General Conditions for Valid Inference in Multi-Way Clustering

Luther Yap *

January 10, 2023

Abstract

This paper proves a new central limit theorem for a sample that exhibits multi-way dependence and heterogeneity across clusters. Statistical inference for situations where there is both multi-way dependence and cluster heterogeneity has thus far been an open issue. Existing theory for multi-way clustering inference requires identical distributions across clusters (implied by the so-called separate exchangeability assumption). Yet no such homogeneity requirement is needed in the existing theory for one-way clustering. The new result therefore theoretically justifies the view that multi-way clustering is a more robust version of one-way clustering, consistent with applied practice. The result is applied to linear regression, where it is shown that a standard plug-in variance estimator is valid for inference.

1 Introduction

Clustering standard errors on multiple dimensions is common and attractive in applied econometrics because it allows observations to be dependent whenever they share a cluster on any dimension.[1] The variance estimator proposed by Cameron et al. (2011) (henceforth CGM) has thus been widely applied to contexts with multi-way dependence. Existing justification for the asymptotic validity of the CGM estimator and other inference procedures in multi-way clustering relies on separate

*Department of Economics, Princeton University. Email: lyap@princeton.edu.

[1] E.g., Dube et al. (2010) clustered on state and border segment when studying the effect of minimum wages on employment; Nunn and Wantchekon (2011) clustered on ethnic groups and district when studying the effect of slave trade on trust; Michalopoulos and Papaioannou (2013) clustered on country and ethnolinguistic family when studying the effect of pre-colonial institutions on development.

exchangeability, which implies the homogeneity of clusters. This paper provides general conditions such that the plug-in mean estimator is asymptotically normal, and the CGM variance estimator is consistent, even when clusters are heterogeneous. These conditions do not include separate exchangeability, and they mimic the conditions in one-way clustering: the only substantive assumption is that two observations are independent when they do not share any cluster. Since asymptotic normality and consistent variance estimation are sufficient for valid inference, the results in this paper provide sufficient general conditions for valid inference in multi-way clustering.

An environment with multi-way clustering permits dependence whenever observations share at least one cluster. To fix ideas, suppose observations can be partitioned on two different dimensions — state and industry. Observations in the same state or in the same industry are plausibly correlated, but two observations in different states and different industries are assumed to be independent.[2] The CGM variance estimator accommodates such dependence, and subsequent literature provided a theoretical basis for its validity (e.g., Davezies et al. (2021); MacKinnon et al. (2021)). Menzel (2021) also showed the validity of a bootstrap procedure for multi-way clustering that is robust to asymptotic non-normalities.[3] The theoretical basis for inference thus far relies on separate exchangeability, the assumption that random variables are exchangeable on either clustering dimension, though not necessarily both.

However, as noted by MacKinnon et al. (2021), separate exchangeability implies identical marginal distributions. Since exchangeability implies identical distribution, separate exchangeability in the state-industry example implies that random variable in Alaska and California must be drawn from the same distribution. In contrast, existing asymptotic theory on one-way clustering (e.g., Hansen and Lee (2019); Djogbenou et al. (2019)) allows the distribution of the random variable to be heterogeneous over clusters. The only substantive assumption is that observations that do not share any cluster are independent. Since the only available conditions for the validity of multi-way clustering require separate exchangeability, the literature lacks general conditions for multi-way clustering that generalize one-way clustering and permit heterogeneity over clusters. This paper fills the gap, and thus justifies multi-way clustering as a more robust version of one-way clustering.

---

[2]This setting permits more general dependence structures than one-way clustering. If there is one-way clustering by state, then two observations from different states are automatically independent. In two-way clustering, two observations from different states are not necessarily independent because they may share the same industry.

[3]Menzel (2021) pointed out that a purely interactive data-generating processes unique to multi-way dependence has an asymptotic distribution that is not normal. Section 2 will consider this process and show how the assumptions of this paper rules it out.

Example 1. *To illustrate separate exchangeability, consider an additive random effects model. Individual i who belongs to cluster $g(i)$ on the G dimension and cluster $h(i)$ on the H dimension has random variable W_i generated from $W_i = \alpha_{g(i)} + \gamma_{h(i)} + \varepsilon_i$, where cluster-specific α_g, γ_h and individual-specific ε_i are independent of each other. If we assume separate exchangeability, then α_g, γ_h, and ε_i are iid.*[4] *In contrast, under one-way cluster asymptotics, the cluster-specific error α_g is allowed to be heteroskedastic. General conditions provided in this paper permits valid inference even when $\alpha_g, \gamma_h, \varepsilon_i$ are heteroskedastic in this model.*

The main result is a central limit theorem for multi-way clustering with heterogeneous cluster sizes and distributions. I apply the theorem to a simple setting of a linear regression, but it is more broadly applicable to many other econometric procedures that exhibit a similar clustering structure.

2 Setting and Main Result

Consider a setup with two-way clustering on dimensions G and H for random vectors $\{W_i\}_{i=1}^n$, where $W_i := (W_{i1}, W_{i2}, \cdots, W_{iK})' \in \mathbb{R}^K$ and i is the unit of observation, for a sequence of populations of size n.[5] For example, G could denote states and H denote industries. This section establishes a central limit theorem (CLT) for a weighted sum of the random vector i.e., $\sum_i \omega_i W_i$, where ω_i are nonstochastic scalar weights, as $n \to \infty$. For $C \in \{G, H\}$, let $\mathcal{N}_c^C$ denote the set of observations in cluster c on dimension C — this partitions the population on the C dimension.

Let $g(i)$ and $h(i)$ denote the cluster that observation i belongs to on the G and H dimensions respectively. These cluster identities are nonstochastic and observed. Let $N_c^C = |\mathcal{N}_c^C|$ denote the cluster size for $C \in \{G, H\}$ and $N_{gh} := |\mathcal{N}_g^G \cap \mathcal{N}_h^H|$. These cluster sizes are allowed to be heterogeneous in a way that will be formalized in the assumptions below. W_i is assumed to be independent of any W_j when $j \notin \mathcal{N}_{g(i)}^G \cup \mathcal{N}_{h(i)}^H =: \mathcal{N}_i$, i.e., when i and j do not share a cluster on either dimension. Hence, $\mathcal{N}_i$ is the set of observations plausibly dependent with i. This environment is stated as Assumption 1, the main substantive assumption.

Assumption 1. *$W_i \perp\!\!\!\perp W_j$ if $g(i) \neq g(j)$ and $h(i) \neq h(j)$.*

[4] To see this, for individuals i and j where $g(i) \neq g(j)$, $h(i) = h(j) = h$, separate exchangeability implies $\alpha_{g(i)} + \gamma_h + \varepsilon_i \overset{d}{=} \alpha_{g(j)} + \gamma_h + \varepsilon_j$. Since α_g, γ_h and ε_i are independent, $\varepsilon_i \overset{d}{=} \varepsilon_j$ and $\alpha_g \overset{d}{=} \alpha_{g'}$.

[5] Clustering in more than two dimensions is possible, and derivations are entirely analogous.

Assumption 1 is agnostic about the dependence structure when W_i and W_j share at least one cluster. It also allows the data generating process to be arbitrarily heterogeneous across different clusters, mimicking the heterogeneity permitted in one-way clustering (e.g., Hansen and Lee (2019)). Since one-way clustering is a special case of two-way clustering where everyone is in their own H cluster, the result here generalizes existing results in one-way clustering. In contrast, existing literature in multi-way clustering assumes separate exchangeability that additionally imposes identical distribution over clusters, so they do not immediately generalize one-way clustering. $\{W_i\}_{i=1}^n$ being separately exchangeable implies Assumption 1 but the converse is not true.[6]

Observations that share a cluster are allowed to be dependent, but they need not be. Hence, let $A_{ij} := 1[W_i \not\perp\!\!\!\perp W_j]$ be a 0-1 indicator for whether W_i and W_j are actually dependent, so $A_{ij} = A_{ji}$, and $A_{ii} = 1$.[7] This notation allows a particular form of misspecification where the researcher is conservative and clusters on dimension G when it is not required. Every observation W_i is weighted by nonstochastic scalar ω_i. For positive definite matrix Q, let $\lambda_{\min}(Q)$ denote the smallest eigenvalue of Q. Then, let $Q_n := Var\left(\sum_{i=1}^n \omega_i W_i\right)$ denote the variance of the sum and $\lambda_n := \lambda_{\min}(Q_n)$ denote its smallest eigenvalue. For example, when $K = 1$ and equal weights are placed on all observations, W_i is a scalar and $\lambda_n = Q_n = Var(\sum_i \omega_i W_i)$. K_0 is used throughout the paper to denote an arbitrary constant.

Assumption 2. *For $C \in \{G, H\}$, and $k \in \{1, 2, \cdots, K\}$, there exists $K_0 < \infty$ such that:*

1. $E[W_{ik}^4] \leq K_0$ *for all* i.
2. $\frac{1}{\lambda_n} \max_c \left(\sum_{i \in \mathcal{N}_c^C} |\omega_i|\right)^2 \to 0.$
3. $\frac{1}{\lambda_n} \sum_c \sum_{i,j \in \mathcal{N}_c^C} A_{ij} |\omega_i \omega_j| \leq K_0.$

Assumption 2.1 requires the fourth moment to be bounded, which is stronger than the moment condition in one-way clustering.[8] The proof in one-way clustering usually verifies a Lindeberg

[6] To illustrate this, let $N_{gh} = 1$ and W_{gh} denote the observation in cluster g and h on the respective dimensions. Due to Kallenberg (2005), $\{W_{gh}\}_{g \geq 1, h \geq 1}$ is separately exchangeable if and only if there exists a representation $W_{gh} = f(\alpha_g, \gamma_h, \varepsilon_{gh})$, where $(\alpha_g, \gamma_h, \varepsilon_{gh}) \overset{iid}{\sim} U[0,1]$. Then, it is obvious that $W_{gh} \perp\!\!\!\perp W_{g'h'}$ for $g \neq g'$, $h \neq h'$. A counterexample for the converse is some $W_{gh} = -W_{gh'}$. These random variables are allowed to be perfectly correlated since they share a cluster under Assumption 1. However, we cannot find a representation $f(.)$, because that representation implies $E[W_{gh}|\alpha_g] \perp\!\!\!\perp E[W_{gh'}|\alpha_g]$.

[7] It is insufficient to define the indicator as $A_{ij} := 1[Cov(W_i, W_j) \neq 0]$, since the proof contains third and fourth moments. For $K = 1$, zero covariance between a pair of observations is insufficient to ensure objects such as $E[W_i W_j W_k]$ and $E[W_i W_j W_k W_l] - E[W_i W_k]E[W_j W_l]$ are zero.

[8] See equation (7) of Hansen and Lee (2019) for the condition in one-way clustering.

condition because blocks of observations are independent of each other. With multi-way dependence, we no longer have independent blocks because each cluster can have observations that are dependent with observations from a different cluster when these observations share a cluster on a different dimension. Hence, a different proof strategy is required. The proof in this paper uses Stein's method, which requires stronger moment restrictions, but provides a non-asymptotic bound on the approximation error — details are in Subsection 2.1.

Assumption 2.2 requires the contribution of the cluster with the largest weight to be small relative to the total variance. In the special case where everyone is equally weighted with $\omega_i = 1$, the condition is simply $(1/\lambda_n)\max_c(N_c^C)^2 \to 0$. Intuitively, this condition is required so that the removal of a cluster does not change the variance substantively. This assumption allows the ratio of any two cluster sizes to diverge to infinity. It is identical to equation (12) of Hansen and Lee (2019) when $C = G = H$. Assumption 2.2 also rules out having components that are perfectly negatively correlated: if the components of the vector were perfectly negatively correlated, $\lambda_n = 0$.

