 6(i), (ii)(c), similar lines of argument to those in the proof of Lemma 3 yield $N\left(\mathcal{Q}'',\|\cdot\|_{Q,2},\epsilon\right)<\infty$ for any finite discrete measure Q and $\epsilon\in(0,1]$. Note that $\mathcal{Q}''$ is a pointwise measurable class of functions by Lemma D.1 under Assumption 6(i), (ii)(c). With Assumptions 1 and 6(iv), Lemma 1 thus yields

$$\sup_{\theta\in\Theta}\left|\frac{1}{\widehat{L}}\sum_{i=1}^{N}\sum_{j=1}^{M}Z_{ij}\frac{\partial^2 q\left(W_{ij},\theta\right)}{\partial\theta_r\partial\theta_l}-E\left[\frac{\partial^2 q\left(W_{11},\theta\right)}{\partial\theta_r\partial\theta_l}\right]\right|\xrightarrow{P}0$$

for each $r,l\in\{1,\ldots,k\}$. Since there are only finite numbers of r and l, we obtain $\sup_{\theta\in\Theta}\left\|\widehat{H}(\theta)-H(\theta)\right\|\xrightarrow{P}0$, where $\widehat{H}(\theta)=\nabla_{\theta\theta^T}\widehat{Q}_{NM}(\theta)$. With Assumption 6(ii)(c), condition 3.1(iv) is satisfied. Applying Theorem 3.1 in Newey and McFadden (1994), we obtain $\sqrt{\underline{C}}\left(\widehat{\theta}-\theta^0\right)\xrightarrow{d}N\left(0,H^{-1}\Sigma H^{-1}\right)$. □

F.3. Proof of Theorem 5

Proof. First, we want to establish $\widetilde{H}\xrightarrow{P}H$ via $\left\|\widetilde{H}-H\right\|\le\left\|\widetilde{H}-H\left(\widehat{\theta}\right)\right\|+\left\|H\left(\widehat{\theta}\right)-H\right\|$, where $H(\theta)=-E\left[\nabla_{\theta\theta^T}q(W_{11},\theta)\right]$. Since the conditions of Lemma 3 are satisfied, we have $\widehat{\theta}\xrightarrow{P}\theta^0$. Under Assumption 7(i), we thus obtain $\left\|H\left(\widehat{\theta}\right)-H\right\|\xrightarrow{P}0$ by the continuous mapping theorem. Note that $\mathcal{Q}''$ is pointwise measurable and $N\left(\mathcal{Q}'',\|\cdot\|_{Q,2},\epsilon\right)<\infty$ for any finite discrete measure Q, $\epsilon\in(0,1]$ by the proof of Theorem 4. Under Assumptions 1 and 6(iv), by applying Lemma 1, we obtain

$$\sup_{\theta\in\Theta}\left|\frac{1}{\widehat{L}}\sum_{i=1}^{N}\sum_{j=1}^{M}Z_{ij}\frac{\partial^2 q\left(W_{ij},\theta\right)}{\partial\theta_r\partial\theta_l}-E\left[\frac{\partial^2 q\left(W_{11},\theta\right)}{\partial\theta_r\partial\theta_l}\right]\right|\xrightarrow{P}0$$

for each $r,l \in \{1,\dots,k\}$. Since there are only finite numbers of r and l, we get

$$\sup_{\theta\in\Theta}\left\|\frac{1}{\widetilde{L}}\sum_{i=1}^{N}\sum_{j=1}^{M} Z_{ij}\nabla_{\theta\theta^T} q\left(W_{ij},\theta\right) - E\left[\nabla_{\theta\theta^T} q\left(W_{11},\theta\right)\right]\right\| \xrightarrow{P} 0.$$

Thus, we obtain $\widetilde{H} \xrightarrow{P} H\left(\widehat{\theta}\right)$. Combining the above result together yields $\widetilde{H} \xrightarrow{P} H$. By the continuity of matrix inversion under Assumption 6(v), it follows that $\widetilde{H}^{-1} \xrightarrow{P} H^{-1}$. Next, we will establish $\widetilde{\Sigma}_2 \xrightarrow{P} \Sigma_2$. Define a new class $\mathcal{Q}'_{\text{sub}} = \left\{(w,z) \mapsto z\partial q(w,\theta)/\partial\theta_r, \theta\in\Theta, r \in \{1,\dots,k\}\right\}$. Under Assumption 6(i), (ii)(b), similar lines of argument to those in the proof of Lemma 3 yield $\sup_Q N\left(\mathcal{Q}'_{\text{sub}}, \|\cdot\|_{Q,2}, \epsilon\left\|q_{\sup}\right\|_{Q,2}\right) < \infty$, for any finite discrete measure Q and $\epsilon \in (0,1]$. By Theorem 9.15 in Kosorok (2008),

