MacKinnon, James G.; Nielsen, Morten Ørregaard; Webb, Matthew

Working Paper

Wild bootstrap and asymptotic inference with multiway clustering

Queen's Economics Department Working Paper, No. 1415

Provided in Cooperation with:
Queen's University, Department of Economics (QED)

Suggested Citation: MacKinnon, James G.; Nielsen, Morten Ørregaard; Webb, Matthew (2019) : Wild bootstrap and asymptotic inference with multiway clustering, Queen's Economics Department Working Paper, No. 1415, Queen's University, Department of Economics, Kingston (Ontario)

This Version is available at:
http://hdl.handle.net/10419/230568

Queen's Economics Department Working Paper No. 1415

Wild Bootstrap and Asymptotic Inference with Multiway Clustering

James G. MacKinnon
Queen's University

Morten Ørregaard Nielsen
Queen's University
and CREATES

Matthew D. Webb
Carleton University

Department of Economics
Queen's University
94 University Avenue
Kingston, Ontario, Canada
K7L 3N6

3-2019

This is a substantially revised version of QED Working Paper No. 1386, which had a slightly different title.

Wild Bootstrap and Asymptotic Inference with Multiway Clustering

James G. MacKinnon*
Queen's University
jgm@econ.queensu.ca

Morten Ørregaard Nielsen
Queen's University and CREATES
mon@econ.queensu.ca

Matthew D. Webb
Carleton University
matt.webb@carleton.ca

September 13, 2019

Abstract

We study two cluster-robust variance estimators (CRVEs) for regression models with clustering in two dimensions and give conditions under which t-statistics based on each of them yield asymptotically valid inferences. In particular, one of the CRVEs requires stronger assumptions about the nature of the intra-cluster correlations. We then propose several wild bootstrap procedures and state conditions under which they are asymptotically valid for each type of t-statistic. Extensive simulations suggest that using certain bootstrap procedures with one of the t-statistics generally performs very well. An empirical example confirms that bootstrap inferences can differ substantially from conventional ones.

Keywords: CRVE, grouped data, clustered data, cluster-robust variance estimator, two-way clustering, robust inference, wild cluster bootstrap.

JEL Codes: C15, C21, C23.

*Corresponding author. Address: Department of Economics, 94 University Avenue, Queen's University, Kingston, Ontario K7L 3N6, Canada. Email: jgm@econ.queensu.ca. Tel. 613-533-2293. Fax 613-533-6668.

1 Introduction

The disturbances (error terms) in regression models often appear to be correlated within clusters. It is typically assumed that there is clustering in just one dimension, such as by jurisdiction or by classroom. In such cases, it is now standard to use a cluster-robust variance estimator, or CRVE, perhaps combined with the wild cluster bootstrap. There is a large and rapidly growing literature on this topic; see the excellent survey of Cameron and Miller (2015). More recent papers include Imbens and Kolesár (2016), Ibragimov and Müller (2016), MacKinnon and Webb (2017, 2018), Carter, Schnepel, and Steigerwald (2017), Pustejovsky and Tipton (2018), and Djogbenou, MacKinnon, and Nielsen (2019).

Methods for one-way clustering are sufficient in many cases, but clustering may also plausibly occur in two or more dimensions. For example, for panel data, there may be correlation both within jurisdictions across time periods and within time periods across jurisdictions. Cameron, Gelbach, and Miller (2011) (CGM hereafter) proposes a method to calculate standard errors that are robust to multiway clustering; see also Thompson (2011), which proposes essentially the same method for the two-way case, and Davezies, D'Haultfœuille, and Guyonvarch (2018), which proposes a variant that is asymptotically equivalent under certain conditions. CGM's "multiway CRVE" is widely used in empirical work, but CGM does not state the conditions under which it is asymptotically valid or provide a formal proof. Moreover, simulations in CGM suggest that using a two-way CRVE does not always work well, especially when the number of clusters in either dimension is small.

In Section 2, we discuss the linear regression model with disturbances that are clustered in two dimensions and the two variants of the multiway CRVE. In Section 3, we prove that t-statistics based on each variant of the multiway CRVE yield asymptotically valid inferences for the case of two-dimensional clustering under precisely stated conditions. Variations of these CRVEs can handle clustering in more than two dimensions, and our proofs could be extended to handle such cases. However, we do not attempt to analyze higher-dimensional clustering, because the notation and analysis would be quite tedious. Moreover, to our knowledge, empirical work with multiway CRVEs very rarely goes beyond the two-dimensional case. Our proof of asymptotic validity builds upon the asymptotic distribution theory with multiway clustering in Davezies et al. (2018) and Menzel (2018).

The second methodological contribution of this paper, discussed in Section 4, is to propose several variants of the wild (cluster) bootstrap. These methods differ in how the bootstrap disturbances are (one-way) clustered, i.e. by the first dimension, second dimension, intersection, or not at all (the ordinary wild bootstrap), by using either of the two multiway CRVEs, and by using either restricted or unrestricted estimates. To our knowledge, this is the first application of wild bootstrap methods to clustering in multiple dimensions. Under various assumptions about the nature of the intra-cluster correlations, that vary across the CRVEs and the methods, we prove which variants are asymptotically valid and which ones are not.

Next, in Section 5, we present the results of an extensive set of simulation experiments, which suggest that wild bootstrap inference tends to be much more reliable than asymptotic inference. Finally, in Section 6, we illustrate our results with an empirical example from Nunn and Wantchekon (2011) where it is possible to cluster both by ethnicity and at different geographic levels. Section 7 concludes. All proofs are given in the appendix.

2 The Model

Consider a linear regression model with two-way clustering written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \tag{1}$$

where $\boldsymbol{y}$ and $\boldsymbol{u}$ are $N \times 1$ vectors of observations and disturbances, $\boldsymbol{X}$ is an $N \times k$ matrix of covariates, and $\boldsymbol{\beta}$ is a $k \times 1$ parameter vector. The model is assumed to have two dimensions of clustering, where the numbers of clusters in the two dimensions are G and H, respectively. We can rewrite (1) as

$$\boldsymbol{y}_{gh} = \boldsymbol{X}_{gh}\boldsymbol{\beta} + \boldsymbol{u}_{gh}, \quad g = 1, \dots, G, \quad h = 1, \dots, H, \tag{2}$$

where the vectors $\boldsymbol{y}_{gh}$ and $\boldsymbol{u}_{gh}$ and the matrix $\boldsymbol{X}_{gh}$ contain, respectively, the rows of $\boldsymbol{y}$, $\boldsymbol{u}$, and $\boldsymbol{X}$ that correspond to both the g^{th} cluster in the first clustering dimension and the h^{th} cluster in the second clustering dimension. The GH clusters into which the data are divided in (2) represent the intersection of the two clustering dimensions.

We need notation for the number of observations in each cluster for each dimension. It would be natural to use N_g^1 for the g^{th} cluster in the first dimension and N_h^2 for the h^{th} cluster in the second dimension. However, to avoid excessive complexity, we omit the superscripts. Thus we use N_g to denote the number of observations in cluster g for the first dimension and N_h to denote the number of observations in cluster h for the second dimension, as well as N_{gh} to denote the number of observations in the intersection of cluster g in the first dimension and cluster h in the second dimension. In the theoretical context, there should be no ambiguity.

Similarly, we use $\boldsymbol{y}_g$, $\boldsymbol{X}_g$, and $\boldsymbol{u}_g$ to denote vectors that contain the rows of $\boldsymbol{y}$, $\boldsymbol{X}$, and $\boldsymbol{u}$ for the g^{th} cluster in the first dimension, and $\boldsymbol{y}_h$, $\boldsymbol{X}_h$, and $\boldsymbol{u}_h$ to denote the corresponding rows for the h^{th} cluster in the second dimension. Note that, in terms of the notation of (2), the vector $\boldsymbol{y}_g$ contains the subvectors $\boldsymbol{y}_{g1}$ through $\boldsymbol{y}_{gH}$.

Since there are N_g observations in a typical cluster for the first dimension, N_h observations in a typical cluster for the second dimension, and N_{gh} observations in a typical cluster for the intersection, the number of observations in the entire sample is

$$N = \sum_{g=1}^{G} N_g = \sum_{h=1}^{H} N_h = \sum_{g=1}^{G}\sum_{h=1}^{H} N_{gh}.$$

We assume that $N_g \geq 1$ and $N_h \geq 1$, but N_{gh} might well equal 0 for some values of g and h.

