Asymptotic Theory for Two-Way Clustering

Luther Yap *

March 28, 2025

Abstract

This paper proves a new central limit theorem for a sample that exhibits two-way dependence and heterogeneity across clusters. Statistical inference for situations with both two-way dependence and cluster heterogeneity has thus far been an open issue. The existing theory for two-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 two-way clustering is a more robust version of one-way clustering, consistent with applied practice. In an application to linear regression, I show that a standard plug-in variance estimator is valid for inference.

Keywords: Two-Way Clustering; Separate Exchangeability

*Department of Economics, Princeton University, 20 Washington Road, Princeton NJ 08540. Email: lyap@princeton.edu. 28 pages. I thank Michal Kolesár, David Lee, and Ulrich Müller for helpful comments and suggestions. This paper supersedes an earlier paper circulated as "General Conditions for Valid Inference in Multi-Way Clustering".

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. Though more broadly applicable, a common instance of two-way clustering is in linear regressions, where a researcher wants to do inference on the coefficient of interest when the residual is two-way clustered. The variance estimator proposed by Cameron et al. (2011) (henceforth CGM) has thus been widely applied to contexts with such two-way dependence.[1] For instance, Nunn and Wantchekon (2011) clustered on ethnic group 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; Jackson (2018) clustered on teacher and student when studying the effect of the teacher on students' skill; Neumark et al. (2019) clustered on resume and job ad when studying the effect of age on getting a call-back. The existing justification for the asymptotic validity of the CGM estimator and other inference procedures in two-way clustering (e.g., MacKinnon et al. (2021); Davezies et al. (2021); Menzel (2021)) relies on separate exchangeability, which implies homogeneity of clusters, a restriction that is not required in one-way clustering. This paper provides sufficient general conditions for valid inference in two-way clustering by proving that, even with cluster heterogeneity, a central limit theorem holds, and the CGM variance estimator is consistent.

An environment with two-way clustering permits dependence whenever observations share at least one cluster. To fix ideas, consider Jackson (2018): observations of the same student or of the same teacher are plausibly correlated, but two observations of different students and different teachers are assumed to be independent.[2] The CGM variance estimator accommodates such dependence, and a subsequent literature provided a theoretical basis for its validity: MacKinnon et al. (2021) obtained sufficient conditions for validity of the CGM estimator in regression models; Davezies et al. (2021) obtained analogous results for empirical processes. Menzel (2021) also showed the validity of a bootstrap procedure for two-way clustering that is robust to asymptotic non-normalities.[3]

[1] CGM has 3886 citations on Google Scholar at the time of writing.

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

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

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, separate exchangeability implies identical marginal distributions. Separate exchangeability in the student-teacher example thus implies the random variables for all students must be drawn from the same distribution, including students of different cohorts over time. As Wooldridge (2010, p. 146) notes in the discussion of pooled data in his graduate textbook, distributions of variables tend to change over time, so the identical distribution assumption is not usually valid. In other examples, separate exchangeability implies that countries (Michalopoulos and Papaioannou, 2013) and jobs (Neumark et al., 2019) are identically distributed. Applied researchers surely would want size to be controlled in such heterogeneous environments, but the existing theories that rely on separate exchangeability do not imply this result. Further, in linear regressions with regressor X_i and residual u_i, asymptotic theory is applied to $X_i u_i$. Separate exchangeability of the product implies that the regressors must also be separately exchangeable, which is not plausible when the regressors include a time trend, say.

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. Since the only available conditions for the validity of two-way clustering require separate exchangeability, the literature lacks conditions for two-way clustering that generalize one-way clustering and permit heterogeneity over clusters. This paper fills the gap, and thus justifies two-way clustering as a more robust version of one-way clustering.

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 is characterized by a random variable W_i generated from $W_i = \alpha_{g(i)} + \gamma_{h(i)} + \varepsilon_i$, where cluster-specific $\alpha_1, \ldots, \alpha_g, \ldots, \alpha_G, \gamma_1, \ldots, \gamma_h, \ldots, \gamma_H$ and individual-specific $\varepsilon_1, \ldots, \varepsilon_i, \ldots, \varepsilon_n$ are mutually independent. 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 need not be identically distributed. The general conditions provided in this paper permit valid inference even when $\alpha_g, \gamma_h, \varepsilon_i$ are not identically distributed in this model.

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

The main result is a central limit theorem for two-way clustering with heterogeneous cluster sizes and distributions. This result is proven using Stein's method. It adapts the strategy from Ross (2011) Theorem 3.6: I first derive an upper bound on the distance between the distribution of a pivotal statistic and the standard normal, then show that this distance converges to zero asymptotically. This proof strategy hence yields intermediate results on non-asymptotic Berry-Esseen type bounds that provide worst-case bounds on the quality of approximation between the pivotal statistic and the standard normal, which may be of independent interest. 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.

This paper contributes to the literature on multi-way clustering and Stein's method. This paper differs from the existing literature on multi-way clustering (e.g., MacKinnon et al. (2021); Davezies et al. (2021); Menzel (2021); Chiang and Sasaki (2023); Chiang et al. (2024)) in that it does not rely on separate exchangeability. Stein's method has been applied to other contexts such as two-way fixed effects (Verdier, 2020), spillover effects (e.g., Chin (2018), Leung (2022) and Braun and Verdier (2023)), and network formation (e.g., Chandrasekhar and Jackson (2016)). Unlike the aforementioned papers, this paper speaks directly to multi-way clustering, and it makes a modification to the proof of Ross (2011) Theorem 3.6 to obtain the result instead of applying the theorem directly.

2 Setting and Main Result

2.1 Setup

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 $i = 1, \ldots, n$ is the unit of observation. For example, G could denote states and H could denote industries. Clustering in more than two dimensions is possible, and derivations are entirely analogous. This section establishes a central limit theorem (CLT) for $\sum_i W_i$, as $n \to \infty$. Here and in the following, sums are over (subsets of) $\{1, 2, \ldots, n\}$. For $C \in \{G, H\}$, let $\mathcal{N}_c^C$ denote the set of observations in cluster c on dimension C — this setup partitions the sample 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 size of cluster c on dimension $C \in \{G, H\}$ 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 the joint distribution of $\{W_j\}$ for $j \notin \mathcal{N}_{g(i)}^G \cup \mathcal{N}_{h(i)}^H =: \mathcal{N}_i$, i.e., when i and j do not share a cluster on either dimension. Hence, $\mathcal{N}_i$ is the set of observations that are arbitrarily dependent with i. This environment is stated as Assumption 1.

Assumption 1. *With* $\mathcal{N}_i = \mathcal{N}_{g(i)}^G \cup \mathcal{N}_{h(i)}^H$,

(a) $W_i \perp\!\!\!\perp \{W_j\}_{j \notin \mathcal{N}_i}$ *for all* i.

(b) For observations i, j *and* $k \in \mathcal{N}_i, l \in \mathcal{N}_j$ *and all nonstochastic* $\mu \in \mathbb{R}^K$, *if* $j, l \notin (\mathcal{N}_i \cup \mathcal{N}_k)$, *then* $Cov(\mu' W_i W_k' \mu, \mu' W_j W_l' \mu) = 0$.