$$\begin{aligned}
&\sup_Q N\left(\mathcal{Q}'_{\text{sub}}\mathcal{Q}'_{\text{sub}}, \|\cdot\|_{Q,2}, \sqrt{2}\epsilon\left\|q^2_{\sup}\right\|_{Q,2}\right)\\
&\leq \sup_Q N\left(\mathcal{Q}'_{\text{sub}}, \|\cdot\|_{Q,2}, \epsilon\left\|q_{\sup}\right\|_{Q,2}\right)\sup_Q N\left(\mathcal{Q}'_{\text{sub}}, \|\cdot\|_{Q,2}, \epsilon\left\|q_{\sup}\right\|_{Q,2}\right) < \infty
\end{aligned}$$

holds for any finite discrete measure Q and $\epsilon \in (0,1]$, where $\mathcal{Q}'_{\text{sub}}\mathcal{Q}'_{\text{sub}}$ is defined as the pointwise product. Note that $\mathcal{Q}'_{\text{sub}}\mathcal{Q}'_{\text{sub}}$ is a pointwise measurable class of functions since $\mathcal{Q}'_{\text{sub}}$ is a pointwise measurable class of functions by the same argument as in the proof of Theorem 4. With Assumption 1 and $E[\dot{q}_{\sup}(W_{ij})^2] < \infty$ under Assumption 6(vi), by applying Lemma D.3, we obtain

$$\begin{aligned}
&E\left[\sup_{\theta\in\Theta}\left\|\frac{1}{L}\sum_{i=1}^{N}\sum_{j=1}^{M}\nabla_\theta q(W_{ij},\theta)Z_{ij}(\nabla_\theta q(W_{ij},\theta)Z_{ij})^T - E[\nabla_\theta q(W_{11},\theta)\nabla_\theta q(W_{11},\theta)^T]\right\|\right]\\
&\quad = o(1).
\end{aligned}$$

Therefore, by the Markov inequality, it follows that

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left\|\frac{1}{L}\sum_{i=1}^{N}\sum_{j=1}^{M}\nabla_\theta q\left(W_{ij},\theta\right)Z_{ij}\left(\nabla_\theta q\left(W_{ij},\theta\right)Z_{ij}\right)^T - E\left[\nabla_\theta q\left(W_{11},\theta\right)\nabla_\theta q\left(W_{11},\theta\right)^T\right]\right\|\\
&\quad \xrightarrow{P} 0.
\end{aligned}$$

In addition, we can write

$$\frac{1}{L}\sum_{i=1}^{N}\sum_{j=1}^{M} Z_{ij}\nabla_\theta q\left(W_{ij},\theta\right)\nabla_\theta q\left(W_{ij},\theta\right)^T = \frac{1}{L}\sum_{i=1}^{N}\sum_{j=1}^{M}\nabla_\theta q\left(W_{ij},\theta\right)Z_{ij}\left(\nabla_\theta q\left(W_{ij},\theta\right)Z_{ij}\right)^T.$$

Thus, Lemma D.2 yields $\widetilde{\Sigma}_2 \xrightarrow{P} \Sigma_2(\widehat{\theta})$, where $\Sigma_2(\theta) = E\left[\nabla_\theta q\left(W_{11},\theta\right)\nabla_\theta q\left(W_{11},\theta\right)^T\right]$. Meanwhile, $\widehat{\theta} \xrightarrow{P} \theta^0$ and we have $\Sigma_2(\widehat{\theta}) \xrightarrow{P} \Sigma_2$ by Assumption 7(iii). Therefore, we establish $\widetilde{\Sigma}_2 \xrightarrow{P} \Sigma_2$.

Finally, we will establish $\widetilde{\Sigma}_1 \xrightarrow{P} \Sigma_1$. Note that $\mathcal{Q}'_{\text{sub}}$ is a pointwise measurable class of functions and $\sup_Q N\left(\mathcal{Q}'_{\text{sub}}, \|\cdot\|_{Q,2}, \epsilon\left\|q_{\sup}\right\|_{Q,2}\right) < \infty$, for any finite discrete measure Q, $\epsilon \in (0,1]$. With Assumptions 1 and 3 and $E[\dot{q}_{sup}(W_{ij})^2] < \infty$ under Assumption 6(vi), Lemma D.4 yields

$$\lim_{\underline{C}\to\infty} E\left[\sup_{\theta\in\Theta}\left\| \frac{\underline{C}}{(NM)^2}\sum_{i=1}^{N}\sum_{1\le j,j'\le M}\nabla_\theta q\left(W_{ij},\theta\right)Z_{ij}\left(\nabla_\theta q\left(W_{ij'},\theta\right)Z_{ij'}\right)^T\right.\right.$$
$$\left.\left. - E\left[\lambda_1\nabla_\theta q(W_{11},\theta)Z_{11}\left(\nabla_\theta q(W_{12},\theta)Z_{12}\right)^T\right]\right\|\right] = 0.$$