Under two-way clustering, the variance matrix of the scores,

$$\boldsymbol{\Sigma} = \mathrm{E}(\boldsymbol{X}^\top \boldsymbol{u}\boldsymbol{u}^\top \boldsymbol{X}) = \sum_{g,g'=1}^{G}\sum_{h,h'=1}^{H} \mathrm{E}\Big(\boldsymbol{X}_{g'h'}^\top \boldsymbol{u}_{g'h'}\boldsymbol{u}_{gh}^\top \boldsymbol{X}_{gh}\Big),$$

has the particular structure

$$\mathrm{E}(\boldsymbol{X}_{g'h'}^\top \boldsymbol{u}_{g'h'}\boldsymbol{u}_{gh}^\top \boldsymbol{X}_{gh}) = \boldsymbol{0} \quad \text{if} \quad g' \neq g \text{ and } h' \neq h \tag{3}$$

and arbitrary covariances if either $g = g'$ or $h = h'$. The variance matrices for the subvectors $\boldsymbol{X}_g^\top \boldsymbol{u}_g$, $\boldsymbol{X}_h^\top \boldsymbol{u}_h$, and $\boldsymbol{X}_{gh}^\top \boldsymbol{u}_{gh}$ are respectively denoted

$$\boldsymbol{\Sigma}_g = \mathrm{E}(\boldsymbol{X}_g^\top \boldsymbol{u}_g \boldsymbol{u}_g^\top \boldsymbol{X}_g), \ \boldsymbol{\Sigma}_h = \mathrm{E}(\boldsymbol{X}_h^\top \boldsymbol{u}_h \boldsymbol{u}_h^\top \boldsymbol{X}_h), \text{ and } \boldsymbol{\Sigma}_{gh} = \mathrm{E}(\boldsymbol{X}_{gh}^\top \boldsymbol{u}_{gh} \boldsymbol{u}_{gh}^\top \boldsymbol{X}_{gh}). \tag{4}$$

With the structure in (3) and notation in (4), we write, by the inclusion-exclusion principle,

$$\boldsymbol{\Sigma} = \sum_{g=1}^{G} \boldsymbol{\Sigma}_g + \sum_{h=1}^{H} \boldsymbol{\Sigma}_h - \sum_{g=1}^{G} \sum_{h=1}^{H} \boldsymbol{\Sigma}_{gh}. \tag{5}$$

As usual, the OLS estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

In large samples, the variance matrix of $\hat{\boldsymbol{\beta}}$ is given by (the limit of) $\boldsymbol{Q}^{-1}\boldsymbol{\Gamma}\boldsymbol{Q}^{-1}$, where $\boldsymbol{Q}$ is defined as $(GH)^{-1}\sum_{g=1}^{G}\sum_{h=1}^{H}\boldsymbol{X}_{gh}^\top\boldsymbol{X}_{gh} = (GH)^{-1}\boldsymbol{X}^\top\boldsymbol{X}$, and $\boldsymbol{\Gamma} = (GH)^{-2}\boldsymbol{\Sigma}$. Based on (5), the cluster-robust variance estimator, i.e. the multiway CRVE, suggested by CGM is

$$\hat{\boldsymbol{V}}_3 = \hat{\boldsymbol{V}}_G + \hat{\boldsymbol{V}}_H - \hat{\boldsymbol{V}}_I, \quad \hat{\boldsymbol{V}}_m = \boldsymbol{Q}^{-1}\hat{\boldsymbol{\Gamma}}_m\boldsymbol{Q}^{-1}, \quad m \in \{G, H, GH\}, \tag{6}$$

where

$$\begin{aligned} \hat{\boldsymbol{\Gamma}}_G &= \frac{1}{(GH)^2}\sum_{g=1}^{G}\boldsymbol{X}_g^\top\hat{\boldsymbol{u}}_g\hat{\boldsymbol{u}}_g^\top\boldsymbol{X}_g, \quad \hat{\boldsymbol{\Gamma}}_H = \frac{1}{(GH)^2}\sum_{h=1}^{H}\boldsymbol{X}_h^\top\hat{\boldsymbol{u}}_h\hat{\boldsymbol{u}}_h^\top\boldsymbol{X}_h, \\ \hat{\boldsymbol{\Gamma}}_I &= \frac{1}{(GH)^2}\sum_{g=1}^{G}\sum_{h=1}^{H}\boldsymbol{X}_{gh}^\top\hat{\boldsymbol{u}}_{gh}\hat{\boldsymbol{u}}_{gh}^\top\boldsymbol{X}_{gh}, \end{aligned} \tag{7}$$

and $\hat{\boldsymbol{u}}_g$, $\hat{\boldsymbol{u}}_h$, and $\hat{\boldsymbol{u}}_{gh}$ denote various subvectors of the vector of OLS residuals. Note that $\hat{\boldsymbol{V}}_m$ is the one-way CRVE with clustering in dimension m. The subscript "3" on $\hat{\boldsymbol{V}}_3$ in (6) emphasizes that this estimator has three terms. A two-term estimator will be discussed below.

In practice, the factor of $(GH)^{-2}$ in (7) is almost always omitted, and $\boldsymbol{Q}$ is replaced by $\boldsymbol{X}^\top\boldsymbol{X}$. This leaves the value of $\hat{\boldsymbol{V}}_3$ unchanged. Moreover, the three matrices in (7) are usually, for example in Stata, multiplied by

$$\frac{G(N-1)}{(G-1)(N-k)}, \quad \frac{H(N-1)}{(H-1)(N-k)}, \quad \text{and} \quad \frac{GH(N-1)}{(GH-1)(N-k)}, \tag{8}$$

respectively, by analogy with the scalar factor that is conventionally employed with the one-way CRVE. We make use of the factors in (8) in our simulations, but for purposes of asymptotic theory we omit them without loss of generality.

One important practical issue is that the matrix $\hat{\boldsymbol{\Gamma}}_3$ defined in (7) is not necessarily positive definite in finite samples, which implies that the diagonal elements of $\hat{\boldsymbol{V}}_3$ may not all be positive. In fact, since the ranks of the three matrices in (7) cannot exceed G, H, and GH, respectively, it seems likely that $\hat{\boldsymbol{V}}_3$ will not be positive definite whenever the model contains significantly more than $\min\{G, H\}$ regressors. Indeed, even when the number of regressors is small relative to G and H, there may very well be samples for which $\hat{\boldsymbol{V}}_3$ is not positive definite; see Sections 5 and 6.

To deal with this problem, CGM suggests calculating the eigenvalues of $\hat{\boldsymbol{V}}_3$, say $\Lambda_1, \ldots, \Lambda_k$. When any of them is not positive, $\hat{\boldsymbol{V}}_3$ is replaced by the eigendecomposition $\hat{\boldsymbol{V}}_3^+ = \boldsymbol{U}\boldsymbol{\Lambda}^+\boldsymbol{U}^\top$, where $\boldsymbol{U}$ is the $k \times k$ matrix of eigenvectors and $\boldsymbol{\Lambda}^+$ is a diagonal matrix with typical diagonal element $\max\{\Lambda_j, 0\}$. The matrix $\hat{\boldsymbol{V}}_3^+$ is guaranteed to be positive semidefinite, so it

could have diagonal elements that equal 0. It is impossible to use $\hat{\boldsymbol{V}}_3^+$ for inference about the coefficients that correspond to those diagonal elements.

Even for the one-way CRVE, it is common to encounter singular variance matrices when there are fixed effects. This problem is most easily dealt with by projecting the regressand and regressors off the fixed effects before running the regression; see Pustejovsky and Tipton (2018) and Djogbenou et al. (2019). That trick could also be used with the two-way CRVE.

An alternative to (6) is proposed in Davezies, D'Haultfœuille, and Guyonvarch (2018) based on the argument that, under the conditions in that paper, the third matrix in (7) is of smaller order of magnitude than the first two matrices. This leads to the two-term CRVE,

$$\hat{\boldsymbol{V}}_2 = \hat{\boldsymbol{V}}_G + \hat{\boldsymbol{V}}_H. \tag{9}$$

Note that $\hat{\boldsymbol{V}}_2$ is actually denoted $\hat{\boldsymbol{V}}_1$ in Davezies et al. (2018). That paper proposes another three-term estimator (there denoted $\hat{\boldsymbol{V}}_2$), which subtracts twice the third term in (7). However, that estimator exacerbates the problem of lack of positive definiteness. It is therefore not recommended in Davezies et al. (2018), and hence we do not consider it further.

The two-term CRVE, $\hat{\boldsymbol{V}}_2$, has the computational advantage that it is guaranteed to be positive semidefinite, and it therefore has merit. However, as we will see in the next section, omitting the third term in (7) is valid only under certain conditions. Moreover, $\hat{\boldsymbol{V}}_2$ is not robust to the possibility that the data-generating process (DGP) does not in fact have clustering in the two dimensions specified (for example, it could have independent observations), whereas $\hat{\boldsymbol{V}}_3$ is robust to such situations. Additionally, in every case that we study in Section 5, bootstrap methods perform better with $\hat{\boldsymbol{V}}_3$ than with $\hat{\boldsymbol{V}}_2$. Nevertheless, in practice it may be useful to employ $\hat{\boldsymbol{V}}_2$ in situations where $\hat{\boldsymbol{V}}_3$ is not positive definite.