Assumption 2.3 is fairly unrestrictive about the convergence rate. To aid exposition, suppose $\omega_i = 1\ \forall i$, $K = 1$, and C is taken to be the clustering dimension that $\lambda_n \asymp \sum_c \sum_{i,j\in\mathcal{N}_c^C} A_{ij}$.[9] With strong dependence, $A_{ij} = 1$ for all $i, j \in \mathcal{N}_c^C$, so $\lambda_n \asymp \sum_c (N_c^C)^2$. However, if the researcher were conservative and clustered on C when the data is indeed iid, then $A_{ij} = 1$ if and only if $i = j$, so $\lambda_n \asymp n$. Assumption 2.3 has implications on λ_n, which then determines how strong Assumption 2.2 is. Namely, when $\lambda_n \asymp n$, Assumption 2.2 requires $\max_c(N_c^C)^2/n \to 0$. When $\lambda_n \asymp \sum_c (N_c^C)^2$, then Assumption 2.2 only requires $\max_c(N_c^C)^2/(\sum_{c'}(N_{c'}^C)^2) \to 0$. The weaker version of Assumption 2.2 permits balanced panels where the unit and time dimensions increase at the same rate, while the stronger version does not.[10] The assumption that $(1/\lambda_n)\sum_c (N_c^C)^2 \le K_0$ matches equation (11) of Hansen and Lee (2019).

Remark 1. *Assumption 2.3 rules out the following purely interactive model. As pointed out by Menzel (2021), this model has an asymptotic distribution that is non-normal, and there is no analog in one-way clustering. For $g \in \{1, \cdots, M\}$, $h \in \{1, \cdots, M\}$ and $N_{gh} = 1$, we observe $W_{gh} = \alpha_g \gamma_h$, where α_g, γ_h are iid with mean zero and variances σ_α^2 and σ_γ^2 respectively, so there are M^2 observations. Then, $\sum_{g,h} W_{gh}/M = \left(\sum_g \alpha_g/\sqrt{M}\right)\left(\sum_h \gamma_h/\sqrt{M}\right) \xrightarrow{d} Z_1 Z_2$, where Z_1 and Z_2*

[9] To be clear about the notation, $a \asymp b$ if and only if there exists $K_0 < \infty$ such that $a/b, b/a \in [-K_0, K_0]$. Since $E[W_i^2]$ is bounded, $\lambda_n \asymp \max_{C\in\{G,H\}} \sum_c \sum_{i,j\in\mathcal{N}_c^C} A_{ij}$.

[10] To see this, let M denote the number of units and time periods, so there are M^2 observations. $\max_c(N_c^C)^2/(\sum_{c'}(N_{c'}^C)^2) = M^2/M^3 = 1/M \to 0$, but $\max_c(N_c^C)^2/n = M^2/M^2 = 1 \neq o(1)$.

are independent standard normal distributions. This limiting distribution is also known as Gaussian chaos. $\sum_g (N_g^G)^2/\lambda_n = M^3/(M^2\sigma_\alpha^2\sigma_\gamma^2) = M/\sigma_\alpha^2\sigma_\gamma^2 \to \infty$ *violates Assumption 2.3.*

Theorem 1. *Under Assumption 1 and 2,* $Q_n^{-1/2}\sum_{i=1}^n \omega_i(W_i - E[W_i]) \xrightarrow{d} N(0, I_K)$. *Further,*

1. *If* $E[W_i] = 0\ \forall i$, *then* $Q_n^{-1}\hat{Q}_n \xrightarrow{p} I_K$, *where* $\hat{Q}_n := \sum_i \sum_{j\in\mathcal{N}_i} \omega_i\omega_j W_i W_j'$.
2. *If* $E[W_i] = \mu\ \forall i$ *and* $\frac{1}{\lambda_n}\sum_c \sum_{i,j\in\mathcal{N}_c^C} |\omega_i\omega_j| \leq K_0$ *for some* $K_0 < \infty$, *then, for* $\bar{W} = (\sum_i \omega_i W_i)/(\sum_j \omega_j)$ *and* $\hat{Q}_n := \sum_i \sum_{j\in\mathcal{N}_i} \omega_i\omega_j(W_i - \bar{W})(W_j - \bar{W})'$, $\bar{W} \xrightarrow{p} \mu$ *and* $Q_n^{-1}\hat{Q}_n \xrightarrow{p} I_K$.

The theorem tells us that, under the aforementioned conditions, $Q_n^{-1/2}\sum_{i=1}^n \omega_i(W_i - E[W_i])$ is asymptotically standard normal and the plug-in variance estimator proposed by CGM is consistent for multi-way clustering. One-way clustering is a special case of this theorem when one dimension is weakly nested within the other: examples include $G = H$ so both dimensions are identical, or if we cluster by county and state (as counties are nested in states), or if everyone is in their own H cluster. A sufficient condition for consistent variance estimation is $E[W_i] = 0$, similiar to theorem 3 of Hansen and Lee (2019). This assumption is sufficient in many applications: for example, linear regressions considered in Section 3 are identified by requiring the expectation of the residual term to be zero. Additionally, the condition $E[W_i] = \mu$ matches theorem 4 of Hansen and Lee (2019) for consistent variance estimation. Theorem 1.2 uses a stronger form of Assumption 2.3 where $A_{ij} = 1$ for all $i, j \in \mathcal{N}_c^C$.

Remark 2. *If* $E[W_i] \neq 0$, *then the variance estimator need not be consistent. Unlike one-way clustering, it may not even be conservative. Suppose* $E[W_i] \neq 0$ *for some* i, *and define* $\tilde{W}_i := W_i - E[W_i]$. *Then,* $Q_n^{-1}\sum_i\sum_{j\in\mathcal{N}_i} W_i W_j' = Q_n^{-1}\left(\sum_i\sum_{j\in\mathcal{N}_i}\tilde{W}_i\tilde{W}_j'\right) + Q_n^{-1}\left(\sum_i\sum_{j\in\mathcal{N}_i} E[W_i]E[W_j]'\right)$. *Since* $Q_n^{-1}\left(\sum_i\sum_{j\in\mathcal{N}_i}\tilde{W}_i\tilde{W}_j'\right) = o_P(1)$ *by Theorem 1.1, and* Q_n *is positive semidefinite, whether the asymptotic variance is over or under estimated depends on whether* $\sum_i\sum_{j\in\mathcal{N}_i} E[W_i]E[W_j]'$ *is positive semidefinite. Let* $K = 1$ *for exposition. In one-way clustering, the variance is weakly over-estimated, so inference is conservative. To see this, let* W_g^G *denote the vector of* W_i *such that* $g(i) = g$. $\sum_i\sum_{j\in\mathcal{N}_i} E[W_i]E[W_j] = \sum_g\sum_{i,j\in\mathcal{N}_g^G} E[W_i]E[W_j] = \sum_g 1'E[W_g^G]E[W_g^G]'1 \geq 0$. *In two-way clustering,* $\sum_i\sum_{j\in\mathcal{N}_i} E[W_i]E[W_j]$ *can be negative. An example is where* $n = 3$: $cov(W_1, W_3) = 0$ *but* $cov(W_1, W_2) \neq 0$ *and* $cov(W_2, W_3) \neq 0$, *so* W_1 *and* W_2 *share a cluster in one dimension and* W_2 *and* W_3 *share a cluster on a different dimension. Further,* $E[W_2] = -1$ *and* $E[W_1] = E[W_3] = 1$. *Then,* $\sum_i\sum_{j\in\mathcal{N}_i} E[W_i]E[W_j] = -1$.

2.1 Proof Sketch

The proof of Theorem 1 proceeds by first proving a CLT for a scalar random variable, then applying the Cramer-Wold device to obtain the multivariate CLT. The scalar CLT is proven using Stein's method. I adapt the proof strategy from Ross (2011) to obtain an upper bound on the Wasserstein distance between a pivotal statistic and the standard normal random variable. By exploiting the multi-way clustering structure, the upper bound on the distance can be shown to converge to zero. All details are in Appendix A.

For ease of exposition, consider a simpler environment where $K = 1$, $\omega_i = 1$ for all i, and $A_{ij} = 1$ whenever $c(i) = c(j)$ for some c, and $E[W_i] = 0$. Lemma 4 in Appendix A provides an explicit bound on the Wasserstein distance. With $d_W(.)$ denoting the Wasserstein distance, $\sigma_n^2 := Q_n$ and $R = \sum_i X_i/\sigma_n$,

$$d_W(R, Z) \leq \frac{1}{\sigma_n^3}\sum_{i=1}^{n}\left|\sum_{j,k\in\mathcal{N}_i} E[W_iW_jW_k]\right| + \frac{\sqrt{2}}{\sqrt{\pi}\sigma_n^2}\sqrt{Var\left(\sum_{i=1}^{n}\sum_{j\in\mathcal{N}_i} W_iW_j\right)}$$

At this point, my proof departs from the proofs in existing statistical literature that employ Stein's method (e.g., Chen and Shao (2004)). Let $N_i := |\mathcal{N}_i|$. Holder's inequality is employed on objects such as $\sum_i |\sum_{j,k\in\mathcal{N}_i} E[W_iW_jW_k]|$. Existing literature uses the L^1 norm of moments $E[W_i^3]$ and L^∞ norm of N_i, resulting in $(\max_m N_m)^2 \sum_i E[W_i^3]$. In contrast, my proof uses the L^∞ norm of $E[W_i^3]$ and L^1 norm of N_i, resulting in $\max_m E[W_m^3]\sum_i N_i^2$. Hence,

$$\frac{1}{\sigma_n^3}\sum_{i=1}^{n}\left|\sum_{j,k\in\mathcal{N}_i} E[W_iW_jW_k]\right| \leq \frac{1}{\sigma_n^3}\max_m E[W_m^3]\sum_i N_i^2$$

Since $\max_m E[W_m^3]$ is bounded by Assumption 2.1, it suffices to show $\sum_i N_i^2/\sigma_n^3 \to 0$. Due to Assumption 1, $N_i \leq N_{g(i)}^G + N_{h(i)}^H$, so

$$\begin{aligned}\frac{1}{\sigma_n^3}\sum_i N_i^2 &\leq \frac{1}{\sigma_n^3}\sum_i (N_{g(i)}^G + N_{h(i)}^H)^2 \leq \frac{1}{\sigma_n^3}\max_{g,h}(N_g^G + N_h^H)\sum_i (N_{g(i)} + N_{h(i)}) \\ &\leq \left[\frac{1}{\sigma_n}\max_{g,h}(N_g^G + N_h^H)\right]\frac{1}{\sigma_n^2}\left(\sum_g (N_g^G)^2 + \sum_h (N_h^H)^2\right)\end{aligned}$$

Since $\lambda_n = \sigma_n$ when $K = 1$, $\max_{g,h}(N_g^G + N_h^H)/\sigma_n \to 0$ by Assumption 2.2 and $\left(\sum_g (N_g^G)^2 + \sum_h (N_h^H)^2\right)/\sigma_n^2$ is bounded by Assumption 2.3. Hence, the term is $o(1)$.

A similar argument is made for the fourth moment that features in $Var\left(\sum_{i=1}^n \sum_{j\in\mathcal{N}_i} W_i W_j\right)$. To complete the proof for variance estimation, observe that since the fourth moments exist, the consistency of the plug-in variance estimator can be proven by using Chebyshev's inequality and existing intermediate results.