As we can write

$$\frac{\underline{C}}{(NM)^2}\sum_{i=1}^{N}\sum_{1\le j,j'\le M}\nabla_\theta q\left(W_{ij},\theta\right)Z_{ij}\left(\nabla_\theta q\left(W_{ij'},\theta\right)Z_{ij'}\right)^T$$
$$=\frac{L^2}{(NM)^2}\frac{\underline{C}}{L^2}\sum_{i=1}^{N}\sum_{1\le j,j'\le M}\nabla_\theta q\left(W_{ij},\theta\right)Z_{ij}\left(\nabla_\theta q\left(W_{ij'},\theta\right)Z_{ij'}\right)^T$$
$$=p^2\frac{\underline{C}}{L^2}\sum_{i=1}^{N}\sum_{1\le j,j'\le M}\nabla_\theta q\left(W_{ij},\theta\right)Z_{ij}\left(\nabla_\theta q\left(W_{ij'},\theta\right)Z_{ij'}\right)^T$$

and

$$E\left[\lambda_1\nabla_\theta q(W_{11},\theta)Z_{11}\left(\nabla_\theta q(W_{12},\theta)Z_{12}\right)^T\right]$$
$$=E[Z_{11}Z_{12}]E\left[\lambda_1\nabla_\theta q(W_{11},\theta)\nabla_\theta q(W_{12},\theta)^T\right]$$
$$=p^2E\left[\lambda_1\nabla_\theta q(W_{11},\theta)\nabla_\theta q(W_{12},\theta)^T\right].$$

Therefore, by the Markov inequality, it follows that

$$\sup_{\theta\in\Theta}\left\|\frac{\underline{C}}{L^2}\sum_{i=1}^{N}\sum_{1\le j,j'\le M}\nabla_\theta q\left(W_{ij},\theta\right)Z_{ij}\left(\nabla_\theta q\left(W_{ij'},\theta\right)Z_{ij'}\right)^T\right.$$
$$\left.-\lambda_1E\left[\nabla_\theta q(W_{11},\theta)\nabla_\theta q(W_{12},\theta)^T\right]\right\|\xrightarrow{P}0,$$

as $\underline{C}\to\infty$. In addition, a symmetric argument also shows that

$$\sup_{\theta\in\Theta}\left\|\frac{\underline{C}}{L^2}\sum_{1\le i,i'\le N}\sum_{j=1}^{M}\nabla_\theta q\left(W_{ij},\theta\right)Z_{ij}\left(\nabla_\theta q\left(W_{i'j},\theta\right)Z_{i'j}\right)^T\right.$$
$$\left.-\lambda_2E\left[\nabla_\theta q(W_{11},\theta)\nabla_\theta q(W_{21},\theta)^T\right]\right\|\xrightarrow{P}0,$$

as $\underline{C}\to\infty$. Also, we can write

$$\frac{\underline{C}}{L^2}\sum_{i=1}^{N}\sum_{1\le j,j'\le M}Z_{ij}Z_{ij'}\nabla_\theta q\left(W_{ij},\theta\right)\nabla_\theta q\left(W_{ij'},\theta\right)^T$$
$$=\frac{\underline{C}}{L^2}\sum_{i=1}^{N}\sum_{1\le j,j'\le M}\nabla_\theta q\left(W_{ij},\theta\right)Z_{ij}\left(\nabla_\theta q\left(W_{ij'},\theta\right)Z_{ij'}\right)^T$$

and

$$\frac{C}{L^2}\sum_{1\le i,i'\le N}\sum_{j=1}^{M} Z_{ij}Z_{i'j}\nabla_\theta q\left(W_{ij},\theta\right)\nabla_\theta q\left(W_{i'j},\theta\right)^T$$
$$=\frac{C}{L^2}\sum_{1\le i,i'\le N}\sum_{j=1}^{M}\nabla_\theta q\left(W_{ij},\theta\right)Z_{ij}\left(\nabla_\theta q\left(W_{i'j},\theta\right)Z_{i'j}\right)^T.$$

Therefore, Lemma D.2 yields $\widetilde{\Sigma}_1 \xrightarrow{P} \Sigma_1\left(\widehat{\theta}\right)$, where

$$\Sigma_1(\theta)=\lambda_1 E\left[\nabla_\theta q(W_{11},\theta)\nabla_\theta q(W_{12},\theta)^T\right]+\lambda_2 E\left[\nabla_\theta q(W_{11},\theta)\nabla_\theta q(W_{21},\theta)^T\right].$$

Meanwhile, since $\widehat{\theta}\xrightarrow{P}\theta^0$, we get $\Sigma_1\left(\widehat{\theta}\right)\xrightarrow{P}\Sigma_1$ by Assumption 7(ii). We thus obtain $\widetilde{\Sigma}_1\xrightarrow{P}\Sigma_1$ and $\widetilde{\Sigma}\xrightarrow{P}\Sigma$ by continuous mapping theorem. Therefore, $\widetilde{H}^{-1}\widetilde{\Sigma}\widetilde{H}^{-1}$ is consistent for $H^{-1}\Sigma H^{-1}$ by the continuous mapping theorem. □