3 Asymptotic Theory

In this section, we derive the asymptotic limit theory for t-statistics based on the CRVEs $\hat{\boldsymbol{V}}_2$ and $\hat{\boldsymbol{V}}_3$. We let $\boldsymbol{\beta}_0$ denote the true value of $\boldsymbol{\beta}$ and consider the cluster-robust t-statistic,

$$t_{a,j} = \frac{\boldsymbol{a}^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)}{\sqrt{\boldsymbol{a}^\top \hat{\boldsymbol{V}}_j \boldsymbol{a}}}, \quad j \in \{2, 3\}, \tag{10}$$

for testing the null hypothesis H_0: $\boldsymbol{a}^\top\boldsymbol{\beta} = \boldsymbol{a}^\top\boldsymbol{\beta}_0$ against either a one-sided or two-sided alternative hypothesis. Here $\boldsymbol{a}$ is a known vector, which if we were testing a hypothesis about one element of $\boldsymbol{\beta}$ would be a unit vector. We impose the normalization that $\boldsymbol{a}^\top\boldsymbol{a} = 1$.

The asymptotic theory for the cluster-robust t-statistic (10) has several precursors in the literature on one-way clustering, although these are obtained under assumptions that are very different from ours. In particular, White (1984, Ch. 6) assumes equal-sized, homogeneous (same variance) clusters, and Hansen (2007) assumes equal-sized, heterogeneous clusters. In contrast, our conditions below allow clusters to be heterogeneous in both size and variance.

More recently, Carter et al. (2017), Djogbenou et al. (2019), and Hansen and Lee (2019) obtain results for one-way clustering that allow cluster heterogeneity. The first of these papers invokes a primitive moment condition and makes some high-level assumptions about cluster-size heterogeneity and interactions between regressors and disturbances. The latter

two papers make weaker assumptions that allow for arbitrary dependence and correlation within clusters. See Djogbenou et al. (2019) for a detailed comparison of these assumptions.

The asymptotic theory in the above papers is based on the insight that, under one-way clustering, $\boldsymbol{a}^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ can be written as a sum of independent random variables, say $\sum_{g=1}^{G} z_g$, so that a standard central limit theorem can be applied. However, that approach is not applicable to multiway clustering, because any partitioning of the data according to one dimension of clustering will be arbitrarily dependent due to the other dimension of clustering. That is, $\boldsymbol{a}^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ cannot be written as $\sum_{g=1}^{G} z_g$ for any independent sequence, z_g.

3.1 Exchangeability and Other Assumptions

A natural stochastic framework for the regression model with multiway clustered data is that of *separately exchangeable* random variables, as pioneered in the clustering context by Davezies, D'Haultfœuille, and Guyonvarch (2018) and Menzel (2018). The asymptotic theory for two-way clustering in this section makes use of key results in those papers.

In one dimension, two random variables, Z_1 and Z_2, are exchangeable if the distribution of (Z_1, Z_2) is the same as that of (Z_2, Z_1), denoted $(Z_1, Z_2) \stackrel{d}{=} (Z_2, Z_1)$. Note that this does not rule out the possibility that the variables are correlated. For example, if (Z_1, Z_2) has a multivariate normal distribution with identical marginal distributions (common means and variances), then Z_1 and Z_2 are exchangeable for any value of the covariance (and hence correlation). In contrast, the sequence of random variables $Z_t, t \geq 1$, generated by the stationary autoregression $Z_t = \alpha Z_{t-1} + \epsilon_t$, with ϵ_t i.i.d. and $|\alpha| < 1$, is not exchangeable. More discussion and examples follow after stating the formal assumptions.

We next formalize the conditions for the asymptotic theory. After establishing the conditions and notation, we give some additional comments and discussion. The conditions are formulated at the intersection level, where, in particular, the infinite sequence of random variables that generate the observations within intersection (g, h) is denoted

$$\mathcal{S}_{gh} = (\mathcal{S}_{gh,i})_{i\geq 1} = (\boldsymbol{X}_{gh,i}^\top, u_{gh,i})_{i\geq 1}^\top, \quad g \geq 1, h \geq 1,$$

where $\boldsymbol{X}_{gh,i}$ denotes the i^{th} row of $\boldsymbol{X}_{gh}$ and $u_{gh,i}$ denotes the i^{th} element of $\boldsymbol{u}_{gh}$. The observed data for intersection (g, h) are the realizations of the first N_{gh} random variables from $\mathcal{S}_{gh}$, and the observed clusters are the realizations of the sequence $\mathcal{S}_{gh}$ for $g = 1, \ldots, G$ and $h = 1, \ldots, H$. Following Davezies et al. (2018), the number of observations within each intersection is randomly determined as the realization of the random variable N_{gh} (without risk of confusion, we use the same notation N_{gh} for the random variable and its realization). For any matrix $\boldsymbol{M}$, let $\|\boldsymbol{M}\| = (\text{Tr}(\boldsymbol{M}^\top \boldsymbol{M}))^{1/2}$ denote the Euclidean (Frobenius) norm.

Assumption 1. $\text{E}(\boldsymbol{X}_{gh}^\top \boldsymbol{u}_{gh}) = \boldsymbol{0}$ and $\text{E}\Big(\Big(\sum_{i=1}^{N_{gh}} \|\mathcal{S}_{gh,i}\|^2\Big)^2\Big) < \infty$.

Assumption 2. The regressors are such that $\boldsymbol{Q}_0 = \text{E}(\boldsymbol{X}_{gh}^\top \boldsymbol{X}_{gh})$ is non-singular.

Assumption 3. $\text{E}(N_{gh}) > 0$.

Assumption 4. $(N_{gh}, \mathcal{S}_{gh})$ is independent of $(N_{g'h'}, \mathcal{S}_{g'h'})$ if $g \neq g'$ and $h \neq h'$.

Assumption 5. $R = \min\{G, H\} \to \infty$.

Assumption 6. $(N_{gh}, \mathcal{S}_{gh}) \overset{d}{=} (N_{\pi_1(g)\pi_2(h)}, \mathcal{S}_{\pi_1(g)\pi_2(h)})$, where $\pi_1(\cdot)$ and $\pi_2(\cdot)$ are permutations of $\mathbb{N}$.

The conditions in the first five assumptions are very mild and nearly minimal. First, Assumption 1 is a simple moment condition, where the existence of at least four moments is assumed. Conditions like Assumption 2 are standard in linear regression models to rule out perfectly collinear regressors. Assumption 3 trivially rules out datasets that are empty almost surely, but it allows the possibility that some (or even many) intersections are empty. Assumption 4 is the assumption of two-way clustering; c.f. (3). Assumption 5 sets up the asymptotic framework under which the number of clusters tends to infinity along both dimensions of clustering. Note that there is no restriction on the relative number of clusters in the two dimensions; that is, there is no restriction on the ratio G/H other than the obvious one that it be non-negative. For example, G and H can grow at different rates.

Finally, the crucial condition of *separate exchangeability* is formalized in Assumption 6. In two dimensions, we can think of a matrix. The separate exchangeability condition in Assumption 6 then means that the distribution of an entry of the matrix is invariant to re-ordering of the rows (but keeping entire rows together) and invariant to re-ordering of the columns (but keeping entire columns together). In more technical terms, this is invariance under permutations of the rows and columns, *separately*. For example, consider a two-dimensional array of random variables, say Z_{ij} for $i \geq 1, j \geq 1$, generated by

$$Z_{ij} = c + a_i + b_j + e_{ij}. \tag{11}$$

When a_i, b_j, and e_{ij} are i.i.d. across i and j and mutually independent, (11) is the random effects model. In that case, the Z_{ij} are easily seen to be separately exchangeable. Note that it is not required that the distribution be invariant under permutations of both rows and columns simultaneously. The latter would be invariance under re-ordering of the entries of the matrix, which is a stronger condition.

Thus, to interpret the separate exchangeability condition, it is important to consider the distribution of the entire array, $(N_{gh}, \mathcal{S}_{gh})_{g\geq 1, h\geq 1}$. We illustrate the relevance of the above points in a simple example. Consider a repeated cross-section clustered by state (g) and year (h). In the matrix analogy, we have states in each row and years in each column. A relevant point in this example is that California is much larger than Rhode Island. Hence, N_{gh} is typically much larger for California than it is for Rhode Island, for all years $h = 1, \ldots, H$. Under the stronger type of exchangeability, this would not be possible. However, because separate exchangeability keeps entire rows together, this type of data can occur when N_{gh} is correlated across h for a given g.

We emphasize three additional important implications of Assumption 6. Although $(N_{gh}, \mathcal{S}_{gh})$ has the same marginal distribution for all $g \geq 1$ and $h \geq 1$, this is not as restrictive as it may appear. First, there is no restriction on the correlation of both N_{gh} and $\mathcal{S}_{gh}$ across g and/or h, where correlation, for example, can generate the behavior illustrated in the previous paragraph. Second, individual elements within clusters, i.e. $\boldsymbol{X}_{gh,i}$ and $u_{gh,i}$, can have different distributions for different i. Third, there is also no restriction on the correlation between N_{gh} and $\mathcal{S}_{gh}$, which implies that, conditional on N_{gh}, the covariance structure of $\mathcal{S}_{gh}$ may differ across g and/or h. These three features thus allow for many types of cluster heterogeneity.