While the dependence structure is implicitly described in the setup of many clustering papers (e.g., Hansen and Lee (2019); Menzel (2021)), Assumption 1 makes the dependence structure explicit. Assumption 1(a) is a dissociation assumption similar to Definition 3.5 of Ross (2011) required to apply Stein's method. Assumption 1(b) is required because, for a scalar W_i, a crucial step of the proof requires $E[W_i W_j W_k W_l] = E[W_i W_k] E[W_j W_l]$ when j, l do not share any cluster with i, k. Even when $W_i \perp\!\!\!\perp (W_j, W_l)$ and $W_k \perp\!\!\!\perp (W_j, W_l)$, we cannot conclude that $E[W_i W_j W_k W_l] = E[W_i W_k] E[W_j W_l]$ in general, because independence of marginal distributions does not imply independence of the joint distribution. Assumption 1(b) hence makes an assumption on the joint distribution. It can alternatively be stated as $(W_i, W_k) \perp\!\!\!\perp (W_j, W_l)$, which is stronger but more interpretable than the zero-covariance assumption. I further discuss the relationship between Assumption 1 and the existing literature in Section 2.3.

Assumption 1 is agnostic about the dependence structure between W_i and W_j when i and 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); Djogbenou et al. (2019)). Since one-way clustering is a special case of two-way clustering where the H cluster consists of single observations, the result here generalizes the existing results in one-way clustering. In contrast, the existing literature on two-way clustering assumes separate exchangeability that additionally imposes identical distribution over clusters, so it does not generalize the results on one-way clustering.

For positive definite matrix Q, let $\lambda_{\min}(Q)$ denote the smallest eigenvalue of Q. Then, let $Q_n := Var\left(\sum_i W_i\right)$ denote the variance of the sum and $\lambda_n := \lambda_{\min}(Q_n)$ denote its smallest eigenvalue. For example, when $K = 1$, W_i is a scalar and $\lambda_n = Q_n = Var(\sum_i W_i)$. 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:*

(a) $E[W_{ik}^4] \leq K_0$ *for all* i.

(b) $\frac{1}{\lambda_n} \max_c (N_c^C)^2 \to 0$.

(c) $\frac{1}{\lambda_n} \sum_c (N_c^C)^2 \leq K_0$.

Since the objective of this paper is to prove a CLT, Assumption 2 imposes restrictions that rule out data generating processes that are asymptotically non-Gaussian. One such example is explained later in Remark 1. Nonetheless, as reflected in Table 1 of Chiang and Sasaki (2023), such a non-Gaussian regime is an exception rather than the norm when considering a generic separately exchangeable process.

Assumption 2(a) requires the fourth moment to be bounded, which is stronger than the moment condition in one-way clustering (e.g., Equation (7) of Hansen and Lee (2019) and Assumption 1 of Djogbenou et al. (2019)). The proof in one-way clustering usually verifies a Lindeberg condition then applies the Lindeberg CLT because blocks of observations are independent of each other. With two-way dependence, we no longer have independent blocks because each cluster can have observations that are dependent on 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 2.4. By using this strategy, a bounded fourth moment is required.

Assumption 2(b) requires the size of the largest cluster to be small relative to the total variance. This condition mimics the sparsity condition in the networks literature (e.g., Graham (2020)). 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) and Assumption 3 of Djogbenou et al. (2019) for one-way clustering. Assumption 2(b) also rules out having components that are perfectly correlated:

if the components of the vector were perfectly correlated (i.e., $\mu' W_i = 0$ for some $\mu \neq (0, \ldots, 0)'$), then $\lambda_n = 0$. If cluster sizes are uniformly bounded, and $\lambda_n \to \infty$, then Assumption 2(b) is satisfied.[5]

Assumption 2(c) is a summability condition that requires λ_n not to be too small, and requires λ_n to be the same order as $\sum_c (N_c^C)^2$, i.e., $\lambda_n \asymp \sum_c (N_c^C)^2$, $C \in \{G, H\}$.[6] With strictly positive covariance within clusters, $\lambda_n \asymp \sum_c (N_c^C)^2$ is satisfied. However, if the researcher were conservative and clustered on C when the data is indeed iid, then $\lambda_n \asymp n$, which then requires $\sum_c (N_c^C)^2 \asymp n$ for the condition to hold. The assumption that $(1/\lambda_n) \sum_c (N_c^C)^2 \leq K_0$ matches Equation (11) of Hansen and Lee (2019) and Assumption 2 of Djogbenou et al. (2019).

In general, the structure of dependence affects λ_n while the structure of clustering affects $\sum_c (N_c^C)^2$. For example, using the common shocks model of Example 1, $\lambda_n \asymp \sum_c (N_c^C)^2$ when the variances of common shocks α_g and γ_h are non-zero, but if the variances of α_g and γ_h are zero, then $\lambda_n \asymp n$. With a balanced clustering structure where $g \in \{1, \cdots, M\}$, $h \in \{1, \cdots, M\}$ and $N_{gh} = 1$, we have $n = M^2$ and $\sum_c (N_c^C)^2 = M^3$. However, if we have one large cluster, say when all observations are the only observation in their H cluster, i.e., $h(i) = i$, and on the G dimension, the first cluster has size $N_1^G = n^{1/4}$, while all other clusters have size 1, then, $\sum_c (N_c^C)^2 \asymp n^{1/2} + (n - n^{1/4}) \asymp n$.

Remark 1. Assumptions 2(b) and 2(c) rule out the following purely interactive model. 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 and γ_h are iid with mean zero and variances σ_α^2 and σ_γ^2 respectively, so there are M^2 observations. As pointed out by Menzel (2021) Example 1.7, this model has an asymptotic distribution that is non-normal, with no analog in one-way clustering. To see this, $\sum_{g,h} W_{gh}/M = \left(\sum_g \alpha_g/\sqrt{M}\right)\left(\sum_h \gamma_h/\sqrt{M}\right) \xrightarrow{d} Z_1 Z_2$, where Z_1 and Z_2 are independent standard normal random variables. This limiting distribution is also known as Gaussian chaos. Since $\max_g (N_g^G)^2/\lambda_n = M^2/(M^2 \sigma_\alpha^2 \sigma_\gamma^2) = 1/(\sigma_\alpha^2 \sigma_\gamma^2)$ does not converge to 0, Assumption 2(b) fails. Further, $\sum_g (N_g^G)^2/\lambda_n = M^3/(M^2 \sigma_\alpha^2 \sigma_\gamma^2) = M/\sigma_\alpha^2 \sigma_\gamma^2 \to \infty$ violates Assumption 2(c).

Remark 2. Assumptions 2(b) and 2(c) mimic the Lindeberg condition as they divide by the variance of the sum. Nonetheless, if we are willing to make stronger assumptions on variances, we

[5] Assumption 2(b) is hence a more general version of sparsity than having the size of the dependency neighborhood (i.e., the number of observations plausibly correlated with some observation i) being bounded above. The conditions are also comparable with Verdier (2020) in the two-way fixed effects literature: when the neighborhood size is bounded, $\lambda_n \asymp n$, which matches his assumption 2(c).