Remark 3. *Due to the proof strategy, the intermediate results are informative about the quality of the normal approximation. With $d_K(.)$ denoting the Kolmogorov distance, proposition 1.2 from Ross (2011) implies that $d_K(R, Z) \leq (2/\pi)^{1/4}\sqrt{d_W(R, Z)}$. Since Z is standard normal in the proof of CLT, the bound on $d_W(.)$ also places a bound on the Kolmogorov distance $d_K(.)$. This is then informative of the maximum distance between the pivotal statistic and the standard normal.*

3 Application

This section applies Theorem 1 to linear regressions, showing that using the normal approximation with the CGM estimator is valid. Consider a linear model where scalar outcome Y_i is generated by

$$Y_i = D_i\theta + W_i'\gamma + u_i =: X_i'\beta + u_i$$

$D_i \in \mathbb{R}$ is the regressor of interest, $W_i \in \mathbb{R}^{K-1}$ is a vector of controls that may include the intercept, and let $X_i = (X_{i1}, X_{i2}, \cdots, X_{iK})' := (D_i, W_i')' \in \mathbb{R}^K$. We are interested in estimating θ. The coefficient vector $\beta := (\theta, \gamma')' \in \mathbb{R}^K$ is the same for all individuals. The stochastic residual term u_i satisfies $E[u_i|X_i] = 0$ for all i, and is allowed to be multi-way clustered. The standard OLS estimator is

$$\hat{\beta} = \left(\sum_{i=1}^n X_i X_i'\right)^{-1}\left(\sum_{i=1}^n X_i Y_i\right) = \beta + \left(\sum_{i=1}^n X_i X_i'\right)^{-1}\left(\sum_{i=1}^n X_i u_i\right)$$

This object is assumed to be well-defined in that $\sum_{i=1}^n X_i X_i'$ is invertible. Using an equivalent representation with data matrices, the model is $Y = D\theta + W\gamma + u = X\beta + u$. Let $M_W = I - W(W'W)^{-1}W'$ denote the annihilator matrix. Let $\tilde{D} := M_W D$ be the D with W's partialled out, and define $\tilde{Y}, \tilde{u}$ in a similar manner. By the Frisch-Waugh-Lovell theorem (FWL),

$$\hat{\theta} = (\tilde{D}'\tilde{D})^{-1}\tilde{D}'\tilde{Y} = \theta + (\tilde{D}'\tilde{D})^{-1}\tilde{D}'\tilde{u} = \theta + \left(\sum_i \tilde{D}_i^2\right)^{-1} \left(\sum_i \tilde{D}_i \tilde{u}_i\right) = \hat{\beta}_1$$

where $\tilde{D}_i$ is the ith component of $\tilde{D}$, so $\sum_i \tilde{D}_i \tilde{u}_i = \tilde{D}'\tilde{u} = D' M_W u = \sum_i \tilde{D}_i u_i$. Let $\sigma_n^2 := Var(\hat{\theta}) = Var\left(\sum_i \tilde{D}_i u_i / (\sum_{i'} \tilde{D}_{i'}^2)\right)$ and $\hat{\sigma}_n^2 := \left(\sum_i \sum_{j \in \mathcal{N}_i} \hat{u}_i \hat{u}_j \tilde{D}_i \tilde{D}_j\right) / \left(\sum_i \tilde{D}_i^2\right)^2$. Estimated residuals are $\hat{u}_i := Y_i - X_i\hat{\beta} = u_i - X_i(\hat{\beta} - \beta)$. Due to FWL, $\hat{u}_i = \tilde{Y}_i - \tilde{D}_i\hat{\theta} = u_i - \tilde{D}_i(\hat{\theta} - \theta)$.

Inference for $\hat{\theta}$, depends on whether we are conditioning on X: the conditions for asymptotic normality differ slightly between random and nonrandom X. I consider each of them in turn.

3.1 Fixed Regressors

First, consider regressions where the X's are nonrandom. An example might be when the object of interest is the difference between male and female wages. Their unobserved error may be correlated by state and industry conditional on X, but the gender status D_i is fixed. This can be viewed as inference on a descriptive object.

With u_i's having a multi-way clustered structure, we can apply Theorem 1 on $\left(\sum_i \tilde{D}_i^2\right)^{-1} \sum_i \tilde{D}_i u_i$, where scalar weights are given by $\omega_i = \tilde{D}_i / (\sum_{i'} \tilde{D}_{i'}^2)$.

Assumption 3. *For $C \in \{G, H\}$ and nonstochastic $\tilde{D}_i$, there exists $K_0 < \infty$ such that:*

1. $E[u_i^4] \leq K_0$, $E[u_i] = 0$.
2. $\dfrac{\max_c \left(\sum_{i \in \mathcal{N}_c^C} |\tilde{D}_i|\right)^2}{\sum_{c'} \left(\sum_{j \in \mathcal{N}_{c'}^C} |\tilde{D}_j|\right)^2} \to 0$.
3. $\dfrac{\sum_{c'} \sum_{i,j \in \mathcal{N}_c^C} |\tilde{D}_i \tilde{D}_j|}{Var\left(\sum_i \tilde{D}_i u_i\right)} \leq K_0$.
4. $u_i \perp\!\!\!\perp u_j$ *if* $g(i) \neq g(j)$ *and* $h(i) \neq h(j)$.

Proposition 1. *Under Assumption 3, $(\hat{\theta} - \theta)/\sigma_n \xrightarrow{d} N(0, 1)$, and $\hat{\sigma}_n^2/\sigma_n^2 \xrightarrow{p} 1$.*

Assumption 3 works in the environment where there is no misspecification, so $A_{ij} = 1$ whenever i, j share at least one cluster. Hence, $\sigma_n^2 \asymp \max_{C \in \{G,H\}} \sum_c \sum_{i,j \in \mathcal{N}_c^C} |\omega_i \omega_j|$, satisfying the conditions of Theorem 1. Consequently, instead of making an assumption on the contribution of the cluster with the largest weight on the total variance, a leverage condition in the form of Assumption 3.2 can be obtained. This condition is also empirically verifiable: the researcher can calculate $L_C := \max_c \left(\sum_{i \in \mathcal{N}_c^C} |\tilde{D}_i|\right)^2 / \left(\sum_{c'} \left(\sum_{j \in \mathcal{N}_{c'}^C} |\tilde{D}_j|\right)^2\right)$, and check if it is small. As a benchmark, when observations are not clustered and all weights $\tilde{D}_i$ are the same, $L_C = 1/n$. Hence, if we believe that $n = 30$ is sufficiently large for asymptotics in the iid case, then $L_C < 1/30$ may be acceptable.

Proposition 1 implies that the usual inference procedure is still valid even when the unobserved component is arbitrarily heterogeneous across different clusters. In contrast, the separate exchangeability of u_i requires u_i to be identically distributed across different clusters (e.g., the unobserved component of wages for women is identically distributed across states) — it is a strong assumption that is no longer required here. If there are fixed effects in the model, the vector of indicators can be collected in W and the argument proceeds as usual.[11]

3.2 Stochastic Regressors

Next, consider stochastic X. This is the relevant case when considering causal regressions. For example, we may be interested in the effect of a randomly assigned opportunity to participate in a job training program D_i on wages Y_i. Both X_i and u_i are plausibly correlated within state and within industry. Although $\hat{\theta} = \hat{\beta}_1$, we can no longer apply Theorem 1 to $\sum_i \tilde{D}_i u_i$ because the multi-way dependence structure breaks once X_i's are random.

Define $S_n := \sum_{i=1}^n E[X_i X_i']$ and $Q_n := Var\left(\sum_{i=1}^n X_i u_i\right)$, and denote their sample analogs as $\hat{S}_n = \sum_i X_i X_i'$ and $\hat{Q}_n := \sum_i \sum_{j \in \mathcal{N}_i} \hat{u}_i \hat{u}_j X_i X_j'$. Let the smallest eigenvalue of Q_n be $\lambda_n := \lambda_{\min}(Q_n)$. The asymptotic variance of $\hat{\beta}$ and its sample analog are $V(\hat{\beta}) := S_n^{-1} Q_n S_n^{-1}$ and $\hat{V}(\hat{\beta}) := \hat{S}_n^{-1} \hat{Q}_n \hat{S}_n^{-1}$ respectively.

Assumption 4 provides sufficient conditions for asymptotic normality of the estimator $\hat{\beta}$ and consistency of the CGM variance estimator. The conditions mimic Assumption 2 so that Theorem 1 is applicable to the random vector $X_i u_i$. The new condition is a weak regularity condition that $\lambda_{\min}(S_n/n) \geq K_1 > 0$, mimicking to the rank condition in OLS.

[11] Fixed effects account for a shift in the unobserved component, so separate exchangeability still makes a restriction on the distribution of the remaining unobserved component.

Assumption 4. *For $C \in \{G, H\}$, and $k \in \{1, 2, \cdots, K\}$, there exists $K_0 < \infty$ and $K_1 > 0$:*

1. $E[u_i^4|X_i] \leq K_0$, $E[X_{ik}^4] \leq K_0$, $E[u_i|X_i] = 0$ *for all* i.
2. $\frac{1}{\lambda_n} \max_c (N_c^C)^2 \to 0$.
3. $\frac{1}{\lambda_n} \sum_c (N_c^C)^2 \leq K_0$.
4. $(X_i', u_i)' \perp\!\!\!\perp (X_j', u_j)'$ *if* $g(i) \neq g(j)$ *and* $h(i) \neq h(j)$.
5. $\lambda_{\min}\left(\frac{1}{n} S_n\right) \geq K_1$.

Proposition 2. *Under Assumption 4,* $Q_n^{-1/2} S_n(\hat{\beta} - \beta) \xrightarrow{d} N(0, I_K)$, *and* $[S_n^{-1} Q_n S_n^{-1}]^{-1} [\hat{S}_n^{-1} \hat{Q}_n \hat{S}_n^{-1}] \xrightarrow{p} I_K$.

Proposition 2 is useful for doing F tests on a subvector of β. The proof of Proposition 2 proceeds by applying Theorem 1 to $\sum_i X_i u_i$, and showing that $S_n^{-1} \hat{S}_n \xrightarrow{p} I_K$. The latter requires the rank condition of Assumption 4.5. It then remains to show that the remainder terms are asymptotically negligible. Nonetheless, if we are only interested in θ, using the residualized objects $\hat{\theta}$ and variance estimator for the residualized object $\hat{\sigma}_n^2$ is still valid. This follows from FWL, and the refinement of FWL for variance estimators in Ding (2021).

Corollary 1. *Under Assumption 4,* $(\hat{\theta} - \theta)/\sigma_n \xrightarrow{d} N(0, 1)$, *and* $\hat{\sigma}_n/\sigma_n \xrightarrow{p} 1$.

The practitioner's takeaway from Proposition 2 is that the existing CGM variance estimator can be used for valid inference with multi-way clustering. With Corollary 1, $\hat{\theta}$ and $\hat{\sigma}_n^2$ can be used as the mean and variance estimators respectively. These results provide the formal theoretical guarantee for using the estimator, under weaker conditions that permits heterogeneity across clusters.

Besides the application mentioned, Theorem 1 also has implications on the conditions required for valid inference when the random variable is multi-way clustered in many other econometric models, including design-based settings and instrument variables models. Inference for estimators based on moment conditions can be done by straightforward application of Theorem 1 as in the linear regression case.

A Proof of Theorem 1

The proof strategy is as follows. I first prove Lemma 1, which is a central limit theorem (CLT) for scalars that permits weights on the random variable. The proof of Lemma 1 relies on Lemmas 2 to 7. Lemmas 2 to 4 derive an upper bound on the Wasserstein distance between a pivotal statistic and standard normal Z. Lemmas 5 to 7 then show that the derived upper bound is $o(1)$. With Lemma 1, the multivariate CLT of Theorem 1 is obtained by using the Cramer-Wold device. The remainder of the proof proceeds in the following order: (i) introduce definitions and notation, (ii) state Lemma 1, (iii) state and prove Lemmas 2 to 7, (iv) prove Lemma 1, (v) state and prove Lemma 8 that is required for consistent variance estimation, then (vi) complete the proof of Theorem 1.