3.2 Asymptotic Distribution

Analogously to each term in (5) and (7), we first define the asymptotic variance matrices

$$\boldsymbol{\Gamma}_G = \lim \frac{1}{GH^2} \sum_{g=1}^{G} \mathrm{E}(\boldsymbol{X}_g^\top \boldsymbol{u}_g \boldsymbol{u}_g^\top \boldsymbol{X}_g) = \lim \frac{1}{GH^2} \sum_{g=1}^{G} \sum_{h,h'=1}^{H} \mathrm{E}(\boldsymbol{X}_{gh}^\top \boldsymbol{u}_{gh} \boldsymbol{u}_{gh'}^\top \boldsymbol{X}_{gh'}), \tag{12}$$

$$\boldsymbol{\Gamma}_H = \lim \frac{1}{G^2H} \sum_{h=1}^{H} \mathrm{E}(\boldsymbol{X}_h^\top \boldsymbol{u}_h \boldsymbol{u}_h^\top \boldsymbol{X}_h) = \lim \frac{1}{G^2H} \sum_{g,g'=1}^{G} \sum_{h=1}^{H} \mathrm{E}(\boldsymbol{X}_{gh}^\top \boldsymbol{u}_{gh} \boldsymbol{u}_{g'h}^\top \boldsymbol{X}_{g'h}), \tag{13}$$

$$\boldsymbol{\Gamma}_I = \lim \frac{1}{GH} \sum_{g=1}^{G} \sum_{h=1}^{H} \mathrm{E}(\boldsymbol{X}_{gh}^\top \boldsymbol{u}_{gh} \boldsymbol{u}_{gh}^\top \boldsymbol{X}_{gh}), \text{ and} \tag{14}$$

$$\boldsymbol{\Gamma}_{NC} = \lim \frac{1}{GH} \sum_{g=1}^{G} \sum_{h=1}^{H} \sum_{i=1}^{N_{gh}} \mathrm{E}(\boldsymbol{X}_{gh,i}^\top \boldsymbol{u}_{gh,i} \boldsymbol{u}_{gh,i}^\top \boldsymbol{X}_{gh,i}), \tag{15}$$

which are all finite and positive semidefinite under our assumptions. We also define

$$\lambda_m = \lim(m^{-1}R) \in [0, \infty) \text{ for } m \in \{G, H, I\},$$
$$\boldsymbol{V}_m = \boldsymbol{Q}^{-1} \boldsymbol{\Gamma}_m \boldsymbol{Q}^{-1} \text{ for } m \in \{G, H, I, NC\}.$$

The matrices $\boldsymbol{V}_G$, $\boldsymbol{V}_H$, and $\boldsymbol{V}_I$, appropriately normalized, are the asymptotic variance matrices of $\hat{\boldsymbol{\beta}}$ under one-way clustering in the first dimension, in the second dimension, and at the intersection level, respectively, while $\boldsymbol{V}_{NC}$ is the asymptotic variance of $\hat{\boldsymbol{\beta}}$ without clustering.

In the first result, we establish the asymptotic normality of $\hat{\boldsymbol{\beta}}$. This is given in Propositions 4.3 and 4.4 of Davezies et al. (2018) under the additional condition that

$$\lambda_G \boldsymbol{V}_G + \lambda_H \boldsymbol{V}_H \ > 0, \tag{16}$$

where "> 0" means positive definite. The condition (16) is an assumption that the DGP does in fact have clustering in at least one of the two dimensions. If that is not the case, the result will be different. Thus, we also give a result under the special case where there is no clustering in the G and H dimensions, but there is clustering at the intersection level, i.e.,

$$(N_{gh}, \mathcal{S}_{gh}) \text{ is independent of } (N_{g'h'}, \mathcal{S}_{g'h'}) \text{ if } g \neq g' \text{ or } h \neq h'. \tag{17}$$

In this case, the matrices $\boldsymbol{\Gamma}_G, \boldsymbol{\Gamma}_H, \boldsymbol{\Gamma}_I$ are identical after appropriate re-normalization, and

$$H\boldsymbol{V}_G = G\boldsymbol{V}_H = \boldsymbol{V}_I \ > 0.$$

Of course, the two cases (16) and (17) are not exhaustive, and there exist some degenerate cases in between in which $\hat{\boldsymbol{\beta}}$ may not even be asymptotically normally distributed; see Menzel (2018, Example 1.7). Finally, there is also the possibility that there is no clustering at all in the DGP, so that all observations are independent:

$$\mathcal{S}_{gh,i} \text{ is independent across } g, h, i. \tag{18}$$

Although asymptotic results for $\hat{\boldsymbol{\beta}}$ in this special case are well known, we consider (18) because it illustrates interesting consequences of using a multiway CRVE when not needed.

Theorem 1. *Suppose Assumptions 1–6 are satisfied and the true value of* $\boldsymbol{\beta}$ *is given by* $\boldsymbol{\beta}_0$.

(i) If (16) is true, then it holds that

$$R^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathrm{N}(0, \lambda_G \boldsymbol{V}_G + \lambda_H \boldsymbol{V}_H). \tag{19}$$

(ii) If (17) is true, then it holds that

$$(GH)^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathrm{N}(0, \boldsymbol{V}_I). \tag{20}$$

(iii) If (18) is true, then it holds that

$$(GH)^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathrm{N}(0, \boldsymbol{V}_{NC}). \tag{21}$$

An important consequence of the results in Theorem 1 is that the relevant notion of sample size in models that have a cluster structure is not in general the number of observations, N. This is seen clearly in the rate of convergence of $\hat{\boldsymbol{\beta}}$ in (19), which is $R^{-1/2}$ instead of the usual $N^{-1/2}$. However, because of assumption (16), the rate of convergence in (19) is based on the premise that clustering is present in the DGP in at least one of the two dimensions. When the DGP in fact only has clustering at the intersection level, the rate of convergence is instead given by $(GH)^{-1/2}$, as may be seen in (20). Indeed, this is why Carter et al. (2017), Djogbenou et al. (2019), and Hansen and Lee (2019), among others, consider asymptotic limit theory only for studentized (or self-normalized) quantities.

3.3 Variance Estimation and t-tests

In the next result, we present the limit theory for both the two-term and three-term CRVEs for each of the three cases considered in Theorem 1.

Theorem 2. *Suppose Assumptions 1–6 are satisfied.*

(i) If (16) is true, then it holds that

$$R\hat{\boldsymbol{V}}_2 \xrightarrow{P} \lambda_G \boldsymbol{V}_G + \lambda_H \boldsymbol{V}_H \quad \textit{and} \quad R\hat{\boldsymbol{V}}_3 \xrightarrow{P} \lambda_G \boldsymbol{V}_G + \lambda_H \boldsymbol{V}_H.$$

(ii) If (17) is true, then it holds that

$$GH\,\hat{\boldsymbol{V}}_2 \xrightarrow{P} 2\boldsymbol{V}_I \quad \textit{and} \quad GH\,\hat{\boldsymbol{V}}_3 \xrightarrow{P} \boldsymbol{V}_I.$$

(iii) If (18) is true, then it holds that

$$GH\,\hat{\boldsymbol{V}}_2 \xrightarrow{P} 2\boldsymbol{V}_{NC} \quad \textit{and} \quad GH\,\hat{\boldsymbol{V}}_3 \xrightarrow{P} \boldsymbol{V}_{NC}.$$

Comparing the results in Theorems 1 and 2, it is immediate that only the three-term CRVE, $\hat{\boldsymbol{V}}_3$, is consistent in all three of the cases considered. In contrast, $\hat{\boldsymbol{V}}_2$ is consistent only in the first case. In the other cases, it is not even valid asymptotically. Instead, it would yield standard errors that are too large, asymptotically, by a factor of $\sqrt{2}$.

The asymptotic distributions of the t-statistics in (10), i.e. $t_{a,2}$ and $t_{a,3}$, follow immediately from Theorems 1 and 2, and we state these as a corollary.

Corollary 1. *Suppose Assumptions 1–6 are satisfied and the null hypothesis, H_0, is true.*

(i) If (16) is true, then it holds that

$$t_{a,2} \xrightarrow{d} \mathrm{N}(0,1) \quad \textit{and} \quad t_{a,3} \xrightarrow{d} \mathrm{N}(0,1).$$

(ii) If either (17) or (18) is true, then it holds that

$$t_{a,2} \xrightarrow{d} \mathrm{N}(0,1/2) \quad \textit{and} \quad t_{a,3} \xrightarrow{d} \mathrm{N}(0,1).$$

The result in Corollary 1 shows the fundamental problem with the two-term CRVE, namely, that it is not robust to cases where the assumed cluster structure is in fact not present in the DGP. In such situations, shown in case (ii) of Corollary 1, a t-test based on $t_{a,2}$ and critical values from the $\mathrm{N}(0,1)$ distribution is not valid, even asymptotically. On the other hand, a t-test based on $t_{a,3}$ is valid in all cases considered. This may or may not outweigh the numerical advantages of the two-term CRVE discussed earlier, depending on how close the (unknown) DGP is to the special cases in (17) and (18).