[6] For sequences a_n and b_n, $a_n \asymp b_n$ if and only if there exists $K_0 < \infty$ such that $a_n/b_n, b_n/a_n \in [-K_0, K_0]$ for all elements in the sequence.

can rewrite the assumptions in terms of primitives. Consider the simple case where W_i is a scalar. If we assume that $E[W_iW_j] \geq c > 0$ for all i and $j \in \mathcal{N}_i$, then Assumption 2(c) is satisfied as $\lambda_n \geq c\left(\sum_g \left(N_g^G\right)^2 + \sum_h \left(N_h^H\right)^2 - \sum_{g,h}\left(N_{(g,h)}^{G\cap H}\right)^2\right) \geq c\sum_g \left(N_g^G\right)^2$ and $\lambda_n \geq c\sum_h (N_h^H)^2$. Then, as long as the largest cluster is small relative to $\sum_c \left(N_c^C\right)^2$, i.e., $\max_c (N_c^C)^2 / \sum_c \left(N_c^C\right)^2 \to 0$, (b) is satisfied. Consequently, a stronger way to state (b) and (c) is that $\max_c (N_c^C)^2 / \sum_c \left(N_c^C\right)^2 \to 0$ and $E[W_iW_j] \geq c > 0$ for all i and $j \in \mathcal{N}_i$.

2.2 Main Result

The main result is that the sum of a sequence of two-way clustered random variables is asymptotically normal. Further, the plug-in variance estimator originally proposed by CGM, $\hat{Q}_n := \sum_i \sum_{j\in\mathcal{N}_i} W_iW_j'$, is consistent. This plug-in expression matches Equation (2.8) of CGM, where W is used here in place of their $\hat{u}$.

Theorem 1. *Under Assumptions 1 and 2,* $Q_n^{-1/2}\sum_i (W_i - E[W_i]) \xrightarrow{d} N(0, I_K)$. *Further, if* $E[W_i] = 0\ \forall i$, *then* $Q_n^{-1/2}\hat{Q}_n Q_n^{-1/2} \xrightarrow{p} I_K$.

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, and clustering by county and state (as counties are nested in states). A sufficient condition for consistent variance estimation is $E[W_i] = 0$, similar 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. If $E[W_i] = \mu$ for all i as in Theorem 4 of Hansen and Lee (2019), consistency can be obtained under the same assumptions.[7]

Remark 3. A double array of random vectors, where the random vector W_{in} is indexed by n, can be accommodated. In this setup, with $K = 1$ for simplicity, we can define $\mathcal{W}_n$ as the class of distributions of n random variables $\{W_{in}\}_{i=1}^n$ that satisfy Assumptions 1, 2(a), 2(c), and that for $C \in \{G, H\}$, there exists $K_0 < \infty$ and $\epsilon > 0$ such that $\frac{1}{\lambda_n}\max_c (N_c^C)^2 \leq K_0 n^{-\epsilon}$ (which is a modification of Assumption 2(b)). Then, for $R_n := Q_n^{-1/2}\sum_i (W_{in} - E[W_{in}])$, d_W denoting the Wasserstein distance[8], and Z denoting the standard normal random variable, we have

[7] Since $(1/n)\sum_i W_i$ consistently estimates μ, the result follows by using $\tilde{W}_i = W_i - \mu$ in place of W_i.
[8] See details in Section 2.4.

$\sup_{\{W_{in}\}_{i=1}^n \in \mathcal{W}_n} d_W(R_n, Z) \to 0$ as $n \to 0$. Consequently, normality holds for a double array uniformly over distributions in $\mathcal{W}_n$. The proof of such a result is the same as the proof of Theorem 1. In the double array, Assumption 2(c) rules out a balanced setting where component variances are of order smaller than one: there are $O(M^3)$ variance and covariance objects in λ_n, so when they are of order r_M, $\lambda_n = O(M^3 r_M)$ while $\sum_c (N_c^C)^2 = M^3$. Then, any r_M that decays at any order of M violates Assumption 2(c).[9]

Remark 4. While the CGM variance estimator is valid in this environment without separate exchangeability, we must be more careful with bootstrap methods that were developed under separate exchangeability (e.g., Menzel (2021), MacKinnon et al. (2021)). Bootstrap methods often resample cluster-specific means, such as $\hat{\alpha}_g = (1/N_g^G) \sum_{i \in \mathcal{N}_g^G} W_i - (1/n) \sum_i W_i$. Consider a data generating process where, with $\alpha_g = (1/N_g^G) \sum_{i \in \mathcal{N}_g^G} [W_i] - (1/n) \sum_i E[W_i]$, odd-numbered g clusters have $\alpha_g = -1$ and even-numbered g clusters have $\alpha_g = 2$, and there are twice as many units in odd-numbered clusters as even-numbered clusters. Such a process is not exchangeable. Resampling $\hat{\alpha}_g$'s with equal probability results in a positive mean, which invalidates naive bootstrap procedures.

The following two subsections discuss technicalities on the dependence structure and the proof sketch. A general-interest audience may wish to proceed immediately to Section 3.

2.3 Discussion of Dependence Structure

To compare the setup used in Assumption 1 to the existing literature, I carefully define a few terms used in Menzel (2021), whose setup uses a *dissociated separately exchangeable* array. Let Y_{gh} denote an infinite array of observations in cluster g on the G dimension and cluster h on the H dimension. Y_{gh} is a *separately exchangeable* array if, for any integers $\tilde{G}, \tilde{H}$ and permutations

[9] These assumptions are primarily used in Lemmas 6 and 7 of the appendix, so an alternative way to proceed with the proof of normality is to assume their conclusions $\frac{1}{\lambda_n^3} \sum_i \sum_{j,k \in \mathcal{N}_i} E\left[|W_i| W_j W_k\right] = o(1)$ and $\frac{1}{\lambda_n^4} Var\left(\sum_i \sum_{j \in \mathcal{N}_i} W_i W_j\right) = o(1)$ directly. In the balanced design where the second, third and fourth moments decay at the same rate r_M, $\sum_i \sum_{j,k \in \mathcal{N}_i} E\left[|W_i| W_j W_k\right] = O\left(M^4 r_M\right)$ and $Var\left(\sum_i \sum_{j \in \mathcal{N}_i} W_i W_j\right) = O\left(M^5 r_M\right)$. Then, $\frac{1}{\sigma_n^3} \sum_i \sum_{j,k \in \mathcal{N}_i} E\left[|W_i| W_j W_k\right] = O\left(M^{-1/2} r_M^{-1/2}\right)$ and $\frac{1}{\sigma_n^4} Var\left(\sum_i \sum_{j \in \mathcal{N}_i} W_i W_j\right) = O\left(M^{-1} r_M^{-1}\right)$. Hence, the conclusions can still hold if these moments decay at a rate slower than M: for instance, if $r_M = M^{-1/2}$, then $O\left(M^{-1} r_M^{-1}\right) = O\left(M^{-1/2}\right) = o(1)$.

$\pi_1 : \{1, \ldots, \tilde{G}\} \to \{1, \ldots, \tilde{G}\}$ and $\pi_2 : \{1, \ldots, \tilde{H}\} \to \{1, \ldots, \tilde{H}\}$, we have:

$$(Y_{\pi_1(g)\pi_2(h)})_{g,h} \overset{d}{=} (Y_{gh})_{g,h},$$

where $\overset{d}{=}$ denotes equality in distribution.[10] Such an array is *dissociated* if, for any $G_0, H_0 \geq 1$, $(Y_{gh})_{g=1,h=1}^{g=G_0,h=H_0}$ is independent of $(Y_{gh})_{g>G_0,h>H_0}$. Dissociation is how the existing literature formally incorporates the multi-way clustering structure. Separate exchangeability implies that the cluster indices are not meaningful, and it is stronger than having identical distributions across clusters. This environment is a special case of Assumption 1, as the following proposition claims.