G. Multiple Observations in a Cluster

In many empirical applications, researchers face situations in which there are multiple observations in some cluster (i,j), and the number of observations may vary across the clusters. In this section, we generalize the results from the main text to accommodate multiple observations per cluster. We shall follow the basic setting in Davezies et al. (2018). Throughout this section, we call a pair (i,j) of indices a cell. The number of observations in a cell is allowed to be random and can be correlated with the observations. This allows for a wide range of heterogeneous cluster structures. We will denote the number of observations in the (i,j)th cell by N_{ij}, which is itself a random variable that takes a nonnegative integer value. The observation that corresponds to the ℓth unit, $1\le\ell\le N_{ij}$, in the (i,j)th cell is a d-dimensional random vector denoted by $W_{\ell,ij}$. With these notations, we consider the following sampling assumption.

Assumption G.1 (Sampling).

(i) The array $\left(N_{ij},(W_{\ell,ij})_{\ell\ge1}\right)_{(i,j)\in\mathbb{N}^2}$ is separately exchangeable.
(ii) $\left(N_{ij},(W_{\ell,ij})_{\ell\ge1}\right)_{(i,j)\in\mathbb{N}^2}$ is dissociated.
(iii) $E[N_{11}]>0$.

This assumption is essentially identical to Assumption 1 in Davezies et al. (2018). Parts (i) and (ii) parallel Assumption 1(i) and (ii), respectively, in the main text except that the cell size is random and can differ across cells here. When $N_{ij}=1$ for all i and j, the conditions reduce to Assumption 1 in the main text. Also, part (ii) allows for a wide range of correlation structures between N_{ij} and $(W_{\ell,ij})_{\ell\ge1}$, and among $(W_{\ell,ij})_{\ell\ge1}$ within (i,j)th cell. Part (iii) excludes the cells that are almost surely empty.

Assumption G.2 (Function class).

(i) $E\left[\sum_{\ell=1}^{N_{ij}}f\left(W_{\ell,ij}\right)\right]=0$.

(ii) $\mathcal{F}$ admits an envelope F satisfying $E\left[\sum_{\ell=1}^{N_{ij}} F(W_{\ell,ij})\right] < \infty$ with $\sup_Q N(\mathcal{F}, \|\cdot\|_{Q,2}, \epsilon\|F\|_{Q,2}) < \infty$ for all $\epsilon > 0$, where Q is any finite discrete measure.

(iii) $\mathcal{F}$ is pointwise measurable.

This assumption generalizes Assumption 2 by allowing for multiple observations within a cell as well as heterogeneous cell sizes.

LEMMA G.1 (Uniform weak law of large numbers). *Suppose that Assumptions G.1 and G.2(ii), (iii) hold. Then we have*

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{\widehat{L}}\sum_{i=1}^{N}\sum_{j=1}^{M}Z_{ij}\sum_{\ell=1}^{N_{ij}}f\left(W_{\ell,ij}\right)-E\left[\sum_{\ell=1}^{N_{11}}f(W_{\ell,11})\right]\right|\xrightarrow{P}0.$$

A proof of Lemma G.1 is a straightforward modification of the proof of Lemma 1 and is therefore omitted.

THEOREM G.1 (Central limit theorem). *Suppose that Assumptions 3, G.1, and G.2(i), (iii) hold. In addition, suppose that any finite function class $\mathcal{F} = \{f_1,\dots,f_k\}$ with a fixed k admits an envelope F satisfying $E\left[\left(\sum_{\ell=1}^{N_{ij}} F(W_{\ell,ij})\right)^2\right] < \infty$. Let $f = (f_1,\dots,f_k)^T$. Then,*

$$\sqrt{\underline{C}}\frac{1}{\widehat{L}}\sum_{i=1}^{N}\sum_{j=1}^{M}Z_{ij}\sum_{\ell=1}^{N_{ij}}f\left(W_{\ell,ij}\right)\xrightarrow{d}N\left(0,\Gamma\right),$$

where $\Gamma = \Gamma_A + \Lambda\Gamma_B$,

$$\Gamma_A=\lambda_1E\left[\left(\sum_{\ell=1}^{N_{11}}f\left(W_{\ell,11}\right)\right)\left(\sum_{\ell=1}^{N_{12}}f\left(W_{\ell,12}\right)\right)^T\right]$$
$$+\lambda_2E\left[\left(\sum_{\ell=1}^{N_{11}}f\left(W_{\ell,11}\right)\right)\left(\sum_{\ell=1}^{N_{21}}f\left(W_{\ell,21}\right)\right)^T\right],$$
$$\Gamma_B=E\left[\left(\sum_{\ell=1}^{N_{11}}f\left(W_{\ell,11}\right)\right)\left(\sum_{\ell=1}^{N_{11}}f\left(W_{\ell,11}\right)\right)^T\right].$$