More generally, the results in Corollary 1 justify the use of critical values and P values from a normal approximation to perform t-tests and construct confidence intervals based on the three-term CRVE, $\hat{\boldsymbol{V}}_3$. However, results in Bester, Conley, and Hansen (2011) suggest that it will often be more accurate to use the $t(G-1)$ distribution in the one-way case; see Cameron and Miller (2015) for a discussion of this issue. In the two-way case, CGM suggests using the $t(R-1)$ distribution (recall from Assumption 5 that $R = \min\{G,H\}$), and we do this in Sections 5 and 6 below.

4 Asymptotic Validity of the Wild (Cluster) Bootstrap

In the context of one-way clustering, it is now well known (e.g., Djogbenou et al. 2019) that asymptotic t-tests often suffer from large size distortions, and that the wild bootstrap, or WB, and in particular the wild cluster bootstrap, or WCB (Cameron, Gelbach, and Miller 2008), can provide more reliable inference. In this section, we therefore consider inference based on several variants of the WB and WCB as alternatives to the t-tests justified in Theorem 1.

For the wild bootstrap, the bootstrap disturbance vectors $\boldsymbol{u}^*$ are obtained by multiplying each residual, either $\tilde{u}_{gh,i}$ (restricted) or $\hat{u}_{gh,i}$ (unrestricted), by independent draws $v^*_{gh,i}$ from an auxiliary random variable v^* that satisfies the condition:

Assumption 7. v^* is independent of $(N_{gh}, \mathcal{S}_{gh})$, with $\mathrm{E}(v^*) = 0$, $\mathrm{E}(v^{*2}) = 1$, and $\mathrm{E}(v^{*4}) < \infty$.

A popular choice is the Rademacher random variable, which takes the values 1 and -1 with equal probabilities; see Davidson and Flachaire (2008). Thus, for the WB, each bootstrap sample uses N draws from the auxiliary distribution.

For the wild cluster bootstrap, the number of draws from the auxiliary distribution is equal to the number of clusters instead of the number of observations. For the two-way model (2), there are three natural ways to cluster the bootstrap disturbances. We can cluster by the first dimension, by the second dimension, or by their intersection. The number of draws would then be G, H, or GH (actually, if any of the possible intersections of the two dimensions

were empty, the number of draws would be less than GH, but for simplicity we ignore this possibility). For each bootstrap sample, every residual within each cluster in the appropriate dimension is multiplied by the same draw from v^*. When the total number of draws is small, roughly less than 10, using a two-point distribution can cause problems; see Webb (2014).

The idea of the WCB is that the bootstrap samples should preserve the pattern of correlations within each cluster. This idea works well for one-way clustering. When clustering is in two dimensions, the best the WCB can do is to preserve some of the intra-cluster correlations. Of course, this does not imply that the WCB will fail when clustering is in multiple dimensions, because we are bootstrapping a pivotal statistic. For example, the subcluster wild bootstrap (MacKinnon and Webb 2018) and the WB are both valid in the context of the one-way clustered model, even though they do not replicate the intra-cluster correlation structure; see Djogbenou, MacKinnon, and Nielsen (2019).

Because the WCB cannot replicate multiway clustering, it is impossible to achieve an asymptotic refinement. Nevertheless, there is still some theoretical rationale for applying the bootstrap. For example, in the context of heteroskedasticity and autocorrelation robust testing, Gonçalves and Vogelsang (2011) show theoretically that the i.i.d. bootstrap is more accurate than the standard normal approximation, even when data are serially correlated, as long as the test statistic is calculated in the same way for the bootstrap data and the original data. In our context, bootstrap tests may yield more accurate inferences than t-tests if the mistakes made in obtaining, say, $\hat{\boldsymbol{V}}_3^*$ for the bootstrap samples are similar to the ones made in obtaining $\hat{\boldsymbol{V}}_3$ for the actual sample, even if the dependence structure in the data cannot be replicated by the bootstrap DGP. It is thus of considerable interest to investigate wild cluster bootstrap tests, and below we give results on the asymptotic validity or failure of several versions of the WCB.

We next describe the algorithms for all variants of the WB and WCB in detail. Both the WB and WCB may be implemented using either restricted (henceforth WR and WCR) or unrestricted (WU and WCU) estimates in the bootstrap DGP. All these procedures are implemented in the (computationally very efficient) Stata package `boottest`; see Roodman, MacKinnon, Nielsen, and Webb (2019). To unify notation, we introduce $\ddot{\boldsymbol{u}}$ and $\ddot{\boldsymbol{\beta}}$, which will represent either restricted or unrestricted quantities as appropriate.

Multiway Wild (Cluster) Bootstrap Algorithms.

1. Regress $\boldsymbol{y}$ on $\boldsymbol{X}$ by OLS to obtain $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{u}}$, and $\hat{\boldsymbol{V}}_j$ for $j \in \{2, 3\}$ defined in (6) and (9). Check whether $\hat{\boldsymbol{V}}_3$ is positive semidefinite, and replace it by $\hat{\boldsymbol{V}}_3^+$ if necessary; see Section 2. For WR and WCR, additionally re-estimate model (1) subject to the restriction $\boldsymbol{a}^\top\boldsymbol{\beta} = \boldsymbol{a}^\top\boldsymbol{\beta}_0$ so as to obtain restricted estimates $\tilde{\boldsymbol{\beta}}$ and restricted residuals $\tilde{\boldsymbol{u}}$.
2. Calculate the cluster-robust t-statistic $t_{a,j}$, given in (10), for H_0: $\boldsymbol{a}^\top\boldsymbol{\beta} = \boldsymbol{a}^\top\boldsymbol{\beta}_0$.
3. For each of B bootstrap replications, indexed by b:
   (a) Generate a new set of bootstrap disturbances given by $\boldsymbol{u}^{*b}$. For the wild bootstrap, set $u^{*b}_{gh,i} = v^{*b}_{gh,i}\ddot{u}_{gh,i}$. For the wild cluster bootstrap, set $\boldsymbol{u}^{*b}_{gh} = v^{*b}_{gh}\ddot{\boldsymbol{u}}_{gh}$, or $\boldsymbol{u}^{*b}_{g} = v^{*b}_{g}\ddot{\boldsymbol{u}}_{g}$, or $\boldsymbol{u}^{*b}_{h} = v^{*b}_{h}\ddot{\boldsymbol{u}}_{h}$, depending on the level of bootstrap clustering.
   (b) Generate the bootstrap dependent variables according to $\boldsymbol{y}^{*b} = \boldsymbol{X}\ddot{\boldsymbol{\beta}} + \boldsymbol{u}^{*b}$.

(c) Obtain the bootstrap estimates $\hat{\boldsymbol{\beta}}^{*b} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}^{*b}$, the bootstrap residuals $\hat{\boldsymbol{u}}^{*b}$, and one of the bootstrap variance matrix estimates $\hat{\boldsymbol{V}}_2^{*b} = \hat{\boldsymbol{V}}_G^{*b} + \hat{\boldsymbol{V}}_H^{*b}$ or $\hat{\boldsymbol{V}}_3^{*b} = \hat{\boldsymbol{V}}_G^{*b} + \hat{\boldsymbol{V}}_H^{*b} - \hat{\boldsymbol{V}}_I^{*b}$, with $\hat{\boldsymbol{V}}_m^{*b} = \boldsymbol{Q}^{-1} \hat{\boldsymbol{\Gamma}}_m^{*b} \boldsymbol{Q}^{-1}$ for $m \in \{G, H, I\}$, where

$$\begin{aligned} \hat{\boldsymbol{\Gamma}}_G^{*b} &= \frac{1}{(GH)^2} \sum_{g=1}^{G} \boldsymbol{X}_g^\top \hat{\boldsymbol{u}}_g^{*b} \hat{\boldsymbol{u}}_g^{*b\top} \boldsymbol{X}_g, \quad \hat{\boldsymbol{\Gamma}}_H^{*b} = \frac{1}{(GH)^2} \sum_{h=1}^{H} \boldsymbol{X}_h^\top \hat{\boldsymbol{u}}_h^{*b} \hat{\boldsymbol{u}}_h^{*b\top} \boldsymbol{X}_h, \\ \hat{\boldsymbol{\Gamma}}_I^{*b} &= \frac{1}{(GH)^2} \sum_{g=1}^{G} \sum_{h=1}^{H} \boldsymbol{X}_{gh}^\top \hat{\boldsymbol{u}}_{gh}^{*b} \hat{\boldsymbol{u}}_{gh}^{*b\top} \boldsymbol{X}_{gh}, \end{aligned} \tag{22}$$

and each term in (22) should be multiplied by the appropriate factor in (8) if the corresponding term in $\hat{\boldsymbol{V}}_2$ or $\hat{\boldsymbol{V}}_3$ is multiplied by it. If $\hat{\boldsymbol{V}}_3^{*b}$ is not positive semidefinite, replace it by $\hat{\boldsymbol{V}}_3^{*b+}$, which is the bootstrap analog of $\hat{\boldsymbol{V}}_3^{+}$.