Proposition 1. *A dissociated separately exchangeable array satisfies Assumption 1.*

One formal generalization of separate exchangeability is relative exchangeability in Crane and Towsner (2018), where exchangeability need not hold for the full sample, but only within each stratum (i.e., relative to some structure), such as within cohorts of students. However, such a generalization is insufficient in finite-population settings with two-way clustered sampling. Suppose there is a finite superpopulation of outcomes $\{Y_i\}_{i=1}^n$ that is nonstochastic, and two-way clustered sampling by ethnic groups and district (e.g., Nunn and Wantchekon (2011)): a subset of districts are independently sampled, a subset of ethnic groups are independently sampled, and units are sampled from the intersections of districts and ethnic groups that are both sampled. We are interested in the mean of Y in the finite superpopulation. With R_i denoting the indicator for whether individual i is sampled and hence observed, the observed random variable is $W_i = R_i Y_i$. Even though R_i is separately exchangeable, $R_i Y_i$ is neither separately exchangeable nor relatively exchangeable due to conditioning on $\{Y_i\}_{i=1}^n$, but Assumption 1 is still satisfied.

2.4 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) Theorem 3.6 to obtain an upper bound on

[10] Due to Kallenberg (2005), $\{Y_{gh}\}_{g\geq 1, h\geq 1}$ is separately exchangeable if and only if there exists a representation $Y_{gh} = f(\alpha_g, \gamma_h, \varepsilon_{gh})$, where $(\alpha_g, \gamma_h, \varepsilon_{gh}) \overset{iid}{\sim} U[0,1]$. The setup in this paper does not require $(\alpha_g, \gamma_h, \varepsilon_{gh}) \overset{iid}{\sim} U[0,1]$, which allows some data generating processes ruled out by separate exchangeability. For example, suppose there is some $Y_{gh} = -Y_{gh'}$. These random variables are allowed to be perfectly correlated under Assumption 1 since they share a cluster. However, the representation $f(.)$ implies $E[Y_{gh}|\alpha_g] \perp\!\!\!\perp E[Y_{gh'}|\alpha_g]$, so no such representation exists.

the Wasserstein distance between a pivotal statistic and the standard normal random variable. By exploiting the two-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 $K = 1$, and $E[W_i] = 0$. Let $\sigma_n^2 := Q_n$, $R = \sum_i W_i/\sigma_n$, and $Z \sim N(0,1)$. Lemma 4 in Appendix A provides an explicit bound on the Wasserstein distance between R and Z. With $d_W(.)$ denoting the Wasserstein distance, and $d_K(.)$ denoting the Kolmogorov distance, Proposition 1.2 from Ross (2011) implies that $d_K(R,Z) \leq (2/\pi)^{1/4}\sqrt{d_W(R,Z)}$.[11] The Kolmogorov distance is the maximal distance between two CDF's, so it is informative of the maximum distance between the distribution of the pivotal statistic and the standard normal. If $d_W(R,Z) \to 0$, then $d_K(R,Z) \to 0$, so the statistic R is asymptotically normal. By using Assumption 1 to adapt the proof of Theorem 3.6 in Ross (2011),

$$d_W(R,Z) \leq \frac{1}{\sigma_n^3}\sum_i \sum_{j,k\in\mathcal{N}_i} E[|W_i|W_jW_k] + \frac{\sqrt{2}}{\sqrt{\pi}\sigma_n^2}\sqrt{Var\left(\sum_i \sum_{j\in\mathcal{N}_i} W_iW_j\right)}. \tag{1}$$

This inequality is informative of the quality of the normal approximation. This bound on the Wasserstein distance (and hence the Kolmogorov distance) is non-asymptotic, and of the Berry-Esseen type, thereby giving a worst-case bound on the distance between the pivotal statistic and the standard normal. Ross (2011) Theorem 3.6 is a corollary of (1): the term with the third moment is immediate, while the term with the fourth moment results from the last line of their proof.

At this point, my proof departs from the proofs in the existing statistical literature that employ Stein's method (e.g., Chen and Shao (2004); Janisch and Lehéricy (2024)). Let $N_i := |\mathcal{N}_i|$. Hölder's inequality is employed on objects such as $\sum_i \sum_{j,k\in\mathcal{N}_i} E[|W_i|W_jW_k]$. The existing literature uses the L^1 norm of moments $E[|W_i|^3]$ and the 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 the L^1 norm of

[11] For completeness, I define both distance metrics using the notation in Ross (2011). For two probability measures μ and ν, and family of test functions $\mathcal{H}$, distances are defined as:

$$d_{\mathcal{H}}(\mu,\nu) = \sup_{h\in\mathcal{H}}\left|\int h(x)d\mu(x) - \int h(x)d\nu(x)\right|.$$

As special cases, the Kolmogorov distance uses $\mathcal{H} = \{1[\cdot \leq x] : x \in \mathbb{R}\}$ and the Wasserstein distance uses $\mathcal{H} = \{h : \mathbb{R} \to \mathbb{R} : |h(x) - h(y)| \leq |x - y|\}$.

N_i, resulting in $\max_m E[|W_m|^3] \sum_i N_i^2$. Hence,

$$\frac{1}{\sigma_n^3} \sum_i \sum_{j,k \in \mathcal{N}_i} E[|W_i| W_j W_k] \leq \frac{1}{\sigma_n^3} \max_m E[|W_m|^3] \sum_i N_i^2.$$

Since $\max_m E[|W_m|^3]$ is bounded by Assumption 2(a), it suffices to show that $\sum_i N_i^2 / \sigma_n^3 \to 0$. Due to Assumption 1(a), $N_i \leq N_{g(i)}^G + N_{h(i)}^H$, 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 $\lambda_n = \sigma_n^2$ when $K = 1$, $\max_{g,h}(N_g^G + N_h^H)/\sigma_n \to 0$ by Assumption 2(b) and the final term $\left(\sum_g (N_g^G)^2 + \sum_h (N_h^H)^2 \right) / \sigma_n^2$ is bounded by Assumption 2(c). Hence, the term is $o(1)$.

A similar argument is made for the fourth moment that features in $Var\left(\sum_i \sum_{j \in \mathcal{N}_i} W_i W_j \right)$. 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 the existing intermediate results.

Remark 5. By modifying the proof of Theorem 3.6 in Ross (2011), the conditions in this paper permit some forms of heterogeneity in cluster sizes that Theorem 3.6 of Ross (2011) does not. The following is one such example. All observations are the only observation in their H cluster, i.e., $h(i) = i$. On the G dimension, the first cluster has size $N_1^G = n^{1/4}$, while all other clusters have size 1. Then, with positive correlation for units within each cluster such that $\lambda_n \asymp \sum_c (N_c^C)^2$, we have $\lambda_n \asymp n^{1/2} + (n - n^{1/4}) \asymp n$ and $(N_1^G)^2 / \lambda_n \asymp n^{1/2}/n = o(1)$, so the conditions of Theorem 1 are satisfied. However, Theorem 3.6 of Ross (2011) bounds the Wasserstein distance by $\left(N_1^2 / \lambda_n^{3/2} \right) \sum_i E|W_i|^3$ and a term that involves the fourth moment. We have $N_1^2 / \lambda_n^{3/2} \sum_i E|W_i|^3 \asymp n^{-1} \sum_i E|W_i|^3 \neq o(1)$, so we may not obtain convergence. This example similarly rules out using results from Janisch and Lehéricy (2024) directly.