Again, a proof is a straightforward modification of that of Theorem 1. Here, we describe the necessary modification without repetitively showing the entire proof. Note that under the current setting, we have

$$\frac{1}{\widehat{L}}\sum_{i=1}^{N}\sum_{j=1}^{M}Z_{ij}\sum_{\ell=1}^{N_{ij}}f\left(W_{\ell,ij}\right)=\frac{L}{\widehat{L}}\left(A_{NM}+\sqrt{1-p_{NM}}B_{NM}\right),$$

where A_{NM} and B_{NM} are defined as

$$A_{NM} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{\ell=1}^{N_{ij}} f\left(W_{\ell,ij}\right) \quad \text{and} \quad B_{NM} = \frac{1}{L}\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{Z_{ij}-p_{NM}}{\sqrt{1-p_{NM}}}\sum_{\ell=1}^{N_{ij}} f\left(W_{\ell,ij}\right),$$

respectively. Then, using a similar argument to the one in the proof of Theorem 1, the asymptotic normality for A_{NM} can be established as

$$\sqrt{\underline{C}}A_{NM} \xrightarrow{d} N\left(0, \lambda_1 E\left[\left(\sum_{\ell=1}^{N_{11}} f\left(W_{\ell,11}\right)\right)\left(\sum_{\ell=1}^{N_{12}} f\left(W_{\ell,12}\right)\right)^T\right]\right.$$
$$\left. + \lambda_2 E\left[\left(\sum_{\ell=1}^{N_{11}} f\left(W_{\ell,11}\right)\right)\left(\sum_{\ell=1}^{N_{21}} f\left(W_{\ell,21}\right)\right)^T\right]\right).$$

Similarly, the variance–covariance matrices for B_{NM} can be calculated as

$$\mathrm{Var}\left(B_{NM}\right) = \frac{1}{L}E\left[\left(\sum_{\ell=1}^{N_{11}} f\left(W_{\ell,11}\right)\right)\left(\sum_{\ell=1}^{N_{11}} f\left(W_{\ell,11}\right)\right)^T\right].$$

We thus obtain Theorem G.1 following from the arguments in the proof of Theorem 1.

G.1. Application to Generalized Method of Moments

We now generalize the results for GMM from Section 3 to allow for multiple observations per cell. Under the current setting, we assume that the true parameter vector $\theta^0 = (\theta_1^0, \ldots, \theta_k^0)^T$ satisfies

$$E\left[\sum_{\ell=1}^{N_{ij}} g\left(W_{\ell,ij}, \theta^0\right)\right] = 0,$$

where $m \geq k$. Multiway algorithmic subsampling GMM estimator $\widehat{\theta}$ can be subsequently defined as

$$\max_{\theta\in\Theta} \widehat{Q}_{NM}(\theta),$$

where $\widehat{Q}_{NM}(\theta) = -\widehat{g}_{NM}(\theta)^T \widehat{V}\widehat{g}_{NM}(\theta)$, $\widehat{g}_{NM}(\theta) = \widehat{L}^{-1}\sum_{i=1}^{N}\sum_{j=1}^{M} Z_{ij}\sum_{\ell=1}^{N_{ij}} g\left(W_{\ell,ij}, \theta\right)$. To establish the asymptotic properties of $\widehat{\theta}$, we impose the following conditions.

Assumption G.3.

(i) V is positive semi-definite, and $VE\left[\sum_{\ell=1}^{N_{ij}} g(W_{\ell,ij}, \theta)\right] = 0$ iff $\theta = \theta^0$.
(ii) $\theta^0 \in \mathrm{int}(\Theta)$, where Θ is a compact subset of $\mathbb{R}^k$.
(iii) (a) $\theta \mapsto g_r(w,\theta)$ is Lipschitz with a universal Lipschitz constant.
  (b) Each component of $\theta \mapsto \nabla_\theta g_r(w,\theta)$ is Lipschitz with a universal Lipschitz constant.
(iv) $E\left[\sup_{\theta\in\Theta}\left\|\sum_{\ell=1}^{N_{ij}} g\left(W_{\ell,ij}, \theta\right)\right\|\right] < \infty$.

(v) $G^T VG$ is nonsingular, where $G = E\left[\sum_{\ell=1}^{N_{ij}} \nabla_\theta g\left(W_{\ell,ij},\theta^0\right)\right]$.

(vi) $E\left[\sup_{\theta\in\Theta}\left\|\sum_{\ell=1}^{N_{ij}} \nabla_\theta g\left(W_{\ell,ij},\theta\right)\right\|\right] < \infty$.