The following definitions and notations are used throughout the proof. Let $d_W(X,Y)$ denote the Wasserstein distance between random variables X and Y, so $d_W(X,Y) = 0$ if and only if the distributions of X and Y are identical. The norms of functions are defined as the sup norm i.e., $||f|| = \sup_{x\in D}|f(x)|$. For vector a, $||a|| = (a'a)^{1/2}$ is the Euclidean norm, and for positive semi-definite matrix A and $\lambda_{\max}(A)$ denoting the largest eigenvalue, $||A|| = \sqrt{\lambda_{\max}(A'A)}$ denotes the spectral norm, and $A^{1/2}$ denotes the symmetric matrix such that $A^{1/2}A^{1/2} = A$. $\sum_{i\in\mathcal{N}_g^G}\sum_{j\in\mathcal{N}_g^G}$ is abbreviated as $\sum_{i,j\in\mathcal{N}_g^G}$. The dependency neighborhood of i, $\mathcal{N}_i \subseteq \{1,\cdots,n\}$, is defined as the set of observations where $i \in \mathcal{N}_i$ and X_i is independent of $\{X_j\}_{j\neq\mathcal{N}_i}$, and $N_i := |\mathcal{N}_i|$ is the number of observations in i's dependency neighborhood. In the rest of this proof, X_i denotes a scalar random variable while $W_i \in \mathbb{R}^K$ as stated in the main text is a random vector.

Every scalar random variable X_i is weighted by nonstochastic ω_i. Denote the variance of the sum as $\sigma_n^2 := Var\left(\sum_{i=1}^n \omega_i X_i\right)$. We are interested in the asymptotic distribution of $(1/\sigma_n)\sum_{i=1}^n \omega_i X_i$. If all observations are equally weighted, $\omega_i = 1\ \forall i$.

Assumption 5. *For $C \in \{G,H\}$, there exists $K_0 < \infty$ such that:*

1. $E[X_i] = 0$ *and* $E[X_i^4] \leq K_0 < \infty$ *for all* i.
2. $\frac{1}{\sigma_n^2}\max_c\left(\sum_{i\in\mathcal{N}_c^C}|\omega_i|\right)^2 \to 0$
3. $\frac{1}{\sigma_n^2}\sum_c\sum_{i,j\in\mathcal{N}_c^C} A_{ij}|\omega_i\omega_j| \leq K_0 < \infty$
4. $X_i \perp\!\!\!\perp X_j$ *if* $g(i)\neq g(j)$ *and* $h(i)\neq h(j)$.

Lemma 1. *Under Assumption 5,* $(1/\sigma_n)\sum_{i=1}^n \omega_i X_i \xrightarrow{d} N(0,1)$*, where* $\sigma_n^2 := Var\left(\sum_{i=1}^n \omega_i X_i\right)$*. Further, using feasible estimator* $\hat{\sigma}_n^2 := \sum_i \sum_{j\in\mathcal{N}_i} \omega_i\omega_j X_i X_j$*,* $\hat{\sigma}_n^2/\sigma_n^2 \xrightarrow{p} 1$*.*

Lemma 2. *If* R *is a random variable and* Z *has a standard normal distribution, and we define the family of functions* $\mathcal{F} = \{f : ||f||, ||f''|| \leq 2, ||f'|| \leq \sqrt{2\pi}\}$*, then* $d_W(R,Z) \leq \sup_{f\in\mathcal{F}} |E[f'(R) - Rf(R)]|$*.*

Proof. See Ross (2011) theorem 3.1. $\square$

Lemma 3. *Let* $X_1, \cdots, X_n$ *be random variables such that* $E[X_i] = 0, \sigma_n^2 = Var(\sum_i X_i)$*, and define* $R = \sum_i X_i/\sigma_n$*. If* $R_i := \sum_{j\neq\mathcal{N}_i} X_j/\sigma_n$*, then*

$$E[Rf(R)] = E\left[\frac{1}{\sigma_n}\sum_{i=1}^n X_i(f(R) - f(R_i) - (R-R_i)f'(R))\right] + E\left[\frac{1}{\sigma_n}\sum_{i=1}^n X_i(R-R_i)f'(R)\right]$$

Proof. Start from right hand side:

$$\begin{aligned}
&E\left[\frac{1}{\sigma_n}\sum_{i=1}^n X_i(f(R) - f(R_i) - (R-R_i)f'(R))\right] + E\left[\frac{1}{\sigma_n}\sum_{i=1}^n X_i(R-R_i)f'(R)\right]\\
&= E\left[\frac{1}{\sigma_n}\sum_{i=1}^n X_i(f(R) - f(R_i))\right] = E\left[\frac{1}{\sigma_n}\sum_{i=1}^n X_i f(R)\right] + E\left[\frac{1}{\sigma_n}\sum_{i=1}^n X_i f(R_i)\right]\\
&= E\left[\frac{1}{\sigma_n}\sum_{i=1}^n X_i f(R)\right] = E[Rf(R)]
\end{aligned}$$

The first equality in the final line comes from the fact that R_i is independent of X_i based on how dependency neighborhoods are defined. Hence, $E[X_i f(R_i)] = 0$. $\square$

Lemma 4. *Let* $X_1, \cdots, X_n$ *be random variables such that,* $E[X_i] = 0, \sigma_n^2 = Var(\sum_i X_i)$*, and define* $R = \sum_i X_i/\sigma_n$*. Let the collection* $(X_1, \cdots, X_n)$ *have dependency neighborhoods* $\mathcal{N}_i$*,* $i = 1, \cdots, n$*. Then for* Z *a standard normal random variable,*

$$d_W(R,Z) \leq \frac{1}{\sigma_n^3}\sum_{i=1}^n \left|\sum_{j,k\in\mathcal{N}_i} E[X_iX_jX_k]\right| + \frac{\sqrt{2}}{\sqrt{\pi}\sigma_n^2}\sqrt{Var\left(\sum_{i=1}^n\sum_{j\in\mathcal{N}_i} X_iX_j\right)} \tag{1}$$

Proof. Due to Lemma 2, to bound $d_W(R,Z)$ from above, it is sufficient to bound $|E[f'(R) - Rf(R)]|$, where $||f||, ||f''|| \leq 2, ||f'|| \leq \sqrt{2/\pi}$. Define $R_i := \sum_{j \neq N_i} X_j/\sigma_n$, so X_i is independent of R_i.

$$
\begin{aligned}
&|E[f'(R) - Rf(R)]| = |E[f'(R)] - E[Rf(R)]| \\
&\leq \left| E[f'(R)] - E\left[\frac{1}{\sigma_n}\sum_{i=1}^{n} X_i(f(R) - f(R_i) - (R - R_i)f'(R))\right] - E\left[\frac{1}{\sigma_n}\sum_{i=1}^{n} X_i(R - R_i)f'(R)\right]\right| \\
&\leq \left| E\left[\frac{1}{\sigma_n}\sum_{i=1}^{n} X_i(f(R) - f(R_i) - (R - R_i)f'(R))\right]\right| + \left| E\left[f'(R)\left(1 - \frac{1}{\sigma_n}\sum_{i=1}^{n} X_i(R - R_i)\right)\right]\right|
\end{aligned}
$$

The first inequality applies Lemma 3, and the second inequality applies the triangle inequality. Consequently, it is sufficient to show that the first term is bounded by the corresponding first term of Equation (1), and the second term is bounded by the corresponding second term.

Consider the first term. By Taylor expansion of $f(R_i)$ around $f(R)$, and the triangle inequality, the term that generates the third moment is:

$$
\begin{aligned}
&|E\left[\frac{1}{\sigma_n}\sum_{i=1}^{n} X_i(f(R) - f(R_i) - (R - R_i)f'(R))\right]| \leq \frac{||f''||}{2\sigma_n}\left|\sum_{i=1}^{n} E[X_i(R - R_i)^2]\right| \\
&= \frac{1}{\sigma_n^3}\left|\sum_{i=1}^{n} E\left[X_i\left(\sum_{j \in \mathcal{N}_i} X_j\right)^2\right]\right| = \frac{1}{\sigma_n^3}\left|\sum_{i=1}^{n}\sum_{j,k \in \mathcal{N}_i} E[X_iX_jX_k]\right| \leq \frac{1}{\sigma_n^3}\sum_{i=1}^{n}\left|\sum_{j,k \in \mathcal{N}_i} E[X_iX_jX_k]\right|
\end{aligned}
$$

Turning now to the second term,

$$
\begin{aligned}
&\left| E\left[f'(R)\left(1 - \frac{1}{\sigma_n}\sum_{i=1}^{n} X_i(R - R_i)\right)\right]\right| \leq \frac{||f'||}{\sigma_n^2}\left| E\left[\sigma_n^2 - \sigma_n\sum_{i=1}^{n} X_i(R - R_i)\right]\right| \\
&\leq \frac{||f'||}{\sigma_n^2} E\left|\sigma_n^2 - \sum_{i=1}^{n} X_i\left(\sum_{j \in N_i} X_j\right)\right| \leq \frac{||f'||}{\sigma_n^2} E\left[\left(\sigma_n^2 - \sum_{i=1}^{n} X_i\left(\sum_{j \in N_i} X_j\right)\right)^2\right]^{1/2} 1^{1/2} \\
&\leq \frac{\sqrt{2}}{\sqrt{\pi}\sigma_n^2}\sqrt{Var\left(\sum_{i=1}^{n}\sum_{j \in N_i} X_iX_j\right)}
\end{aligned}
$$

□

Lemma 5. $E[|X_iX_jX_k|] \leq \max_m E[|X_m|^3]$, $E[|X_iX_jX_kX_l|] \leq \max_m E[|X_m|^4]$, *and* $|E[X_iX_k]E[X_jX_l]| \leq \max_m E[|X_m|^4]$.

Proof. By the arithmetic mean — geometric mean (AM-GM) inequality,

$$E|X_iX_jX_k| \leq \frac{1}{3}\left(E|X_i|^3 + E|X_j|^3 + E|X_k|^3\right) \leq \max_m E[|X_m|^3]$$

A similar argument yields $E[|X_iX_jX_kX_l|] \leq \max_m E[|X_m|^4]$. For the final result, first observe that $E[X_iX_k]^2 \pm 2E[X_iX_k]E[X_jX_l] + E[X_jX_l]^2 = (E[X_iX_k] \pm E[X_jX_l])^2 \geq 0$. Hence,

$$\begin{aligned} |E[X_iX_k]E[X_jX_l]| &\leq \frac{1}{2}(E[X_iX_k]^2 + E[X_jX_l]^2) \leq \frac{1}{2}(E[X_i^2X_k^2] + E[X_j^2X_l^2]) \\ &\leq \frac{1}{4}(E[X_i^4] + E[X_j^4] + E[X_k^4] + E[X_l^4]) \leq \max_m E[X_m^4] \end{aligned}$$

□

Lemma 6. *Under Assumption 5,* $\frac{1}{\sigma_n^3}\sum_{i=1}^n \left|\sum_{j,k\in\mathcal{N}_i} E[\omega_i\omega_j\omega_k X_iX_jX_k]\right| \to 0.$

Proof. Note that $E[X_iX_jX_k] = 0$ whenever one of $\{X_i, X_j, X_k\}$ is independent of the other two, so $E[\omega_i\omega_j\omega_kX_iX_jX_k]$ is nonzero only if A_{ij}, A_{ik}, or A_{jk} is nonzero. Apply the triangle inequality and push the absolute value into the expectation.