Remark 6. There are several early papers in probability theory that deliver similar results, but are insufficient for Theorem 1. For instance, Theorem 2 of Janson (1988) is a central limit theorem

that uses the condition (with $m = 3$):

$$\left(\frac{n}{\max_i N_i}\right)^{1/3} \frac{(\max_i N_i)\max_i |W_i|}{\sigma_n} = \left(\frac{n}{\sigma_n^3}\left(\max_i N_i\right)^2\right)^{1/3} \max_i |W_i| \to 0.$$

In this proof sketch, I have shown that $\sum_i N_i^2/\sigma_n^3 \to 0$, but $n(\max_i N_i)^2/\sigma_n^3 \geq \sum_i N_i^2/\sigma_n^3$, so the Janson (1988) condition need not hold in this environment.

3 Theory for Least Squares Regression

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

$$Y_i = X_i'\beta + u_i.$$

with $X_i \in \mathbb{R}^K$. We are interested in estimating β. Suppose $E[X_i u_i] = 0$ for all i, and (X_i', u_i) is allowed to be two-way clustered. The standard OLS estimator is:

$$\hat{\beta} = \left(\sum_i X_i X_i'\right)^{-1}\left(\sum_i X_i Y_i\right) = \beta + \left(\sum_i X_i X_i'\right)^{-1}\left(\sum_i X_i u_i\right).$$

This object is assumed to be well-defined in that $\sum_i X_i X_i'$ is invertible. Define $S_n := \sum_i E[X_i X_i']$ and $Q_n := Var\left(\sum_i X_i u_i\right)$, 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 3 provides sufficient conditions for the estimator $\hat{\beta}$ to be asymptotically normal and for the CGM variance estimator to be consistent. 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 the rank condition in OLS.

Assumption 3. *For $C \in \{G, H\}$, and $k \in \{1, 2, \cdots, K\}$, there exists $K_0 < \infty$ and $K_1 > 0$ such that:*

(a) $E[u_i^4|X_i] \leq K_0$, $E[X_{ik}^4] \leq K_0$, $E[X_i u_i] = 0$ *for all* i.

(b) $\frac{1}{\lambda_n} \max_c (N_c^C)^2 \to 0$.

(c) $\frac{1}{\lambda_n} \sum_c (N_c^C)^2 \leq K_0$.

(d) $(X_i', u_i)' \perp\!\!\!\perp \{(X_j', u_j)'\}_{j \notin \mathcal{N}_i}$. *For observations* i, j *and* $k \in \mathcal{N}_i, l \in \mathcal{N}_j$ *and all nonstochastic* $\mu \in \mathbb{R}^K$, *if* $j, l \notin (\mathcal{N}_i \cup \mathcal{N}_k)$, *then* $(X_i', u_i, X_k', u_k)' \perp\!\!\!\perp (X_j', u_j, X_l', u_l)'$.

(e) $\lambda_{\min}\left(\frac{1}{n} S_n\right) \geq K_1$.

Proposition 2. *Under Assumption 3,* $Q_n^{-1/2} S_n (\hat{\beta} - \beta) \xrightarrow{d} N(0, I_K)$, *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 performing F tests on a subvector of β. The proof of Proposition 2 proceeds by applying Theorem 1 to $\sum_i X_i u_i$, then showing that $S_n^{-1} \hat{S}_n \xrightarrow{p} I_K$, which uses the rank condition of Assumption 3(e). It then remains to show that the remainder terms are asymptotically negligible.

The practitioner's takeaway from Proposition 2 is that the existing CGM variance estimator can be used for valid inference with two-way clustering. The result provides the formal theoretical guarantee for using the estimator, under conditions that permit heterogeneity across clusters.

Besides the application mentioned, Theorem 1 also has implications on the conditions required for valid inference when the random variable is two-way clustered in many other econometric models, including design-based settings and instrumental variables models. This theory is especially relevant for design-based settings where the researcher conditions on potential outcomes, so the random variable cannot be separately exchangeable by construction — see Xu and Yap (2024), for instance. Inference for estimators based on moment conditions can be done by straightforward application of Theorem 1 as in linear regression. Practically, this paper has shown that the popular CGM estimator is robust in an environment without separate exchangeability, but practitioners should exercise caution when applying bootstrap methods to environments that are not separately exchangeable. While the results are presented for two-way clustering, they can be easily extended to clustering on three or more dimensions.

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. 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, then (v) 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. $1[A]$ is an indicator function that takes value 1 if A is true and 0 otherwise. 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. Denote the variance of the sum of the scalar random variable X_i as $\sigma_n^2 := Var\left(\sum_i X_i\right)$. We are interested in the asymptotic distribution of $(1/\sigma_n)\sum_i X_i$.

Assumption 4. *For $C \in \{G,H\}$, there exists $K_0 < \infty$ such that:*

(a) $E[X_i] = 0$ *and* $E[X_i^4] \leq K_0 < \infty$ *for all i;*

(b) $\frac{1}{\sigma_n^2}\max_c \left(N_c^C\right)^2 \to 0$*;*

(c) $\frac{1}{\sigma_n^2}\sum_c \left(N_c^C\right)^2 \leq K_0 < \infty$*;*

(d) $X_i \perp\!\!\!\perp \{X_j\}_{j\notin\mathcal{N}_i}$*; and*

(e) for observations $i,j,k \in \mathcal{N}_i, l \in \mathcal{N}_j$, if $(\mathcal{N}_i \cup \mathcal{N}_k) \cap (\mathcal{N}_j \cup \mathcal{N}_l) = \emptyset$, then $Cov(X_iX_k, X_jX_l) = 0$.

Lemma 1. *Under Assumption 4,* $(1/\sigma_n)\sum_i X_i \xrightarrow{d} N(0,1)$*, where* $\sigma_n^2 := Var\left(\sum_i X_i\right)$*. Further, using feasible estimator* $\hat{\sigma}_n^2 := \sum_i \sum_{j\in\mathcal{N}_i} X_i X_j$*,* $\hat{\sigma}_n^2/\sigma_n^2 \xrightarrow{p} 1$*.*

Lemma 2. *(Theorem 3.1 of Ross (2011)) If* R *is a random variable,* 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* $d_W(R,Z) \leq \sup_{f\in\mathcal{F}} |E[f'(R) - Rf(R)]|$*.*

The proofs of Lemmas 3 and 4 follow Ross (2011) Theorem 3.6 up to Equations (3.11) and (3.12).

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* $R_i := \sum_{j\notin\mathcal{N}_i} X_j/\sigma_n$*, then, for all* $f \in \mathcal{F}$*,*

$$E[Rf(R)] = E\left[\frac{1}{\sigma_n}\sum_i X_i(f(R) - f(R_i) - (R - R_i)f'(R))\right] + E\left[\frac{1}{\sigma_n}\sum_i X_i(R - R_i)f'(R)\right].$$

Proof. Start from right-hand side:

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

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, $E[X_i f(R_i)] = 0$. □

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,*

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

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 $R_i := \sum_{j \notin \mathcal{N}_i} X_j/\sigma_n$, so X_i is independent of R_i. Then,

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

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 (2), 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:

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

Turning now to the second term,

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

□

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 4,* $\frac{1}{\sigma_n^3}\sum_i \sum_{j,k\in\mathcal{N}_i} E\left[|X_i|X_jX_k\right] = o(1)$.