(vii) $g_{\sup}(\cdot) = \max_{r\in\{1,\ldots,m\}} |g_r(\cdot,\theta)|$ satisfies $E\left[\left(\sum_{\ell=1}^{N_{ij}} g_{\sup}(W_{\ell,ij})\right)^2\right] < \infty$.

LEMMA G.2 (Consistency of multiway algorithmic subsampling GMM estimator). *If Assumptions G.1 and G.3(i)–(iv) hold, and that $\widehat{V} \xrightarrow{P} V$, then $\widehat{\theta} \xrightarrow{P} \theta^0$.*

A proof of Lemma G.2 follows analogously from the proof of Lemma 2 and an application of Lemma G.1. We omit the proof.

THEOREM G.2 (Asymptotic normality of multiway algorithmic subsampling GMM estimator). *If Assumptions 3, G.1, and G.3 hold, and that $\widehat{V} \xrightarrow{P} V$, then*

$$\sqrt{\underline{C}}\left(\widehat{\theta}-\theta^0\right) \xrightarrow{d} N\left(0, \left(G^T VG\right)^{-1} G^T V\Omega VG\left(G^T VG\right)^{-1}\right),$$

where $G = E\left[\sum_{\ell=1}^{N_{11}} \nabla_\theta g\left(W_{\ell,11},\theta^0\right)\right]$ *and* $\Omega = \Gamma_1 + \Lambda\Gamma_2$, *with*

$$\Gamma_1 = \lambda_1 E\left[\left(\sum_{\ell=1}^{N_{11}} g\left(W_{\ell,11},\theta^0\right)\right)\left(\sum_{\ell=1}^{N_{12}} g\left(W_{\ell,12},\theta^0\right)\right)^T\right]$$
$$+\lambda_2 E\left[\left(\sum_{\ell=1}^{N_{11}} g\left(W_{\ell,11},\theta^0\right)\right)\left(\sum_{\ell=1}^{N_{21}} g\left(W_{\ell,21},\theta^0\right)\right)^T\right]$$

and $\Gamma_2 = E\left[\left(\sum_{\ell=1}^{N_{11}} g\left(W_{\ell,11},\theta^0\right)\right)\left(\sum_{\ell=1}^{N_{11}} g\left(W_{\ell,11},\theta^0\right)\right)^T\right]$.

A proof of Theorem G.2 follows straightforwardly from the proof of Theorem 2 and an application of Theorem G.1 in place of Theorem 1. To avoid repetition, the proof is omitted.

G.2. Application to M-Estimation

We now generalize the results for M-Estimation from Section 4 to allow for multiple observations per cell. Under this setting, multiway algorithmic subsampling M-estimator $\widehat{\theta}$ is defined as the solution to

$$\max_{\theta\in\Theta} -\frac{1}{\widehat{L}}\sum_{i=1}^{N}\sum_{j=1}^{M} Z_{ij}\sum_{\ell=1}^{N_{ij}} q\left(W_{\ell,ij},\theta\right).$$

The true parameter vector θ^0 is assumed to be the unique solution of $\max_{\theta\in\Theta} -E[\sum_{\ell=1}^{N_{ij}} q(W_{\ell,ij},\theta)]$. For each $\theta\in\Theta$, let $-\widehat{L}^{-1}\sum_{i=1}^{N}\sum_{j=1}^{M} Z_{ij}\sum_{\ell=1}^{N_{ij}} q(W_{\ell,ij},\theta)$ and $-E[\sum_{\ell=1}^{N_{ij}} q(W_{\ell,ij},\theta)]$ be denoted by $\widehat{Q}_{NM}(\theta)$ and $Q_0(\theta)$, respectively. To establish the asymptotic properties of $\widehat{\theta}$, we restate Assumption 6 as follows.

Assumption G.4.

(i) $\theta^0 \in \operatorname{int}(\Theta)$, where Θ is a compact subset of $\mathbb{R}^k$. Also, $E\left[\sum_{\ell=1}^{N_{ij}} q(W_{\ell,ij},\theta^0)\right] < E\left[\sum_{\ell=1}^{N_{ij}} q(W_{\ell,ij},\theta)\right]$ holds for all $\theta \in \Theta \backslash \{\theta_0\}$.

(ii) (a) $\theta \mapsto q(w,\theta)$ is Lipschitz with a universal Lipschitz constant.
(b) Each coordinate of $\theta \mapsto \nabla_\theta q(w,\theta)$ is Lipschitz with a universal Lipschitz constant.
(c) Each coordinate of $\theta \mapsto \nabla_{\theta\theta^T} q(w,\theta) = \partial^2 q(w,\theta)/\partial\theta\partial\theta^T$ is Lipschitz with a universal Lipschitz constant.

(iii) $E[\sup_{\theta\in\Theta} \sum_{\ell=1}^{N_{ij}} q\left(W_{\ell,ij},\theta\right)] < \infty$.