(d) Calculate the bootstrap t-statistic

$$t_{a,j}^{*b} = \frac{\boldsymbol{a}^\top (\hat{\boldsymbol{\beta}}^{*b} - \ddot{\boldsymbol{\beta}})}{\sqrt{\boldsymbol{a}^\top \hat{\boldsymbol{V}}_j^{*b} \boldsymbol{a}}}. \tag{23}$$

4. Depending on whether the alternative hypothesis is $\mathrm{H_L}$: $\boldsymbol{a}^\top \boldsymbol{\beta} < \boldsymbol{a}^\top \boldsymbol{\beta}_0$, $\mathrm{H_R}$: $\boldsymbol{a}^\top \boldsymbol{\beta} > \boldsymbol{a}^\top \boldsymbol{\beta}_0$, or $\mathrm{H_2}$: $\boldsymbol{a}^\top \boldsymbol{\beta} \neq \boldsymbol{a}^\top \boldsymbol{\beta}_0$, compute one of the following bootstrap P values:

$$\hat{P}_{\mathrm{L}}^* = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}(t_a^{*b} < t_a), \quad \hat{P}_{\mathrm{R}}^* = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}(t_a^{*b} > t_a) \quad \text{or} \quad \hat{P}_S^* = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}(|t_a^{*b}| > |t_a|),$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. If the alternative hypothesis is $\mathrm{H_2}$, then the symmetric P value $\hat{P}_S^*$ could be replaced by the equal-tail P value, $2\min(\hat{P}_{\mathrm{L}}^*, \hat{P}_{\mathrm{R}}^*)$.

The above algorithm presents the steps needed to implement the WR, WCR, WU, and WCU bootstraps for testing the hypothesis $\mathrm{H_0}$. If interest focuses on confidence intervals for $\boldsymbol{a}^\top \boldsymbol{\beta}$, there are two approaches. Studentized bootstrap confidence intervals based on WU or WCU can easily be constructed by calculating lower-tail and upper-tail quantiles of the t_a^{*b} instead of P values; see Davidson and MacKinnon (2004, Sec. 5.3). Restricted studentized bootstrap confidence intervals that usually have better finite-sample properties can be obtained by inverting equal-tail bootstrap tests based on WR or WCR, as discussed in Hansen (1999) and MacKinnon (2015). Unless GH is very large, the implementation of these tests in `boottest` is so computationally efficient that test inversion is feasible even for extremely large data sets; see Roodman et al. (2019).

The asymptotic validity of the WB and WCB tests is investigated in our next results, where we derive the properties of the bootstrap test statistics, $t_{a,j}^*$. We use an asterisk to denote the bootstrap probability measure and associated expectation and variance, noting that these are different depending on the choice of bootstrap; c.f. step 3(a).

To state the asymptotic results for the bootstrap test statistics, we need to consider not only the cases (16)–(18) but also intermediate cases where there is clustering along one dimension, but not the other. Specifically, we sometimes divide condition (16) into the additional two conditions,

$$\boldsymbol{V}_G \ > 0, \tag{24}$$

$$\boldsymbol{V}_H \ > 0 \text{ and also } (N_{gh}, \mathcal{S}_{gh}) \text{ is independent of } (N_{g'h'}, \mathcal{S}_{g'h'}) \text{ if } h \neq h'. \tag{25}$$

The condition in (24) guarantees that there is clustering in the first (G) dimension, but it is silent about whether or not there is clustering in the second (H) dimension. On the other hand, under (25), there is no clustering in the first dimension, but there is clustering in the second dimension (ensured by the condition $\boldsymbol{V}_H > 0$). The reverse case is symmetric.

The following theorem is the bootstrap analog of Corollary 1. It establishes the asymptotic distribution of the WB and WCB t-statistics.

Theorem 3. *Suppose Assumptions 1–7 are satisfied and that H_0 is true.*

(i) Suppose the bootstrap DGP in step 3(a) is clustered along the first (G) dimension (results for bootstrap clustering along the second dimension are symmetric).

(a) If (24) holds, so that the DGP is clustered along the first dimension, then

$$t^*_{a,2} \xrightarrow{d^*} \mathrm{N}(0,1) \quad \textit{and} \quad t^*_{a,3} \xrightarrow{d^*} \mathrm{N}(0,1), \ \textit{in probability.}$$

(b) If (25) holds, so that the DGP is not clustered along the first dimension, but it is clustered along the second dimension, then

$$t^*_{a,2} \xrightarrow{d^*} \left(\frac{W_1^2}{W_1^2 + W_0 + \boldsymbol{a}^\top \boldsymbol{V}_I \boldsymbol{a}}\right)^{1/2} Z \quad \textit{and} \quad t^*_{a,3} \xrightarrow{d^*} \left(\frac{W_1^2}{W_1^2 + W_0}\right)^{1/2} Z,$$

in probability, where Z, W_1^2, and W_0 are mutually independent random variables satisfying $Z \sim \mathrm{N}(0,1)$, $W_1^2 > 0$ almost surely, and $\mathrm{E}(W_0) = 0$.

(c) If either (17) or (18) holds, so that the DGP is clustered by intersections or not clustered at all, then

$$t^*_{a,2} \xrightarrow{d^*} \mathrm{N}(0,1/2) \quad \textit{and} \quad t^*_{a,3} \xrightarrow{d^*} \mathrm{N}(0,1), \ \textit{in probability.}$$

(ii) If the bootstrap DGP in step 3(a) is either clustered by intersections or is the WB, then

$$t^*_{a,2} \xrightarrow{d^*} \mathrm{N}(0,1/2) \quad \textit{and} \quad t^*_{a,3} \xrightarrow{d^*} \mathrm{N}(0,1), \ \textit{in probability.}$$

Note that, as usual, all the results in Theorem 3 are conditional on the original sample, and hence also conditional on $t_{a,j}$. This implies that the results in Theorem 3 hold for any possible realization of the original sample, and therefore also any possible realization of $t_{a,j}$, which is the crucial requirement for asymptotic validity of the bootstrap.

It is well-known that, by Polya's theorem and the triangle inequality, if the asymptotic distribution of the bootstrap t-statistic $t^*_{a,j}$ in (23) correctly replicates that of the original sample t-statistic $t_{a,j}$ in (10), then the bootstrap is asymptotically valid in the sense that

$$\sup_x |P^*(t^*_{a,j} \leq x) - P(t_{a,j} \leq x)| = o_p(1), \tag{26}$$

where $P(t_{a,j} \leq x)$ denotes the cumulative distribution function (CDF) of $t_{a,j}$ and $P^*(\cdot)$ the corresponding bootstrap CDF. When (26) holds, the P values computed in step 4 of the WB and WCB algorithms are asymptotically valid, as are studentized bootstrap confidence intervals. The cases in which (26) holds are summarized in Table 1.

Table 1: Asymptotic validity of bootstrap tests

	WCB_G		WCB_I		WB	
DGP clustering	$t^*_{a,2}$	$t^*_{a,3}$	$t^*_{a,2}$	$t^*_{a,3}$	$t^*_{a,2}$	$t^*_{a,3}$
By first dimension (G), (24)	valid	valid	OR	valid	OR	valid
By second dimension (H) only, (25)	see text	see text	OR	valid	OR	valid
By intersections (I) only, (17)	valid	valid	valid	valid	valid	valid
Independent observations, (18)	valid	valid	valid	valid	valid	valid

Notes: WCB_G and WCB_I denote the WCB with bootstrap clustering along the first (G) dimension and by intersections, respectively, and WB denotes the ordinary wild bootstrap; c.f. step 3(a). The results for WCB_H are symmetric to those of WCB_G. "valid" indicates that (26) holds, "OR" denotes asymptotic over-rejection, and "see text" refers to case (i)(b) of Theorem 3 discussed in the main text.

From Theorem 3 and Table 1, we see that the distribution of the bootstrap t-statistic correctly replicates that of the original sample t-statistic in many cases. If so, the bootstrap test is asymptotically valid. This essentially follows from the fact that the t-statistic is asymptotically pivotal. Even though $\boldsymbol{a}^\top \hat{\boldsymbol{\beta}}^*$ does not have the same variance as $\boldsymbol{a}^\top \hat{\boldsymbol{\beta}}$, as shown in Theorems A.1 and A.2 in the appendix, this does not cause a failure of the bootstrap, because $t^*_{a,j}$ still has the correct asymptotic distribution in those cases.