Proof. Using Lemma 5,

$$\begin{aligned}\frac{1}{\sigma_n^3}\sum_i \sum_{j,k\in\mathcal{N}_i} E\left[|X_i|X_jX_k\right] &\leq \frac{1}{\sigma_n^3}\sum_i \sum_{j,k\in\mathcal{N}_i} E\left[||X_i|X_jX_k|\right] \\ &\leq \frac{\max_m E[|X_m|^3]}{\sigma_n^3}\sum_i \sum_{j,k\in\mathcal{N}_i} 1 = \frac{\max_m E[|X_m|^3]}{\sigma_n^3}\sum_i N_i^2.\end{aligned}$$

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 $o(1)$. Due to Assumption 1, $N_i \leq N_{g(i)}^G + N_{h(i)}^H$, so

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

$\max_{g,h}(N_g^G + N_h^H)/\sigma_n \to 0$ by Assumption 2(b) and the final term $\left(\sum_g (N_g^G)^2 + \sum_h (N_h^H)^2\right)/\sigma_n^2$ is bounded by Assumption 2(c). Hence, the term is $o(1)$. □

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

Proof. Observe that:

$$\frac{1}{\sigma_n^4} Var\left(\sum_i \sum_{j\in\mathcal{N}_i} X_i X_j\right) = \frac{1}{\sigma_n^4} E\left[\left(\sum_i \sum_{j\in\mathcal{N}_i} X_i X_j\right)^2\right] - \frac{1}{\sigma_n^4}\left(\sum_i \sum_{j\in\mathcal{N}_i} E[X_i X_j]\right)^2$$
$$= \frac{1}{\sigma_n^4} \sum_i \sum_j \sum_{k\in\mathcal{N}_i} \sum_{l\in\mathcal{N}_j} (E[X_i X_j X_k X_l] - E[X_i X_k] E[X_j X_l]).$$

Due to Assumption 1(b), when j, l do not share any cluster with i, k, $E[X_i X_j X_k X_l] = E[X_i X_k] E[X_j X_l]$. Hence, we only have to consider terms where there is at least one pair that shares a cluster. Let $A_{ij} := 1[j \in \mathcal{N}_i]$. With finite 4th moment and Lemma 5, using the same argument as the proof of Lemma 6, it is sufficient to show

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

It is sufficient to consider the A_{ij} term because the other terms are symmetric. In particular,

$$\sum_i \sum_j \sum_{k\in\mathcal{N}_i} \sum_{l\in\mathcal{N}_j} A_{il} = \sum_i \sum_{k\in\mathcal{N}_i} \sum_l \sum_{j\in\mathcal{N}_l} A_{il} = \sum_i \sum_j \sum_{k\in\mathcal{N}_i} \sum_{l\in\mathcal{N}_j} A_{ij},$$
$$\sum_i \sum_j \sum_{k\in\mathcal{N}_i} \sum_{l\in\mathcal{N}_j} A_{kj} = \sum_k \sum_{i\in\mathcal{N}_k} \sum_j \sum_{l\in\mathcal{N}_j} A_{kj} = \sum_i \sum_j \sum_{k\in\mathcal{N}_i} \sum_{l\in\mathcal{N}_j} A_{ij}, \text{ and}$$
$$\sum_i \sum_j \sum_{k\in\mathcal{N}_i} \sum_{l\in\mathcal{N}_j} A_{kl} = \sum_k \sum_l \sum_{i\in\mathcal{N}_k} \sum_{j\in\mathcal{N}_l} A_{kl} = \sum_i \sum_j \sum_{k\in\mathcal{N}_i} \sum_{l\in\mathcal{N}_j} A_{ij}.$$

Considering the A_{ij} term,

$$\sum_i \sum_j \sum_{k\in\mathcal{N}_i} \sum_{l\in\mathcal{N}_j} A_{ij} \leq \sum_i \left(\sum_{j\in\mathcal{N}^G_{g(i)}} + \sum_{j\in\mathcal{N}^H_{h(i)}}\right)\left(\sum_{k\in\mathcal{N}^G_{g(i)}} + \sum_{k\in\mathcal{N}^H_{h(i)}}\right)\left(\sum_{l\in\mathcal{N}^G_{g(j)}} + \sum_{l\in\mathcal{N}^H_{h(j)}}\right) A_{ij}.$$

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

$$\sum_i \sum_{j\in\mathcal{N}^G_{g(i)}} \sum_{k\in\mathcal{N}^G_{g(i)}} \sum_{l\in\mathcal{N}^G_{g(j)}} A_{ij} = \sum_g \sum_{i,j,k,l\in\mathcal{N}^G_g} A_{ij} = \sum_g (N^G_g)^4.$$

The first equality in the equation above follows from how $\sum_i = \sum_g \sum_{i\in\mathcal{N}^G_g}$ and that if $j \in \mathcal{N}^G_{g(i)}$,

then i and j share the same g and hence $\sum_{l\in\mathcal{N}^G_{g(j)}} = \sum_{l\in\mathcal{N}^G_{g(i)}}$. The second equality occurs as $A_{ij} = 1$ when i and j share the same g cluster. With this equality, observe that $\sum_g (N^G_g)^4 = \left(\max_g (N^G_g)^2\right)\sum_g (N^G_g)^2$. Since $\frac{1}{\sigma_n^2}\max_g \left(N^G_g\right)^2 = o(1)$ and $\frac{1}{\sigma_n^2}\sum_g \sum_{i,j\in\mathcal{N}^G_g} A_{ij} \leq \frac{1}{\sigma_n^2}\sum_g (N^G_g)^2 < \infty$ by Assumption 4, these terms are $o(1)$ when divided by σ_n^4.

An upper bound can similarly be derived for the interactive terms. To explain the steps carefully, I label the equalities and inequalities (i) to (iv):

$$
\begin{aligned}
&\sum_i \sum_{j\in\mathcal{N}^G_{g(i)}} \sum_{k\in\mathcal{N}^G_{g(i)}} \sum_{l\in\mathcal{N}^H_{h(j)}} A_{ij} \\
&\quad \overset{(i)}{=} \sum_{i,j,k}\sum_g 1[i\in\mathcal{N}^G_g]1[j\in\mathcal{N}^G_g]1[k\in\mathcal{N}^G_g]\sum_l\sum_h 1[j\in\mathcal{N}^H_h]1[l\in\mathcal{N}^H_h]A_{ij} \\
&\quad \overset{(ii)}{=} \sum_g\sum_{i,j,k} 1[i\in\mathcal{N}^G_g]1[j\in\mathcal{N}^G_g]1[k\in\mathcal{N}^G_g]A_{ij}\sum_h\sum_l 1[j\in\mathcal{N}^H_h]1[l\in\mathcal{N}^H_h] \\
&\quad \overset{(iii)}{\leq} \left(\max_j \sum_h\sum_l 1[j\in\mathcal{N}^H_h]1[l\in\mathcal{N}^H_h]\right)\left(\sum_g \sum_{i,j,k\in\mathcal{N}^G_g} A_{ij}\right) \\
&\quad \overset{(iv)}{\leq} \left(\max_h \sum_{l\in\mathcal{N}^H_h} 1\right)\left(\max_g \sum_{k\in\mathcal{N}^G_g} 1\right)\left(\sum_g \sum_{i,j\in\mathcal{N}^G_g} A_{ij}\right) = \left(\max_h N^H_h\right)\left(\max_g N^G_g\right)\left(\sum_g (N^G_g)^2\right).
\end{aligned}
$$