(iv) $E\left[\sup_{\theta\in\Theta}\left\|\sum_{\ell=1}^{N_{ij}} \nabla_{\theta\theta^T} q\left(W_{\ell,ij},\theta\right)\right\|\right] < \infty$.

(v) $H = H\left(\theta^0\right)$ is nonsingular, where $H(\theta) = -E\left[\sum_{\ell=1}^{N_{ij}} \nabla_{\theta\theta^T} q\left(W_{\ell,ij},\theta\right)\right]$.

(vi) $\dot{q}_{\sup}(\cdot) = \max_{r\in\{1,\dots,k\}} |\partial q(\cdot,\theta)/\partial\theta_r|$ satisfies $E\left[\left(\sum_{\ell=1}^{N_{ij}} \dot{q}_{\sup}(W_{\ell,ij})\right)^2\right] < \infty$.

LEMMA G.3 (Consistency of multiway algorithmic subsampling M-estimator). *If Assumptions G.1 and G.4(i)–(iii) hold, then,* $\widehat{\theta} \xrightarrow{P} \theta^0$.

A proof of Lemma G.3 follows analogously from the proof of Lemma 3 and an application of Lemma G.1. We omit the proof.

THEOREM G.3 (Asymptotic normality of multiway algorithmic subsampling M-estimator). *If Assumptions 3, G.1, and G.4 hold, then*

$$\sqrt{\underline{C}}\left(\widehat{\theta}-\theta^0\right) \xrightarrow{d} N\left(0, H^{-1}\Sigma H^{-1}\right),$$

where $H = -E\left[\sum_{\ell=1}^{N_{11}} \nabla_{\theta\theta^T} q\left(W_{\ell,11},\theta^0\right)\right]$, $\Sigma = \Sigma_1 + \Lambda\Sigma_2$,

$$\begin{aligned}\Sigma_1 =& \lambda_1 E\left[\left(\sum_{\ell=1}^{N_{11}} \nabla_\theta q\left(W_{\ell,11},\theta^0\right)\right)\left(\sum_{\ell=1}^{N_{12}} \nabla_\theta q\left(W_{\ell,12},\theta^0\right)\right)^T\right] \\ &+ \lambda_2 E\left[\left(\sum_{\ell=1}^{N_{11}} \nabla_\theta q\left(W_{\ell,11},\theta^0\right)\right)\left(\sum_{\ell=1}^{N_{21}} \nabla_\theta q\left(W_{\ell,21},\theta^0\right)\right)^T\right],\end{aligned}$$

and $\Sigma_2 = E\left[\left(\sum_{\ell=1}^{N_{11}} \nabla_\theta q\left(W_{\ell,11},\theta^0\right)\right)\left(\sum_{\ell=1}^{N_{11}} \nabla_\theta q\left(W_{\ell,11},\theta^0\right)\right)^T\right]$.

A proof of Theorem G.3 follows straightforwardly from the proof of Theorem 4 and an application of Theorem G.1 in place of Theorem 1. To avoid repetition, the proof is omitted.

H. Alternative Subsampling Methods

Although we have thus far focused on Bernoulli subsampling as the default subsampling method, other algorithmic subsampling schemes are also applicable. In Lee and Ng (2020b), two classes of algorithmic subsampling schemes are considered, namely, random subsampling methods and random projection methods, with the Bernoulli subsampling considered throughout this paper belonging to the former category. It is possible to adapt other random subsampling methods under multiway cluster sampling setting, such as uniform sampling with/without replacement. In fact, the robustness against possible degeneracy remains valid when either of these two alternative random subsampling schemes is substituted. We are going to illustrate such adaptations in the rest of this section. On the other hand, as random projection-based methods produce rather different decompositions in the asymptotic terms, their validity and asymptotic behaviors under the current setting remain unclear to us, and are therefore not discussed here.

To implement uniform subsampling without replacement, the researcher sets L randomly chosen $\{Z_{ij} : i = 1,\ldots,N, j = 1,\ldots,M\}$ to 1 and the rest to 0. To implement uniform subsampling with replacement, the researcher generates $\{Z_{ij} : i = 1,\ldots,N, j = 1,\ldots,M\}$ following a multinomial distribution with L trials, mutually exclusive events $\{1,\ldots,NM\}$ and equal event probabilities $1/(NM)$. For both subsampling schemes, the total number of subsampled units, $\sum_{i=1}^{N}\sum_{j=1}^{M} Z_{ij} = L$, is deterministic, while for the Bernoulli random sampling, $\sum_{i=1}^{N}\sum_{j=1}^{M} Z_{ij} = \widehat{L}$ is stochastic. Despite such a discrepancy, the uniform subsampling without replacement yields asymptotically the same result as Bernoulli subsampling, since Bernoulli subsampling can be considered as a uniform subsampling without replacement with random sample size and also $\widehat{L}/L \overset{P}{\to} 1$ due to Lemma D.2. We state the following two propositions for the uniform subsampling with and without replacement.