In other cases, however, most notably for bootstrap tests based on $t_{a,2}$, the distribution of the bootstrap t-statistic does not coincide with that of the original sample t-statistic. In several cases, $t_{a,2}$ is asymptotically distributed as $\text{N}(0,1)$, while the bootstrap distribution of $t^*_{a,2}$ is asymptotically $\text{N}(0,1/2)$. Conducting inference based on the distribution of $t^*_{a,2}$ then leads to over-rejection, and these cases are indicated by "OR" in Table 1.

A particularly challenging situation arises in case (i)(b) in Theorem 3, which is labelled "see text" in Table 1. In this case, both $t^*_{a,2}$ and $t^*_{a,3}$ are asymptotically mixed normal, and the conditional variances are the random variables given in the large parentheses in case (i)(b) in Theorem 3. There are several interesting remarks. First, the conditional variance of $t^*_{a,2}$ is always less than that of $t^*_{a,3}$ because of the positive term, $\boldsymbol{a}^\top \boldsymbol{V}_I \boldsymbol{a}$, in the denominator. Thus, asymptotically, the bootstrap test based on $t^*_{a,2}$ rejects more often than the one based on $t^*_{a,3}$.

Second, although both conditional variances can in principle be either greater than or less than one, for $t^*_{a,2}$ it seems unlikely to be greater than one because of the additional positive term in the denominator. Specifically, when $\boldsymbol{a}^\top \boldsymbol{V}_I \boldsymbol{a}$ is large relative to the zero-mean random variable W_0, the conditional variance of $t^*_{a,2}$ will be less than one, leading to over-rejection of the bootstrap test based on $t^*_{a,2}$. However, when the conditional variance is centered at one, the bootstrap test will tend to under-reject, because the mixed normal distribution has heavier tails than the standard normal distribution.

Third, the random variable W_0 is an infinite weighted sum of centered χ^2_1 distributions. When the weights are approximately equal (which happens when the clusters are not too heterogeneous), W_0 is well approximated by a normal distribution with mean zero. In such situations, the conditional variance of $t^*_{a,3}$ has median equal to one, and the bootstrap test based on $t^*_{a,3}$ is nearly valid in the sense that the mixed normal distribution of $t^*_{a,3}$ has conditional variance with the correct median, although the test will still under-reject because

the mixed normal distribution has heavier tails than the normal distribution.

It is generally desirable that a bootstrap DGP should replicate the key features of the true (unknown) DGP. In the context of cluster-robust inference, this means that in step 3(a) the bootstrap DGP should replicate the clustering structure to the extent possible. With the WCB considered in this paper, it is impossible to replicate multiway clustering, but it is possible to replicate clustering in one dimension. This consideration would tend to favor either WCB_G or WCB_H, which replicate as much as possible of the clustering structure in the DGP. However, there is to some extent a tradeoff between the desire to replicate the DGP clustering structure and the asymptotic validity results summarized in Table 1. We investigate this issue via extensive Monte Carlo simulations in Section 5.

Similarly, it is generally desirable to impose the null restrictions on the bootstrap DGP (Davidson and MacKinnon 1999), and there is compelling evidence in MacKinnon and Webb (2017) and Djogbenou et al. (2019) that, for one-way clustering, the restricted versions of both the WB and WCB, and particularly the latter, outperform the unrestricted ones. However, in the context of multiway clustering, it is not clear which variant of the WCB in step 3(a) is likely to perform best in any given case, or even whether any variant is likely to outperform the WB or the asymptotic t-test. This undoubtedly depends on G, H, $\boldsymbol{\Sigma}$, and so on. We use Monte Carlo simulations to investigate these issues in Section 5.

5 Simulation Experiments

We performed an extensive set of simulation experiments and report the most interesting results in this section. The objectives of these experiments are to confirm the theoretical predictions made in Sections 3 and 4 and to guide choices among the numerous ways in which inferences can be made when there is two-way clustering, including the choice of bootstrap.

In the experiments that we report, the DGP is

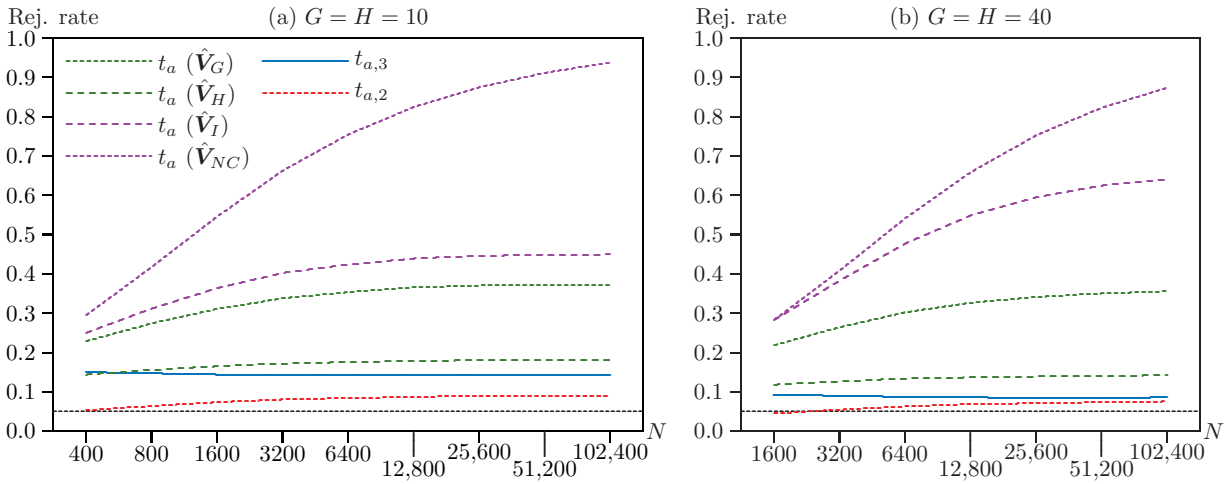
$$y_{gh,i} = \beta_0 + \beta_1 X_{gh,i} + u_{gh,i}, \quad u_{gh,i} = \sigma_1 v_g + \sigma_2 v_h + \sigma_\epsilon \epsilon_{gh,i}, \tag{27}$$

where v_g, v_h, and $\epsilon_{gh,i}$ are mutually independent standard normals. The values of σ_1, σ_2, and σ_ϵ are chosen so that the correlation between any two disturbances that belong to the same cluster in the G (or H) dimension is ρ_1 (or ρ_2). This implies that the correlation is $\rho_1 + \rho_2$ for disturbances that belong to the same cluster in both dimensions, and zero for ones that do not belong to the same cluster in either dimension.

The regressor $X_{gh,i}$ is lognormally distributed to avoid the risk that our results may be artifacts of an experimental design in which both the regressor and the disturbances are normally distributed. Specifically, $\log(X_{gh,i})$ is generated in the same way as $u_{gh,i}$, but with correlations ϕ_1 and ϕ_2. If $X_{gh,i}$ were normally distributed, many tests would perform somewhat better, but the overall pattern of the results would not change.

In our experiments, we replaced $\hat{\boldsymbol{V}}_3$ by $\hat{\boldsymbol{V}}_3^+$ when necessary, as discussed in Section 2. When $G = H = 10$, $\hat{\boldsymbol{V}}_3$ quite often had negative eigenvalues, in extreme cases as much as a quarter of the time. Negative eigenvalues were much less common for larger values of G and H. Even when $G = H = 10$, we only very rarely encountered cases in which the standard error of $\hat{\beta}_1$ based on $\hat{\boldsymbol{V}}_3^+$ was not positive. In those rare cases, we set the standard error to a small number such that $t_{a,3}$ was very large. We did the same thing when calculating bootstrap test statistics. This happened so rarely that it should have a negligible effect on the results.