$$\begin{aligned} \frac{1}{\sigma_n^3}\sum_{i=1}^n \left|\sum_{j,k\in\mathcal{N}_i} E[\omega_i\omega_j\omega_kX_iX_jX_k]\right| &\leq \frac{1}{\sigma_n^3}\sum_{i=1}^n \left|\sum_{j,k\in\mathcal{N}_i}(A_{ij}+A_{jk}+A_{ik})E[\omega_i\omega_j\omega_kX_iX_jX_k]\right| \\ &\leq \frac{1}{\sigma_n^3}\sum_{i=1}^n \sum_{j,k\in\mathcal{N}_i}(A_{ij}+A_{jk}+A_{ik})|\omega_i\omega_j\omega_k|E[|X_iX_jX_k|] \\ &\leq \frac{\max_m E[|X_m|^3]}{\sigma_n^3}\sum_{i=1}^n \sum_{j,k\in\mathcal{N}_i}|\omega_i\omega_j\omega_k|(A_{ij}+A_{jk}+A_{ik}) \end{aligned}$$

The last inequality applies Lemma 5. Observe $\max_m E[|X_m|^3] \leq K_0$ since the 4th moment exists, so it remains to show that the remaining terms are $o(1)$.

$$\frac{1}{\sigma_n^3}\sum_{i=1}^n \sum_{j,k\in\mathcal{N}_i}(A_{ij}+A_{jk}+A_{ik})|\omega_i\omega_j\omega_k| \leq \frac{1}{\sigma_n^3}\sum_{i=1}^n \left(\sum_{j,k\in\mathcal{N}^G_{g(i)}} + \sum_{j,k\in\mathcal{N}^H_{h(i)}}\right)(A_{ij}+A_{jk}+A_{ik})|\omega_i\omega_j\omega_k|$$

It is sufficient to consider the G dimension as the H dimension is analogous.

$$\frac{1}{\sigma_n^3}\sum_{i=1}^{n}\sum_{j,k\in\mathcal{N}^G_{g(i)}}(A_{ij}+A_{jk}+A_{ik})|\omega_i\omega_j\omega_k| = \frac{3}{\sigma_n^3}\sum_g\sum_{i,j,k\in\mathcal{N}^G_g}A_{ij}|\omega_i\omega_j\omega_k|$$

$$\frac{1}{\sigma_n^3}\sum_g\sum_{i,j,k\in\mathcal{N}^G_g}A_{ij}|\omega_i\omega_j||\omega_k| \le \left(\frac{\max_g\sum_{k\in\mathcal{N}^G_g}|\omega_k|}{\sigma_n}\right)\frac{1}{\sigma_n^2}\sum_g\sum_{i,j\in\mathcal{N}^G_g}A_{ij}|\omega_i\omega_j| = o(1)$$

Convergence occurs because $(1/\sigma_n^2)\sum_g\sum_{i,j\in\mathcal{N}^G_g}A_{ij}|\omega_i\omega_j| < \infty$ by Assumption 5.3 and $\max_g\sum_{k\in\mathcal{N}^G_g}|\omega_k|/\sigma_n = \left(\max_g\left(\sum_{k\in\mathcal{N}^G_g}|\omega_k|\right)^2/\sigma_n^2\right)^{1/2} = o(1)$ by Assumption 5.2. □

Lemma 7. *Under Assumption 5,* $\frac{1}{\sigma_n^4}Var\left(\sum_{i=1}^n\sum_{j\in\mathcal{N}_i}\omega_i\omega_jX_iX_j\right) = o(1)$.

Proof.

$$\frac{1}{\sigma_n^4}Var\left(\sum_i\sum_{j\in\mathcal{N}_i}\omega_i\omega_jX_iX_j\right) = \frac{1}{\sigma_n^4}E\left[\left(\sum_i\sum_{j\in\mathcal{N}_i}\omega_i\omega_jX_iX_j\right)^2\right] - \frac{1}{\sigma_n^4}\left(\sum_i\sum_{j\in\mathcal{N}_i}E[\omega_i\omega_jX_iX_j]\right)^2$$
$$= \frac{1}{\sigma_n^4}\sum_i\sum_j\sum_{k\in\mathcal{N}_i}\sum_{l\in\mathcal{N}_j}(E[\omega_i\omega_j\omega_k\omega_lX_iX_jX_kX_l] - E[\omega_i\omega_kX_iX_k]E[\omega_j\omega_lX_jX_l])$$
$$= \frac{1}{\sigma_n^4}\sum_i\sum_j\sum_{k\in\mathcal{N}_i}\sum_{l\in\mathcal{N}_j}\omega_i\omega_j\omega_k\omega_l(E[X_iX_jX_kX_l] - E[X_iX_k]E[X_jX_l])$$

When $(X_i,X_k)\perp\!\!\!\perp(X_j,X_l)$, $E[X_iX_jX_kX_l] = E[X_iX_j]E[X_kX_l]$. Hence, we only have to consider where there is at least one pair that is correlated i.e., when A_{ij}, A_{il}, A_{kj}, or A_{kl} is not zero. As before, with finite 4th moment and Lemma 5, it is sufficient to show

$$\frac{1}{\sigma_n^4}\sum_i\sum_j\sum_{k\in\mathcal{N}_i}\sum_{l\in\mathcal{N}_j}|\omega_i\omega_j\omega_k\omega_l|(A_{ij}+A_{il}+A_{kj}+A_{kl}) = o(1)$$

It is sufficient to consider the A_{ij} term because everything else is analogous.

$$\sum_i\sum_j\sum_{k\in\mathcal{N}_i}\sum_{l\in\mathcal{N}_j}|\omega_i\omega_j\omega_k\omega_l|A_{ij} \le \sum_i\left(\sum_{j\in\mathcal{N}^G_{g(i)}}+\sum_{j\in\mathcal{N}^H_{h(i)}}\right)\left(\sum_{k\in\mathcal{N}^G_{g(i)}}+\sum_{k\in\mathcal{N}^H_{h(i)}}\right)\left(\sum_{l\in\mathcal{N}^G_{g(j)}}+\sum_{l\in\mathcal{N}^H_{h(j)}}\right)|\omega_i\omega_j\omega_k\omega_l|A_{ij}$$

The first and last terms of the summation take the form:

$$\sum_i \sum_{j\in\mathcal{N}^G_{g(i)}} \sum_{k\in\mathcal{N}^G_{g(i)}} \sum_{l\in\mathcal{N}^G_{g(j)}} |\omega_i\omega_j\omega_k\omega_l|A_{ij} = \sum_g \sum_{i,j,k,l\in\mathcal{N}^G_g} |\omega_i\omega_j\omega_k\omega_l|A_{ij} \leq \left(\max_g \sum_{k,l\in\mathcal{N}^G_g} |\omega_k||\omega_l|\right) \sum_g \sum_{i,j\in\mathcal{N}^G_g} |\omega_i\omega_j|A_{ij}$$

Since $\frac{1}{\sigma_n^2}\max_h \sum_{i,k\in\mathcal{N}^H_h} |\omega_i||\omega_k| = o(1)$ and $\frac{1}{\sigma_n^2}\sum_g\sum_{i,j\in\mathcal{N}^G_g} |\omega_i\omega_j|A_{ij} < \infty$ by Assumption 5, these terms are $o(1)$ when divided by σ_n^4.

The interactive terms have the form:

$$\begin{aligned}
&\sum_i \sum_{j\in\mathcal{N}^G_{g(i)}} \sum_{k\in\mathcal{N}^G_{g(i)}} \sum_{l\in\mathcal{N}^H_{h(j)}} |\omega_i\omega_j\omega_k\omega_l|A_{ij} \\
&\quad= \sum_{i,j,k}\sum_g 1[i\in\mathcal{N}^G_g]1[j\in\mathcal{N}^G_g]1[k\in\mathcal{N}^G_g]\sum_l\sum_h 1[j\in\mathcal{N}^H_h]1[l\in\mathcal{N}^H_h]|\omega_i\omega_j\omega_k\omega_l|A_{ij} \\
&\quad= \sum_j\sum_{i,k}\sum_g 1[i\in\mathcal{N}^G_g]1[j\in\mathcal{N}^G_g]1[k\in\mathcal{N}^G_g]A_{ij}|\omega_i\omega_j\omega_k|\sum_h\sum_l 1[j\in\mathcal{N}^H_h]1[l\in\mathcal{N}^H_h]|\omega_l| \\
&\quad\leq \left(\max_j \sum_h\sum_l 1[j\in\mathcal{N}^H_h]1[l\in\mathcal{N}^H_h]|\omega_l|\right)\left(\sum_g\sum_{i,j,k\in\mathcal{N}^G_g}|\omega_i\omega_j\omega_k|A_{ij}\right) \\
&\quad= \left(\max_h \sum_{l\in\mathcal{N}^H_h}|\omega_l|\right)\left(\sum_g\sum_{i,j,k\in\mathcal{N}^G_g}|\omega_i\omega_j\omega_k|A_{ij}\right) \\
&\quad= \left(\max_h \sum_{l\in\mathcal{N}^H_h}|\omega_l|\right)\left(\max_g \sum_{k\in\mathcal{N}^G_g}|\omega_k|\right)\left(\sum_g\sum_{i,j\in\mathcal{N}^G_g}|\omega_i\omega_j|A_{ij}\right)
\end{aligned}$$

Since $\sum_g\sum_{i,j\in\mathcal{N}^G_g}|\omega_i\omega_j|A_{ij}/\sigma_n^2 \leq K_0$ and $\max_g\sum_{k\in\mathcal{N}^G_g}|\omega_k|/\sigma_n = o(1)$,

$$\begin{aligned}
&\frac{1}{\sigma_n^4}\sum_i \sum_{j\in\mathcal{N}^G_{g(i)}} \sum_{k\in\mathcal{N}^G_{g(i)}} \sum_{l\in\mathcal{N}^H_{h(j)}} |\omega_i\omega_j\omega_k\omega_l|A_{ij} \\
&\quad\leq \left(\frac{1}{\sigma_n}\max_h\sum_{l\in\mathcal{N}^H_h}|\omega_l|\right)\left(\frac{1}{\sigma_n}\max_g\sum_{k\in\mathcal{N}^G_g}|\omega_k|\right)\left(\frac{1}{\sigma_n^2}\sum_g\sum_{i,j\in\mathcal{N}^G_g}|\omega_i\omega_j|A_{ij}\right) = o(1)
\end{aligned}$$

$\square$

Proof of Lemma 1. Apply Lemma 4 on random variable $\omega_i X_i$ to obtain:

$$d_W(R,Z) \leq \frac{1}{\sigma_n^3}\sum_{i=1}^{n}\left|\sum_{j,k\in\mathcal{N}_i} E[\omega_i\omega_j\omega_k X_i X_j X_k]\right| + \frac{\sqrt{2}}{\sqrt{\pi}\sigma_n^2}\sqrt{Var\left(\sum_{i=1}^{n}\sum_{j\in\mathcal{N}_i}\omega_i\omega_j X_i X_j\right)}$$

Applying Lemma 6 and 7 on each of the two terms, $d_W(R,Z) = o(1)$. Proof for consistency of the variance estimator is equivalent to proving that $(\hat{\sigma}_n^2 - \sigma_n^2)/\sigma_n^2 = o_P(1)$. By Chebyshev's inequality,

$$P\left(\frac{\hat{\sigma}_n^2 - \sigma_n^2}{\sigma_n^2} > \epsilon\right) \leq \frac{1}{\epsilon^2}\frac{1}{\sigma_n^4}E[(\hat{\sigma}_n^2 - \sigma_n^2)^2] = \frac{Var\left(\sum_i \sum_{j\in\mathcal{N}_i}\omega_i\omega_j X_i X_j\right)}{\epsilon^2\sigma_n^4} = o_P(1)$$

The convergence in the last step occurs by Lemma 7. □

Lemma 8. *Under Assumption 1, 2.1 and 2.2,* $\forall i$, $||(1/(\sum_i \omega_i))\sum_i \omega_i(W_i - E[W_i])|| \xrightarrow{p} 0$.