Proposition H.1. *Consider the uniform subsampling without replacement. If the conditions for Theorem 1 hold, then*

$$\sqrt{\underline{C}}\frac{1}{L}\sum_{i=1}^{N}\sum_{j=1}^{M} Z_{ij}f\left(W_{ij}\right) \overset{d}{\to} N\left(0,\Gamma_{UN}\right), \tag{H.1}$$

where Γ_{UN} has the same form as Γ in Theorem 1.

Proposition H.2. *Consider the uniform subsampling with replacement. If the conditions for Theorem 1 hold, then*

$$\sqrt{\underline{C}}\frac{1}{L}\sum_{i=1}^{N}\sum_{j=1}^{M} Z_{ij}f\left(W_{ij}\right) \overset{d}{\to} N\left(0,\Gamma_{UR}\right), \tag{H.2}$$

where $\Gamma_{UR} = \Gamma_A + \lim_{N,M\to\infty}(\underline{C}/(NMp))\Gamma_B$, where Γ_A and Γ_B are as defined in Theorem 1.

Proofs of Propositions H.1 and H.2 follow from a straightforward adaptation of arguments in the proof of Theorem 1 with Lemma 2 and Theorem 1 of Janson (1984). Here, we describe the required modifications rather than reproducing the repetitive proofs. First, note that under either of these subsampling schemes, one can proceed with the following proof of Theorem 1 to obtain the decomposition of equation (C.1) with the factor $L/\widehat{L}$ replaced by 1. In addition, the B_{NM} term in the decomposition of equation (C.1) has a different conditional (on observations) distribution that depends on the subsampling scheme and thus a different conditional variance. The propositions then follow from calculating the alternative conditional variance of B_{NM} and applying the law of total variance.

REFERENCES

Cameron, A.C., J.B. Gelbach, & D.L. Miller (2011) Robust inference with multiway clustering. *Journal of Business & Economic Statistics* 29, 238–249.

Cameron, A.C. & D.L. Miller (2014) *Robust Inference for Dyadic Data*. University of California-Davis.

Chen, J. & J. Rao (2007) Asymptotic normality under two-phase sampling designs. *Statistica Sinica* 17, 1047–1064.

Chiang, H.D., K. Kato, Y. Ma, & Y. Sasaki (2022) Multiway cluster robust double/debiased machine learning. *Journal of Business & Economic Statistics* 40, 1046–1056.

Chiang, H.D., K. Kato, & Y. Sasaki (2021) Inference for high-dimensional exchangeable arrays. *Journal of the American Statistical Association*, forthcoming.

Davezies, L., X. D'Haultfoeuille, & Y. Guyonvarch (2018) Asymptotic results under multiway clustering. Preprint, arXiv:1807.07925.

Davezies, L., X. D'Haultfoeuille, & Y. Guyonvarch (2021) Empirical process results for exchangeable arrays. *Annals of Statistics* 49, 845–862.

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

Janson, S. (1984) The asymptotic distributions of incomplete U-statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 66, 495–505.

Kosorok, M.R. (2008) *Introduction to Empirical Processes and Semiparametric Inference*. Springer.

Lee, S. & S. Ng (2020a) An econometric perspective on algorithmic subsampling. *Annual Review of Economics* 12, 45–80.

Lee, S. & S. Ng (2020b) Sketching for two-stage least squares estimation. Preprint, arXiv:2007.07781.

MacKinnon, J.G., M.Ø. Nielsen, & M. Webb (2023) Testing for the appropriate level of clustering in linear regression models. *Journal of Econometrics*, forthcoming.

MacKinnon, J.G., M.Ø. Nielsen, & M.D. Webb (2021) Wild bootstrap and asymptotic inference with multiway clustering. *Journal of Business & Economic Statistics* 39, 505–519.

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

MacKinnon, J.G. & M.D. Webb (2018) The wild bootstrap for few (treated) clusters. *Econometrics Journal* 21, 114–135.

Menzel, K. (2021) Bootstrap with cluster-dependence in two or more dimensions. *Econometrica* 89(5), 2143–2188.

Nevo, A. (2000) A practitioner's guide to estimation of random-coefficients logit models of demand. *Journal of Economics & Management Strategy* 9, 513–548.

Newey, W.K. & D. McFadden (1994) *Handbook of Econometrics , Volume IV*, R.F. Engle and D.L. McFadden (eds.), pp. 2111–2245. Elsevier Science.

Thompson, S.B. (2011) Simple formulas for standard errors that cluster by both firm and time. *Journal of Financial Economics* 99, 1–10.

Van der Vaart, A.W. (2000) *Asymptotic Statistics*, vol. 3. Cambridge University Press.

van der Vaart, A.W. & J.A. Wellner (1996) *Weak Convergence and Empirical Processes*. Springer.