The equality in (i) is obtained by transforming the conditional sums into sums over products of indicators. The equality in (ii) is obtained from commutative and associative properties of additional and multiplication. In step (iii), the inequality is obtained by using the upper bound on the innermost sum over h and l. In step (iv), to see how $\max_j \sum_h \sum_l 1[j\in\mathcal{N}^H_h]1[l\in\mathcal{N}^H_h] = \max_h \sum_{l\in\mathcal{N}^H_h} 1$, observe that once we choose the index j, the indicator $1[j\in\mathcal{N}^H_h]$ can only take value 1 for one particular h, so the maximum occurs when we choose a corresponding h that results in the largest $\sum_{l\in\mathcal{N}^H_h} 1$. The inequality in (iv) is due to extracting $\left(\max_g \sum_{k\in\mathcal{N}^G_g} 1\right)$ from $\left(\sum_g \sum_{i,j,k\in\mathcal{N}^G_g} A_{ij}\right)$. Since $\sum_g (N^G_g)^2/\sigma_n^2 \leq K_0$ and $\max_g N^G_g/\sigma_n = o(1)$,

$$
\frac{1}{\sigma_n^4}\sum_i \sum_{j\in\mathcal{N}^G_{g(i)}} \sum_{k\in\mathcal{N}^G_{g(i)}} \sum_{l\in\mathcal{N}^H_{h(j)}} A_{ij} \leq \left(\frac{1}{\sigma_n}\max_h N^H_h\right)\left(\frac{1}{\sigma_n}\max_g N^G_g\right)\left(\frac{1}{\sigma_n^2}\sum_g (N^G_g)^2\right) = o(1).
$$

□

Proof of Lemma 1. Apply Lemma 4 to obtain:

$$d_W(R,Z) \leq \frac{1}{\sigma_n^3}\sum_i \sum_{j,k\in\mathcal{N}_i} E[|X_i|X_jX_k] + \frac{\sqrt{2}}{\sqrt{\pi}\sigma_n^2}\sqrt{Var\left(\sum_i\sum_{j\in\mathcal{N}_i} X_iX_j\right)}.$$

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,

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

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

Proof of Theorem 1. To show that $Q_n^{-1/2}\sum_i(W_i - E[W_i]) \xrightarrow{d} N(0, I_K)$, due to the Cramer-Wold device, it suffices to show that $\forall\mu \in \mathbb{R}^K$, $\mu' Q_n^{-1/2}\sum_i(W_i - E[W_i]) \xrightarrow{d} \mu' N(0, I_K)$. If μ is a vector of zeroes, then $\mu' Q_n^{-1/2}\sum_i(W_i - E[W_i]) \xrightarrow{d} \mu' N(0, I_K)$ is immediate. For $||\mu|| > 0$, it suffices to show $(1/||\mu||)\mu' Q_n^{-1/2}\sum_i(W_i - E[W_i]) \xrightarrow{d} (1/||\mu||)\mu' N(0, I_K) = N(0,1)$. Without loss of generality, we can set $||\mu|| = 1$. For all nonstochastic $\mu \in \mathbb{R}^K\backslash\{0\}$, let $\sigma_n^2(\mu) := Var\left(\sum_i \mu'\left(Q_n/\lambda_n\right)^{-1/2}(W_i - E[W_i])\right)$, so the following hold:

1. $E\left[\left(\mu'\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}(W_i - E[W_i])\right)\right] = 0$ and $E\left[\left(\mu'\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}(W_i - E[W_i])\right)^4\right] \leq K_0$ for all i.
2. $\frac{1}{\sigma_n^2(\mu)}\max_c\left(N_c^C\right)^2 \to 0$.
3. $\frac{1}{\sigma_n^2(\mu)}\sum_c (N_c^C)^2 \leq K_0$.
4. $\left(\mu'\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}(W_i - E[W_i])\right) \perp\!\!\!\perp \left\{\left(\mu'\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}W_j\right)\right\}_{j\notin\mathcal{N}_i}$.
5. For observations $i,j,k \in \mathcal{N}_i, l \in \mathcal{N}_j$, if $(\mathcal{N}_i \cup \mathcal{N}_k) \cap (\mathcal{N}_j \cup \mathcal{N}_l) = \emptyset$, then

$$Cov\left(\mu'\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}W_i\mu'\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}W_k, \mu'\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}W_j\mu'\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}W_l\right) = 0.$$

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, which implies $|\mu'(Q_n/\lambda_n)^{-1/2}| \leq K_1$ for

some arbitrary constant $K_1 < \infty$. Item 1 then follows from Assumption 2(a). Observe that $\sigma_n^2(\mu) = \mu'(Q_n/\lambda_n)^{-1/2}Q_n(Q_n/\lambda_n)^{-1/2}\mu = \lambda_n$. Then, Assumption 2(b) yields item 2, and Assumption 2(c) yields item 3. Item 4 is immediate from Assumption 1(a), and item 5 from Assumption 1(b). By applying Lemma 1, $(1/\sigma_n(\mu))\mu'(Q_n/\lambda_n)^{-1/2}\sum_i(W_i - E[W_i]) \xrightarrow{d} N(0,1)$. By using $\sigma_n^2(\mu) = \lambda_n$, this result is equivalent to $\mu' Q_n^{-1/2}\sum_i(W_i - E[W_i]) \xrightarrow{d} N(0,1)$ as required.

Turning to consistent variance estimation, I first 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 $\hat{Q}_n - Q_n = \sum_i \sum_{j\in\mathcal{N}_i} W_i W_j' - E[W_i W_j']$, 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 $m, p \in \{1, 2, \cdots, K\}$. Then,

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

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

$$
\mu'(Q_n^{-1/2}(\hat{Q}_n - Q_n)Q_n^{-1/2})\mu = \mu_0'\frac{1}{\lambda_n}(\hat{Q}_n - Q_n)\mu_0 \xrightarrow{p} 0.
$$

where μ_0 is a vector whose entries are all bounded above by some arbitrary constant $K_1 < \infty$ by a similar argument as before. Convergence occurs because $(1/\lambda_n)(\hat{Q}_n - Q_n) \xrightarrow{p} 0_{K\times K}$.

□

B Proof of Propositions

Proof of Proposition 1. For (a), take any observation i and its associated clusters $g(i), h(i)$. Use the permutation function $\pi_1(g(i)) = 1$ and $\pi_2(h(i)) = 1$ so the array has the same distribution as before due to separate exchangeability. Since the array is dissociated, by setting $G_0 = H_0 = 1$, W_i

is independent of all observations that are not in $g(i)$ or $h(i)$, verifying (a).

For (b), take any i and $k \in \mathcal{N}_i$. Without loss of generality, suppose that $g(i) = g(k)$. Consider the case where $h(i) \neq h(k)$. Use the permutation function $\pi_1(g(i)) = 1$ and $\pi_2(h(i)) = 1, \pi_2(h(k)) = 2$ to get another array that has the same distribution. Since the array is dissociated, by setting $G_0 = 1, H_0 = 2$, (W_i, W_k) is independent of all observations that are not in $(\mathcal{N}_i \cup \mathcal{N}_k)$. Since $j, l \notin (\mathcal{N}_i \cup \mathcal{N}_k)$, $(W_i, W_k) \perp\!\!\!\perp (W_j, W_l)$, which yields (b). If $h(k) = h(i)$, set $\pi_2(h(k)) = 1$ and $G_0 = 1, H_0 = 1$. The same argument applies. □

For Proposition 2, I first prove a consistency result.