Proof. It suffices to show convergence elementwise. Let X_i denote a scalar components of W_i, i.e., $X_i = W_{im}$, where $m \in \{1, 2, \cdots, K\}$. By Chebyshev's inequality, and $\max_{m,k} E[W_{mk}^2] < K_0$,

$$\begin{aligned}
&P\left(\frac{1}{\sum_i \omega_i}\sum_i \omega_i(X_i - E[X_i]) > \epsilon\right) \leq \frac{1}{\epsilon^2}\frac{1}{\left(\sum_i \omega_i\right)^2}E\left(\sum_i\sum_{j\in\mathcal{N}_i}\omega_i\omega_j(X_i - E[X_i])(Y_i - E[Y_i])\right)\\
&\leq \frac{K_0}{\epsilon^2\left(\sum_j \omega_j\right)^2}\sum_i\sum_{j\in\mathcal{N}_i}\omega_i\omega_j
\end{aligned}$$

Hence, it suffices to show $(\sum_i\sum_{j\in\mathcal{N}_i}\omega_i\omega_j)/\left(\sum_j \omega_j\right)^2 = o(1)$. Observe

$$\frac{\sum_i\sum_{j\in\mathcal{N}_i}\omega_i\omega_j}{\left(\sum_j \omega_j\right)^2} \leq \frac{\max_i \sum_{j\in\mathcal{N}_i}|\omega_j|}{\left(\sum_j \omega_j\right)}\frac{\left(\sum_j \omega_j\right)}{\left(\sum_j \omega_j\right)}$$

so it suffices to show $\max_i \sum_{j\in\mathcal{N}_i}|\omega_j|/\left(\sum_j \omega_j\right) = o(1)$. Since $\lambda_n \leq \sum_i\sum_{j\in\mathcal{N}_i}|\omega_i\omega_j|\max_m E[W_{mk}^2] \leq \left(\sum_j |\omega_j|\right)^2 \max_m E[W_{mk}^2]$,

$$\frac{\left(\max_i \sum_{j\in\mathcal{N}_i}|\omega_j|\right)^2}{\left(\sum_j \omega_j\right)^2} = \frac{\left(\max_i \sum_{j\in\mathcal{N}_i}|\omega_j|\right)^2 \max_m E[W_{mk}^2]}{\left(\sum_j \omega_j\right)^2 \max_m E[W_{mk}^2]} \leq \max_m E[W_{mk}^2]\frac{\left(\max_i \sum_{j\in\mathcal{N}_i}|\omega_j|\right)^2}{\lambda_n} = o(1)$$

Convergence occurs due to Assumption 2.2 and $\max_m E[W_{mk}^2] < K_0$. $\square$

Proof of Theorem 1. To show that $Q_n^{-1/2} \sum_{i=1}^n \omega_i(W_i - E[W_i]) \xrightarrow{d} N(0, I_K)$, due to the Cramer-Wold device, it suffices to show that $\forall l \in \mathbb{R}^K$, $l'Q_n^{-1/2} \sum_{i=1}^n \omega_i(W_i - E[W_i]) \xrightarrow{d} l'N(0, I_K)$. If l is a vector of zeroes, then $l'Q_n^{-1/2} \sum_{i=1}^n \omega_i(W_i - E[W_i]) \xrightarrow{d} l'N(0, I_K)$ is immediate. For $||l|| > 0$, it suffices to show $(1/||l||)l'Q_n^{-1/2} \sum_{i=1}^n \omega_i(W_i - E[W_i]) \xrightarrow{d} (1/||l||)l'N(0, I_K) = N(0,1)$. For all nonstochastic $l \in \mathbb{R}^K \backslash \{0\}$, let $\sigma_n^2(l) := Var\left(\sum_i (l/||l||)' (Q_n/\lambda_n)^{-1/2} \omega_i(W_i - E[W_i])\right)$, so the following hold:

1. $E\left[\left(\left(\frac{l}{||l||}\right)'\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}(W_i - E[W_i])\right)\right] = 0$ and $E\left[\left(\left(\frac{l}{||l||}\right)'\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}(W_i - E[W_i])\right)^4\right] \leq K_0$ for all i.
2. $\frac{1}{\sigma_n^2(l)} \max_c \left(\sum_{i \in \mathcal{N}_c^C} |\omega_i|\right)^2 \to 0$.
3. $\frac{1}{\sigma_n^2(l)} \sum_c \sum_{i,j \in \mathcal{N}_c^C} A_{ij}|\omega_i \omega_j| \leq K_0$.
4. $\left(\left(\frac{l}{||l||}\right)'\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}(W_i - E[W_i])\right) \perp\!\!\!\perp \left(\left(\frac{l}{||l||}\right)'\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2} W_j\right)$ if $g(i) \neq g(j)$ and $h(i) \neq h(j)$.

For item 1, since $\lambda_n := \lambda_{\min}(Q_n)$, all eigenvalues of Q_n/λ_n must be at least 1. Hence, all eigenvalues of $(Q_n/\lambda_n)^{-1/2}$ are bounded above by 1. This implies $|(l/||l||)'(Q_n/\lambda_n)^{-1/2}| \leq K_1$ for some arbitrary constant $K_1 < \infty$. Item 1 then follows from Assumption 2.1. Observe that $\sigma_n^2(l) = (l/||l||)'(Q_n/\lambda_n)^{-1/2}Q_n(Q_n/\lambda_n)^{-1/2}(l/||l||) = 1/\lambda_n$. Then, Assumption 2.2 yields item 2 and Assumption 2.3 yields item 3. Item 4 is immediate from Assumption 1. By applying Lemma 1, $(1/\sigma_n(l))(l/||l||)'(Q_n/\lambda_n)^{-1/2} \sum_{i=1}^n \omega_i(W_i - E[W_i]) \xrightarrow{d} N(0,1)$. By using $\sigma_n^2(l) = 1/\lambda_n$, this is equivalent to $(l/||l||)'Q_n^{-1/2} \sum_{i=1}^n \omega_i(W_i - E[W_i]) \xrightarrow{d} N(0,1)$ as required.

Proof of Theorem 1.1

Turning to consistent variance estimation, it suffices to show that for all $l \in \mathbb{R}^K$ such that $||l|| = 1$, $P(l'Q_n^{-1}(\hat{Q}_n - Q_n)l > \epsilon) \to 0$. Now, impose the assumption that $E[W_i] = 0$.

$$
\begin{aligned}
P(l'Q_n^{-1}(\hat{Q}_n - Q_n)l > \epsilon) &\leq \frac{1}{\epsilon^2} E\left[\left(l'(Q_n^{-1}(\hat{Q}_n - Q_n))\right)^2\right] \\
&= \frac{1}{\epsilon^2} E\left[\left(l'\left(\frac{1}{\lambda_n}Q_n\right)^{-1}\frac{1}{\lambda_n}(\hat{Q}_n - Q_n)\right)^2\right] \leq \frac{1}{\epsilon^2} E\left[\left(l_0' \frac{1}{\lambda_n}(\hat{Q}_n - Q_n)\right)^2\right]
\end{aligned}
$$

where l_0 is a vector whose entries are all bounded above by some arbitrary constant $K_1 < \infty$ by a similar argument as before. Hence, it suffices to show that $(1/\lambda_n)(\hat{Q}_n - Q_n) \xrightarrow{p} 0_{K\times K}$, where $0_{K\times K}$ is a $K \times K$ matrix of zeroes. Since $\hat{Q}_n - Q_n = \sum_i \sum_{j\in\mathcal{N}_i} \omega_i\omega_j W_i W_j' - E[\omega_i\omega_j W_i W_j']$, it suffices to show convergence elementwise. Let X_i and Y_i denote scalar components of W_i, i.e., $X_i = W_{im}, Y_i = W_{ip}$, where $m, p \in \{1, 2, \cdots, K\}$.

$$
\begin{aligned}
P\left(\frac{1}{\lambda_n}\sum_i \sum_{j\in\mathcal{N}_i} \omega_i\omega_j(X_iY_j - E[X_iY_j]) > \epsilon\right) &\leq \frac{1}{\epsilon^2}\frac{1}{\lambda_n^2} Var\left(\sum_i \sum_{j\in\mathcal{N}_i} \omega_i\omega_j X_iY_j\right) \\
&\leq \frac{1}{\epsilon^2\lambda_n^2}\sum_i \sum_j \sum_{k\in\mathcal{N}_i}\sum_{l\in\mathcal{N}_j} |E[\omega_i\omega_j\omega_k\omega_l X_iX_jY_kY_l] - E[\omega_i\omega_k X_iY_k]E[\omega_j\omega_l X_jY_l]| \\
&\leq \frac{K_0}{\lambda_n^2}\sum_i \sum_j \sum_{k\in\mathcal{N}_i}\sum_{l\in\mathcal{N}_j} |\omega_i\omega_j\omega_k\omega_l|(A_{ij} + A_{il} + A_{kj} + A_{kl}) = o(1)
\end{aligned}
$$

The inequality in the last line is obtained due to Holder's inequality and finite moments. An argument similar to that of Lemma 7 yields the $o(1)$ equality.

Proof of Theorem 1.2

Now assume $E[W_i] = \mu$. Using Lemma 8, $\bar{W} \xrightarrow{p} \mu$ is immediate, i.e., $\bar{W} = \mu + o_P(1)$. To ease notation, let $\tilde{W}_i := W_i - \mu$. Hence, $Q_n = \sum_i \sum_{j\in\mathcal{N}_i} \omega_i\omega_j E[\tilde{W}_i\tilde{W}_j']$.

$$
\begin{aligned}
\hat{Q}_n &= \sum_i \sum_{j\in\mathcal{N}_i} \omega_i\omega_j(W_i - \bar{W})(W_j - \bar{W})' = \sum_i \sum_{j\in\mathcal{N}_i} \omega_i\omega_j(\tilde{W}_i + o_P(1))(\tilde{W}_j + o_P(1))' \\
&= \sum_i \sum_{j\in\mathcal{N}_i} \omega_i\omega_j\tilde{W}_i\tilde{W}_j' + 2\sum_i \sum_{j\in\mathcal{N}_i} \omega_i\omega_j\tilde{W}_i 1_K' o_P(1) + \sum_i \sum_{j\in\mathcal{N}_i} \omega_i\omega_j 1_K 1_K' o_P(1)
\end{aligned}
$$

Since $Q_n^{-1}\sum_i \sum_{j\in\mathcal{N}_i} \omega_i\omega_j\tilde{W}_i\tilde{W}_j' = 1 + o_P(1)$ by Theorem 1.1, it then remains to show that each of the two remaining terms are $o_P(1)$ when pre-multiplied by Q_n^{-1}.

$$
\left(\frac{1}{\lambda_n}Q_n\right)^{-1}\frac{1}{\lambda_n}\sum_i \sum_{j\in\mathcal{N}_i} \omega_i\omega_j 1_K 1_K' \leq K_0\left(\frac{1}{\lambda_n}Q_n\right)^{-1} 1_K 1_K' = O(1)1_K 1_K'
$$

The first inequality is due to the assumption that $(1/\lambda_n)\sum_i \sum_{j\in\mathcal{N}_i} |\omega_i\omega_j| \leq K_0$, and the $O(1)$ term occurs due to the eigenvalues of $(Q_n/\lambda_n)^{-1}$ being bounded above by 1. Take some component $\tilde{X}_i$

of $\tilde{W}_i$. For all $\epsilon > 0$, there exists $M_\epsilon = K_0^2/\epsilon < \infty$ such that:

$$P\left(\left|\frac{1}{\lambda_n}\sum_i \sum_{j\in\mathcal{N}_i} \omega_i\omega_j\tilde{X}_i\right| \geq M_\epsilon\right) \leq \frac{1}{\lambda_n M_\epsilon} E\left[\left|\sum_i \sum_{j\in\mathcal{N}_i} \omega_i\omega_j\tilde{X}_i\right|\right]$$
$$\leq \frac{1}{M_\epsilon}\max_i E[|\tilde{X}_i|]\frac{1}{\lambda_n}\sum_i \sum_{j\in\mathcal{N}_i} |\omega_i\omega_j| \leq \frac{K_0}{K_0^2/\epsilon} = \epsilon$$

Hence, $Q_n^{-1}\sum_i \sum_{j\in\mathcal{N}_i}\omega_i\omega_j\tilde{W}_i 1_K' = 1_K 1_K' O_P(1)$. Since $O_P(1)o_P(1) = o_P(1)$, the result is obtained.

□

B Proof of Propositions

Proof of Proposition 1. We have $\hat{\theta}-\theta = \left(\sum_i \tilde{D}_i^2\right)^{-1}\left(\sum_i \tilde{D}_i u_i\right) = \sum_i \omega_i u_i$, where $\omega_i := \tilde{D}_i/(\sum_j \tilde{D}_j^2)$. Let $\sigma_n^2 := Var(\omega_i u_i)$. Apply Theorem 1 with $K = 1$ to $\sum_i \omega_i u_i$. Assumption 1 and Assumption 2.1 are automatically satisfied for clustered random variable u_i and weight ω_i. Assumption 2.2 is satisfied because

$$\frac{1}{\sigma_n^2}\max_c\left(\sum_{i\in\mathcal{N}_c^C}|\omega_i|\right)^2 \leq \frac{\frac{1}{(\sum_i \tilde{D}_i^2)^2}\max_c\left(\sum_{i\in\mathcal{N}_c^C}|\tilde{D}_i|\right)^2}{K_0\frac{1}{(\sum_i \tilde{D}_i^2)^2}\sum_{c'}\left(\sum_{j\in\mathcal{N}_{c'}^C}|\tilde{D}_j|\right)^2} = \frac{\frac{1}{(\sum_i \tilde{D}_i^2)^2}\max_c\left(\sum_{i\in\mathcal{N}_c^C}|\tilde{D}_i|\right)^2}{\frac{1}{(\sum_i \tilde{D}_i^2)^2}\sum_{c'}\left(\sum_{j\in\mathcal{N}_{c'}^C}|\tilde{D}_j|\right)^2} \to 0$$

where the first inequality comes from Assumption 3.3 and convergence occurs due to Assumption 3.2. Assumption 2.3 is satisfied because

$$\frac{1}{\sigma_n^2}\sum_c \sum_{i,j\in\mathcal{N}_c^C} A_{ij}|\omega_i\omega_j| = \frac{\frac{1}{(\sum_i \tilde{D}_i^2)^2}\sum_{c'}\sum_{i,j\in\mathcal{N}_c^C}|\tilde{D}_i\tilde{D}_j|}{\frac{1}{(\sum_i \tilde{D}_i^2)^2}Var\left(\sum_i \tilde{D}_i u_i\right)} < \infty$$

Hence, Theorem 1 yields $(\hat{\theta}-\theta)/\sigma_n \xrightarrow{d} N(0,1)$.

To prove consistent variance estimation, it suffices to show $(\hat{\sigma}_n^2 - \sigma_n^2)/\sigma_n^2 = o_P(1)$.

$$\hat{\sigma}_n^2 = \sum_i \sum_{j\in\mathcal{N}_i}\omega_i u_i \omega_j u_j - 2\left(\sum_i \sum_{j\in\mathcal{N}_i}\omega_i^2\omega_j u_j\right)(\hat{\theta}-\theta) + \left(\sum_i \sum_{j\in\mathcal{N}_i}\omega_i^2\omega_j^2\right)(\hat{\theta}-\theta)^2$$

By Theorem 1, $\left(\sum_i \sum_{j\in\mathcal{N}_i} \omega_i u_i \omega_j u_j - \sigma_n^2\right)/\sigma_n^2 = o_P(1)$. Since $(\hat{\theta}-\theta)^2/\sigma_n^2 \xrightarrow{d} Z^2 = \chi_1^2$,

$$\frac{\left(\sum_i \sum_{j\in\mathcal{N}_i} \omega_i^2\omega_j^2\right)(\hat{\theta}-\theta)^2}{\sigma_n^2} = \left(\sum_i \sum_{j\in\mathcal{N}_i} \omega_i^2\omega_j^2\right) O_P(1)$$

$$\begin{aligned}\frac{\left(\sum_i \sum_{j\in\mathcal{N}_i} \tilde{D}_i^2\tilde{D}_j^2\right)}{\left(\sum_i \tilde{D}_i^2\right)^4} &\le \frac{\left(\max_i \sum_{j\in\mathcal{N}_i} \tilde{D}_j^2\right)}{\left(\sum_i \tilde{D}_i^2\right)^3}\frac{\sum_i \tilde{D}_i^2}{\sum_i \tilde{D}_i^2} \le \frac{\left(\max_i \sum_{j\in\mathcal{N}_i} \tilde{D}_j^2\right)}{\left(\sum_i \tilde{D}_i^2\right)^2} O(1)\\ &\le \left(\frac{\max_g \left(\sum_{j\in\mathcal{N}_g^G} |\tilde{D}_j|\right)^2}{\sum_{g'}\left(\sum_{j\in\mathcal{N}_{g'}^G} |\tilde{D}_j|\right)^2} + \frac{\max_h \left(\sum_{j\in\mathcal{N}_h^H} |\tilde{D}_j|\right)^2}{\sum_{h'}\left(\sum_{j\in\mathcal{N}_{h'}^H} |\tilde{D}_j|\right)^2}\right) O(1) = o(1)\end{aligned}$$

Convergence occurs due to Assumption 3.2, so $\left(\sum_i \sum_{j\in\mathcal{N}_i} \omega_i^2\omega_j^2\right)(\hat{\theta}-\theta)^2/\sigma_n^2 = o_P(1)$. Finally,

$$\frac{\left(\sum_i \sum_{j\in\mathcal{N}_i} \omega_i^2\omega_j u_j\right)(\hat{\theta}-\theta)}{\sigma_n^2} = \frac{\left(\sum_i \sum_{j\in\mathcal{N}_i} \tilde{D}_i^2\tilde{D}_j u_j\right)}{\sigma_n\left(\sum_i \tilde{D}_i^2\right)^3} O_P(1)$$

Applying Markov and Minkowski inequalities,

$$\begin{aligned}P\left(\frac{\left|\sum_i \sum_{j\in\mathcal{N}_i} \tilde{D}_i^2\tilde{D}_j u_j\right|}{\left(\sum_i \tilde{D}_i^2\right)^3\sigma_n} > \epsilon\right) &\le \frac{1}{\epsilon}\frac{1}{\left(\sum_i \tilde{D}_i^2\right)^3\sigma_n} E\left[\left|\sum_i \sum_{j\in\mathcal{N}_i} \tilde{D}_i^2\tilde{D}_j u_j\right|\right]\\ &\le \frac{1}{\epsilon}\frac{1}{\left(\sum_i \tilde{D}_i^2\right)^3\sigma_n}\sum_i \sum_{j\in\mathcal{N}_i} E|\tilde{D}_i^2\tilde{D}_j u_j| \le \frac{1}{\epsilon}\frac{\max_i \sum_{j\in\mathcal{N}_i} E|\tilde{D}_j u_j|}{\left(\sum_i \tilde{D}_i^2\right)^2\sigma_n}\frac{\sum_i \tilde{D}_i^2}{\sum_i \tilde{D}_i^2} = o(1)\end{aligned}$$

Convergence occurs because

$$\frac{\max_i \left(\sum_{j\in\mathcal{N}_i} E|\tilde{D}_j u_j|\right)^2}{\left(\sum_i \tilde{D}_i^2\right)^2} \le \frac{\max_j E|u_j|^2 \max_i \left(\sum_{j\in\mathcal{N}_i} |\tilde{D}_j|\right)^2}{\left(\sum_i \tilde{D}_i^2\right)^2} = o(1)$$

□

For Proposition 2, I first prove a consistency result.

Lemma 9. *Under Assumption 1, 2.1 and 2.2, and* $E[W_i] = 0\ \forall i$, $||(1/(\sum_i \omega_i))\sum_i \omega_i(W_iW_i' -$

$E[W_iW_i'])|| \xrightarrow{p} 0$.

Proof. It suffices to show convergence elementwise. Let X_i and Y_i denote scalar components of W_i, i.e., $X_i = W_{im}, Y_i = W_{ip}$, where $m, p \in \{1, 2, \cdots, K\}$. By Chebyshev's inequality, and $\max_{m,k} E[W_{mk}^4] < K_0$,

$$P\left(\frac{1}{\sum_i \omega_i}\sum_i \omega_i(X_iY_i - E[X_iY_i]) > \epsilon\right)$$
$$\leq \frac{1}{\epsilon^2}\frac{1}{(\sum_i \omega_i)^2}E\left(\sum_i\sum_{j\in\mathcal{N}_i}\omega_i\omega_j(X_iY_i - E[X_iY_i])(X_jY_j - E[X_jY_j])\right) \leq \frac{K_0}{\epsilon^2\left(\sum_j \omega_j\right)^2}\sum_i\sum_{j\in\mathcal{N}_i}\omega_i\omega_j$$

Hence, it suffices to show $(\sum_i \sum_{j\in\mathcal{N}_i}\omega_i\omega_j)/\left(\sum_j \omega_j\right)^2 = o(1)$. This follows from a similar argument as Lemma 8. □

Proof of Proposition 2. $E[u_i|X_i] = 0$ implies $E[X_iu_i] = 0$ by law of iterated expectations. Since $E[u_i^4|X_i] \leq K_0$, $E[u_i^4X_{ik}^4] = E[E[u_i^4|X_i]X_{ik}^4] \leq K_0E[X_{ik}^4] \leq K_0^2$ is bounded. By Theorem 1, $Q_n^{-1/2}\sum_{i=1}^n X_iu_i \xrightarrow{d} N(0, I_K)$.

To complete the normality result, I show that $S_n^{-1}\hat{S}_n \xrightarrow{p} I_K$, which is the same as showing that $||S_n^{-1}(\hat{S}_n - S_n)|| \xrightarrow{p} 0$. By applying Lemma 9 with $\omega_i = 1$, $(1/n)(\hat{S}_n - S_n) = (1/n)\sum_i(X_iX_i' - E[X_iX_i']) = o_P(1)$. Hence, it suffices that $(S_n/n)^{-1}$ has bounded eigenvalues, i.e., $\lambda_{\min}(S_n/n) \geq K_1 > 0$, which is true by Assumption 4.5. Since $\hat{\beta} - \beta = \hat{S}_n^{-1}\sum_i X_iu_i$, by Slutsky's lemma, $Q_n^{-1/2}S_n(\hat{\beta} - \beta) \xrightarrow{d} N(0, I_K)$.