Lemma 8. *Under Assumptions 1, 2(a) and 2(b), and* $E[W_i] = 0 \; \forall i$, $||(1/n \sum_i (W_i W_i' - E[W_i W_i'])|| \xrightarrow{p} 0$.

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 Assumption 2(a) that $\max_{m,k} E[W_{mk}^4] < K_0$,

$$P\left(\frac{1}{n}\sum_i (X_i Y_i - E[X_i Y_i]) > \epsilon\right)$$
$$\leq \frac{1}{\epsilon^2}\frac{1}{n^2} E\left(\sum_i \sum_{j \in \mathcal{N}_i} (X_i Y_i - E[X_i Y_i])(X_j Y_j - E[X_j Y_j])\right) \leq \frac{K_0}{\epsilon^2 n^2} \sum_i \sum_{j \in \mathcal{N}_i} 1.$$

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

$$\frac{\sum_i \sum_{j \in \mathcal{N}_i} 1}{n^2} \leq \frac{\max_i N_i}{n} \frac{(\sum_i 1)}{n},$$

so it suffices to show $\max_i N_i / n = o(1)$. Since

$$\lambda_n \leq \sum_i \sum_{j \in \mathcal{N}_i} \max_m E[W_{mk}^2] \leq n^2 \max_m E[W_{mk}^2],$$

we have:

$$\frac{(\max_i N_i)^2}{n^2} = \frac{(\max_i N_i)^2 \max_m E[W_{mk}^2]}{n^2 \max_m E[W_{mk}^2]} \leq \max_m E[W_{mk}^2] \frac{(\max_i N_i)^2}{\lambda_n} = o(1).$$

Convergence occurs due to Assumption 2(b) and $\max_m E[W_{mk}^2] < K_0$. □

Proof of Proposition 2. Since $E[u_i^4|X_i] \leq K_0$, $E[u_i^4 X_{ik}^4] = E[E[u_i^4|X_i]X_{ik}^4] \leq K_0 E[X_{ik}^4] \leq K_0^2$ is bounded. By Theorem 1, $Q_n^{-1/2}\sum_i X_i u_i \xrightarrow{d} N(0, I_K)$.

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 8, $(1/n)(\hat{S}_n - S_n) = (1/n)\sum_i (X_i X_i' - E[X_i X_i']) = o_P(1)$. 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 3(e). Since $\hat{\beta} - \beta = \hat{S}_n^{-1}\sum_i X_i u_i$, 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, $\forall \mu \in \mathbb{R}^K$, $\mu'\left(Q_n^{-1/2}(\hat{Q}_n - Q_n)Q_n^{-1/2}\right)\mu = o_P(1)$. Expanding $\hat{Q}_n$,

$$\begin{aligned}\hat{Q}_n &:= \sum_i \sum_{j\in\mathcal{N}_i} \hat{u}_i \hat{u}_j X_i X_j' = \sum_i \sum_{j\in\mathcal{N}_i} (u_i - X_i'(\hat{\beta} - \beta))(u_j - X_j'(\hat{\beta} - \beta)) X_i X_j' \\ &= \sum_i \sum_{j\in\mathcal{N}_i} u_i u_j X_i X_j' - 2\left(\sum_i \sum_{j\in\mathcal{N}_i} u_i X_j'(\hat{\beta} - \beta) X_i X_j'\right) + \left(\sum_i \sum_{j\in\mathcal{N}_i} X_i'(\hat{\beta} - \beta) X_j'(\hat{\beta} - \beta) X_i X_j'\right).\end{aligned}$$

By Theorem 1, $\mu' Q_n^{-1/2}(\sum_i \sum_{j\in\mathcal{N}_i} u_i u_j X_i X_j' - Q_n) Q_n^{-1/2}\mu = o_P(1)$. Hence, it remains to show:

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

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 and $Z_K \sim N(0, I_K)$. 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 equality implies:

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

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 3, $\lambda_nK_0\geq\sum_c(N_c^C)^2$, which implies $\lambda_n\geq(\sum_c(N_c^C)^2)/K_0$. Hence,

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

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, and the same argument as the proof of Theorem 1 for $\mu\in\mathbb{R}^K$, $||\mu||=1$,

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

where $K_0\in\mathbb{R}$ is an arbitrary (finite) constant. Convergence occurs due to Assumption 3(b), which implies $\max_i N_i/n\to 0$, since $\max_i\sum_{j\in\mathcal{N}_i}N_i/n=o(1)$ in the proof of Lemma 8.

Going back to the first term,

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

By using Markov and Minkowski inequalities,

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

The penultimate inequality occurs due to Hölder'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 3(b). Convergence in the last step occurs because $\max_i N_i/\sqrt{\lambda_n}=o(1)$, and the moments are finite.

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. □

References

Braun, M. and V. Verdier (2023): "Estimation of spillover effects with matched data or longitudinal network data," *Journal of Econometrics*, 233, 689–714.

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

Chandrasekhar, A. G. and M. O. Jackson (2016): "A network formation model based on subgraphs," *arXiv preprint arXiv:1611.07658*.

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

Chiang, H. D., B. E. Hansen, and Y. Sasaki (2024): "Standard errors for two-way clustering with serially correlated time effects," *Review of Economics and Statistics*, 1–40.

Chiang, H. D. and Y. Sasaki (2023): "On Using The Two-Way Cluster-Robust Standard Errors," *arXiv preprint arXiv:2301.13775*.

Chin, A. (2018): "Central limit theorems via Stein's method for randomized experiments under interference," *arXiv preprint arXiv:1804.03105*.

Crane, H. and H. Towsner (2018): "Relatively exchangeable structures," *The Journal of Symbolic Logic*, 83, 416–442.

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

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.

Graham, B. S. (2020): "Sparse network asymptotics for logistic regression," *arXiv preprint arXiv:2010.04703*.

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

Jackson, C. K. (2018): "What do test scores miss? The importance of teacher effects on non–test score outcomes," *Journal of Political Economy*, 126, 2072–2107.

Janisch, M. and T. Lehéricy (2024): "Berry–Esseen-Type Estimates for Random Variables with a Sparse Dependency Graph," *Journal of Theoretical Probability*, 37, 3627–3653.

Janson, S. (1988): "Normal convergence by higher semiinvariants with applications to sums of dependent random variables and random graphs," *The Annals of Probability*, 305–312.

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

Leung, M. P. (2022): "Rate-optimal cluster-randomized designs for spatial interference," *The Annals of Statistics*, 50, 3064–3087.

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.

Neumark, D., I. Burn, and P. Button (2019): "Is it harder for older workers to find jobs? New and improved evidence from a field experiment," *Journal of Political Economy*, 127, 922–970.

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.

Verdier, V. (2020): "Estimation and inference for linear models with two-way fixed effects and sparsely matched data," *Review of Economics and Statistics*, 102, 1–16.

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

Xu, R. and L. Yap (2024): "Clustering with Potential Multidimensionality: Inference and Practice," *arXiv preprint arXiv:2411.13372*.