Next, proceed to consistent variance estimation. Showing that $||Q_n^{-1}\hat{Q}_n - I_K|| = o_P(1)$ is equivalent to showing that, $\forall l \in \mathbb{R}^K$, $l'\left(Q_n^{-1}(\hat{Q}_n - Q_n)\right)l = o_P(1)$.

$$\hat{Q}_n := \sum_i\sum_{j\in\mathcal{N}_i}\hat{u}_i\hat{u}_jX_iX_j' = \sum_i\sum_{j\in\mathcal{N}_i}(u_i - X_i'(\hat{\beta} - \beta))(u_j - X_j'(\hat{\beta} - \beta))X_iX_j'$$
$$= \sum_i\sum_{j\in\mathcal{N}_i}u_iu_jX_iX_j' - 2\left(\sum_i\sum_{j\in\mathcal{N}_i}u_iX_j'(\hat{\beta} - \beta)X_iX_j'\right) + \left(\sum_i\sum_{j\in\mathcal{N}_i}X_i'(\hat{\beta} - \beta)X_j'(\hat{\beta} - \beta)X_iX_j'\right)$$

By Theorem 1, $l'Q_n^{-1}(\sum_i \sum_{j\in\mathcal{N}_i} u_i u_j X_i X_j' - Q_n)l = o_P(1)$. Hence, it remains to show:

$$\left\|Q_n^{-1}\left[-2\left(\sum_i \sum_{j\in\mathcal{N}_i} u_i X_j'(\hat{\beta}-\beta)X_i X_j'\right) + \left(\sum_i \sum_{j\in\mathcal{N}_i} X_i'(\hat{\beta}-\beta)X_j'(\hat{\beta}-\beta)X_i X_j'\right)\right]\right\| = o_P(1)$$

Observe that $X_i'(\hat{\beta}-\beta) = \left(X_i' S_n^{-1} Q_n^{1/2}\right)\left(Q_n^{-1/2} S_n(\hat{\beta}-\beta)\right) = \left(X_i' S_n^{-1} Q_n^{1/2}\right)(Z_K + 1_K o_P(1))$, where 1_K is a K-vector of ones. Hence, addressing the second term,

$$\begin{aligned}
X_i'(\hat{\beta}-\beta)X_j'(\hat{\beta}-\beta) &= \left(X_i' S_n^{-1} Q_n^{1/2}\right)(Z_K + 1_K o_P(1))(Z_K + 1_K o_P(1))'\left(X_j' S_n^{-1} Q_n^{1/2}\right)' \\
&= \left(X_i' S_n^{-1} Q_n^{1/2}\right)(I_K O_P(1) + o_P(1))\left(X_j' S_n^{-1} Q_n^{1/2}\right)' \\
&= X_i' S_n^{-1} Q_n S_n^{-1} X_j O_P(1)
\end{aligned}$$

This implies

$$\begin{aligned}
Q_n^{-1}\left(\sum_i \sum_{j\in\mathcal{N}_i} X_i'(\hat{\beta}-\beta)X_j'(\hat{\beta}-\beta)X_i X_j'\right) &= Q_n^{-1}\left(\sum_i \sum_{j\in\mathcal{N}_i}\left(X_i' S_n^{-1} Q_n S_n^{-1} X_j\right) X_i X_j'\right) O_P(1) \\
= \frac{1}{n^2}\left(\frac{1}{\lambda_n}Q_n\right)^{-1}&\left(\sum_i \sum_{j\in\mathcal{N}_i}\left(X_i'\left(\frac{1}{n}S_n\right)^{-1}\left(\frac{1}{\lambda_n}Q_n\right)\left(\frac{1}{n}S_n\right)^{-1} X_j\right) X_i X_j'\right) O_P(1)
\end{aligned}$$

The eigenvalues of (Q_n/λ_n) are bounded. To see this, it suffices to show that there exists $K_0 < \infty$ such that $\lambda_{\max}(Q_n)/\lambda_n \leq K_0$. Due to finite moments, $Q_n := Var(\sum_i X_i) \leq K_0 1_{K\times K}\sum_c (N_c^C)^2$. Since $(\sum_c (N_c^C)^2)/\lambda_n \leq K_0$ by Assumption 4, $\lambda_n K_0 \geq \sum_c (N_c^C)^2$, which implies $\lambda_n \geq (\sum_c (N_c^C)^2)/K_0$. Hence,

$$\frac{\lambda_{\max}(Q_n)}{\lambda_n} \leq \frac{\sum_c (N_c^C)^2 K_0}{\sum_c (N_c^C)^2 \frac{1}{K_0}} = K_0^2$$

Recall that $(S_n/n)^{-1}$ has bounded eigenvalues. The proof of Theorem 1 also showed that $(Q_n/\lambda_n)^{-1}$

has bounded eigenvalues. By using Markov and Minkowski inequalities,

$$
\begin{aligned}
&P\left(\frac{1}{n^2}\left|l'\left(\frac{1}{\lambda_n}Q_n\right)^{-1}\left(\sum_i\sum_{j\in\mathcal{N}_i}\left(X_i'\left(\frac{1}{n}S_n\right)^{-1}\left(\frac{1}{\lambda_n}Q_n\right)\left(\frac{1}{n}S_n\right)^{-1}X_j\right)X_iX_j'\right)l\right|>\epsilon\right)\\
&\leq\frac{1}{n^2\epsilon}E\left[\left|l'\left(\frac{1}{\lambda_n}Q_n\right)^{-1}\left(\sum_i\sum_{j\in\mathcal{N}_i}\left(X_i'\left(\frac{1}{n}S_n\right)^{-1}\left(\frac{1}{\lambda_n}Q_n\right)\left(\frac{1}{n}S_n\right)^{-1}X_j\right)X_iX_j'\right)l\right|\right]\\
&\leq\frac{1}{n^2\epsilon}\sum_i N_i\max_{m,k}E[X_{mk}^4]K_0\leq\frac{\max_i N_i}{n}\frac{n}{n}K_0\to 0
\end{aligned}
$$

where $K_0\in\mathbb{R}$ is an arbitrary (finite) constant. Convergence occurs due to Assumption 4.2, which implies $\max_i N_i/n\to 0$. This occurs due to the result that $\max_i\sum_{j\in\mathcal{N}_i}|\omega_j|/\left(\sum_j\omega_j\right)=o(1)$ in the proof of Lemma 8, using $\omega_i=1$.

Going back to the first term,

$$
\begin{aligned}
Q_n^{-1}\sum_i\sum_{j\in\mathcal{N}_i}u_iX_j'(\hat\beta-\beta)X_iX_j'&=Q_n^{-1}\sum_i\sum_{j\in\mathcal{N}_i}u_i\left(X_i'S_n^{-1}Q_n^{1/2}\right)(Z_K+1_Ko_P(1))X_iX_j'\\
&=\frac{1}{n\sqrt{\lambda_n}}\left(\frac{1}{\lambda_n}Q_n\right)^{-1}\sum_i\sum_{j\in\mathcal{N}_i}u_i\left(X_i'\left(\frac{1}{n}S_n\right)^{-1}\left(\frac{1}{\lambda_n}Q_n\right)^{1/2}\right)X_iX_j'O_P(1)
\end{aligned}
$$

By using Markov and Minkowski inequalities,

$$
\begin{aligned}
&P\left(\frac{1}{n\sqrt{\lambda_n}}\left|l'\left(\frac{1}{\lambda_n}Q_n\right)^{-1}\sum_i\sum_{j\in\mathcal{N}_i}u_i\left(X_i'\left(\frac{1}{n}S_n\right)^{-1}\left(\frac{1}{\lambda_n}Q_n\right)^{1/2}\right)X_iX_j'l\right|>\epsilon\right)\\
&\leq\frac{1}{n\sqrt{\lambda_n}\epsilon}E\left[\left|l'\left(\frac{1}{\lambda_n}Q_n\right)^{-1}\sum_i\sum_{j\in\mathcal{N}_i}u_i\left(X_i'\left(\frac{1}{n}S_n\right)^{-1}\left(\frac{1}{\lambda_n}Q_n\right)^{1/2}\right)X_iX_j'l\right|\right]\\
&\leq\frac{1}{n\sqrt{\lambda_n}\epsilon}\sum_i\sum_{j\in\mathcal{N}_i}\max_{m_1,m_2,k}E\left[|X_{m_1k}u_{m_1}X_{m_2}^2|\right]K_0\\
&\leq\frac{1}{n\sqrt{\lambda_n}\epsilon}\sum_i N_i\max_{m_1,m_2,k}E\left[|X_{m_1k}u_{m_1}|^2\right]^{1/2}E\left[|X_{m_2}^2|^2\right]^{1/2}K_0\\
&\leq\frac{\max_i N_i}{\sqrt{\lambda_n}}\frac{1}{\epsilon}\max_{m_1,m_2,k}E[X_{m_1k}^2u_{m_1}^2]^{1/2}E[X_{m_2}^4]^{1/2}K_0=o(1)
\end{aligned}
$$

The penultimate inequality occurs due to Holder's inequality. Observe that $\max_i N_i/\sqrt{\lambda_n}=o(1)$ if and only if $\max_c(N_c^C)^2/\lambda_n=o(1)$, which is given by Assumption 4.2. Convergence in the last step occurs because $\max_i N_i/\sqrt{\lambda_n}=o(1)$, and finite moments.

Hence, it has been shown that $Q_n^{-1}\hat{Q}_n \xrightarrow{p} I_K$. Then, $[S_n^{-1}Q_nS_n^{-1}]^{-1}[\hat{S}_n^{-1}\hat{Q}_n\hat{S}_n^{-1}] \xrightarrow{p} I_K$ by the continuous mapping theorem. $\square$

Proof of Corollary 1. By Proposition 2, $(\hat{\beta}_1 - \beta_1)/[V(\hat{\beta})]_{11}^{1/2} \xrightarrow{d} N(0,1)$. Since $\hat{\theta} = \hat{\beta}_1$, $[V(\hat{\beta})]_{11} = V(\hat{\beta}_1) = V(\hat{\theta}) = \sigma_n^2$. Hence, $(\hat{\theta} - \theta)/\sigma_n \xrightarrow{d} N(0,1)$.

A further implication of Proposition 2 is that $[\hat{V}(\hat{\beta})]_{11}/[V(\hat{\beta})]_{11} \xrightarrow{p} 1$. Using theorem 3 of Ding (2021), the Liang-Zeger estimators (Liang and Zeger, 1986) are numerically equivalent regardless of whether the long regression or the residualized regression were used. Since the CGM estimator is a function of the Liang-Zeger estimators, $\hat{\sigma}_n^2 = [\hat{V}(\hat{\beta})]_{11}$. Hence, $\hat{\sigma}_n^2/\sigma_n^2 \xrightarrow{p} 1$. $\square$

References

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

Chen, L. H. and Q.-M. Shao (2004): "Normal approximation under local dependence," *The Annals of Probability*, 32, 1985–2028.

Davezies, L., X. D'Haultfœuille, and Y. Guyonvarch (2021): "Empirical process results for exchangeable arrays," *The Annals of Statistics*, 49, 845–862.

Ding, P. (2021): "The Frisch–Waugh–Lovell theorem for standard errors," *Statistics & Probability Letters*, 168, 108945.

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

Dube, A., T. W. Lester, and M. Reich (2010): "Minimum wage effects across state borders: Estimates using contiguous counties," *The review of economics and statistics*, 92, 945–964.

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

Kallenberg, O. (2005): *Probabilistic symmetries and invariance principles*, vol. 9, Springer.

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

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

Menzel, K. (2021): "Bootstrap With Cluster-Dependence in Two or More Dimensions," *Econometrica*, 89, 2143–2188.

Michalopoulos, S. and E. Papaioannou (2013): "Pre-colonial ethnic institutions and contemporary African development," *Econometrica*, 81, 113–152.

Nunn, N. and L. Wantchekon (2011): "The slave trade and the origins of mistrust in Africa," *American Economic Review*, 101, 3221–52.

Ross, N. (2011): "Fundamentals of Stein's method," *Probability Surveys*, 8, 210–293.