Figure 1: Rejection frequencies for six forms of t-statistic

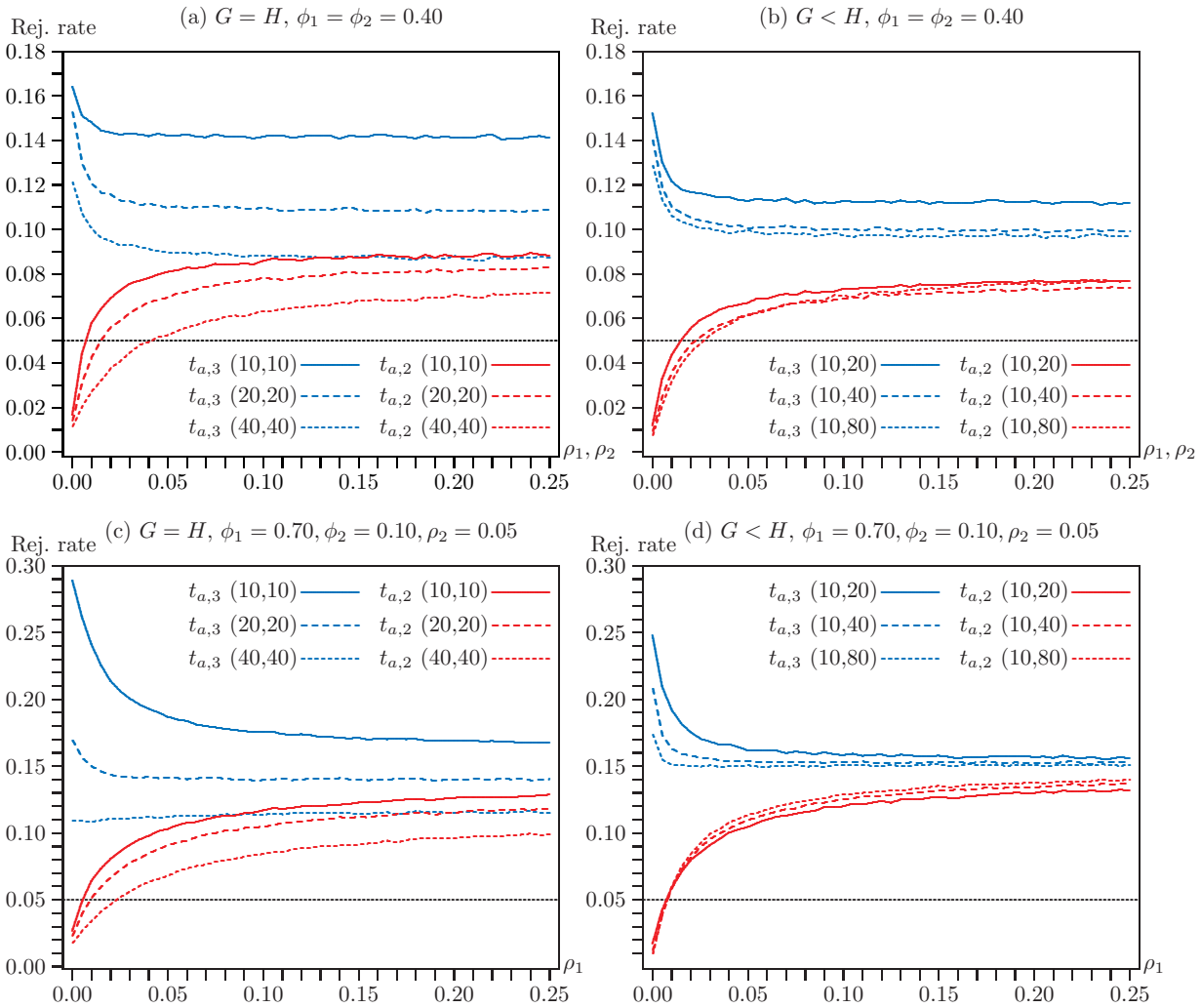


Notes: There are 400,000 replications. All tests are at the 5% nominal level. Parameter values are $\rho_1 = 0.05$, $\rho_2 = 0.15$, $\phi_1 = 0.40$, and $\phi_2 = 0.40$. $\hat{\boldsymbol{V}}_m$ for $m = G$, H, and I denote one-way CRVEs with clustering by G, by H, and by intersection, respectively. $\hat{\boldsymbol{V}}_{NC}$ denotes the heteroskedasticity-consistent variance matrix called HC_1 in MacKinnon and White (1985), which is the default in Stata. Rejection frequencies for $t_{a,2}$, $t_{a,3}$, and the t-statistics that use $\hat{\boldsymbol{V}}_G$ and $\hat{\boldsymbol{V}}_H$ are based on the $t(G-1) = t(H-1)$ distribution. For the t-statistics that use $\hat{\boldsymbol{V}}_I$ and $\hat{\boldsymbol{V}}_{NC}$, they are based on the $t(GH-1)$ distribution and the $t(N-2)$ distribution, respectively.

Our experiments cover a much wider range of cases than those of CGM, and they differ from the latter in two important respects. CGM's DGP has two regressors, each correlated in just one dimension, instead of one regressor that is correlated in both dimensions. In addition, the values of ρ_1 and ρ_2 in our experiments are generally smaller than the values of ϕ_1 and ϕ_2. That is because, in our experience, intra-cluster correlations for residuals tend to be small, while intra-cluster correlations for at least some regressors can be large. A number of our experiments deal with cases in which ρ_1 and/or ρ_2 is zero, which are of great interest in view of Theorem 2 and the practical possibility that we may be (multiway) clustering even when it is not needed. In contrast, CGM implicitly sets $\rho_1 = \rho_2 = 1/3$ in many experiments.

Figure 1 illustrates that it is essential to use two-way clustered standard errors when there actually is two-way clustering. It shows rejection frequencies for six different t-tests. The t-statistics $t_{a,2}$ and $t_{a,3}$ were defined in (10) and use two-way clustering. The others use either one-way clustering or no clustering at all. With the exception of $t_{a,3}$, all methods over-reject more severely as N increases. This is most apparent for the t-test that makes no allowance for clustering, which rejects over 90% of the time for the largest sample sizes when $G = H = 10$. The t-test based on clustering by intersection also over-rejects very severely for large N, even more so when there are 40 clusters in each dimension than when there are only 10. This makes sense, because the intersections become smaller as the number of clusters increases. It is also not surprising that the t-test based on clustering by G performs much worse than the one based on clustering by H. Because $\rho_2 > \rho_1$, the correlations that are ignored by the former test statistic are larger than the ones ignored by the latter. If we had set $\rho_1 = \rho_2$, these tests would have had the same rejection frequencies, because the data would have been completely symmetric in the G and H dimensions.

Figure 2: Rejection frequencies for two-way t-tests

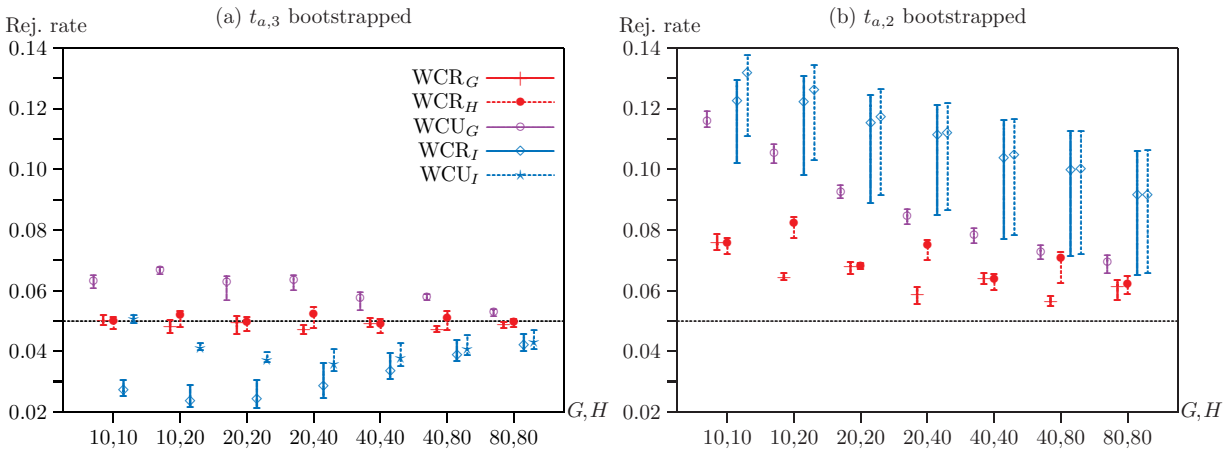


Notes: There are 400,000 replications, and the sample size N is always 6400. All tests are at the 5% nominal level. Rejection frequencies are based on the 0.975 quantile of the $t(R-1)$ distribution.

The t-statistics based on two-way clustering clearly outperform all the others in Figure 1, especially in large samples, because they are the only ones that are asymptotically valid. $t_{a,2}$ always rejects less often than $t_{a,3}$, because $\hat{\boldsymbol{V}}_2$ equals $\hat{\boldsymbol{V}}_3$ minus a positive semidefinite matrix; see (6). The difference between the rejection frequencies for $t_{a,3}$ and $t_{a,2}$ is much greater in Panel (a) than in Panel (b), because the intersections contain more observations in the former case. In both panels, this difference diminishes as N increases.

The most important implication of Figure 1 is that failing to use two-way clustering when there is actually correlation in two dimensions can lead to very severe errors of inference, and these errors become more severe as the sample size increases. In the remainder of this section, we therefore focus exclusively on tests based on the statistics $t_{a,2}$ and $t_{a,3}$. Even though two-way t-tests work less badly than other methods in Panel (a), and reasonably well in Panel (b), they still over-reject and do not yield particularly reliable inferences.

Figure 3: Rejection frequencies for wild cluster bootstrap tests



Notes: There are 100,000 replications, and $N = 6400$. All bootstrap tests use $B = 399$ and reject whenever $\hat{P}_S^* < 0.05$. In all cases, $\phi_1 = \phi_2 = 0.40$. For each method and each pair of G, H values, the top of the vertical line shows the largest observed rejection frequency across the cases $\rho_1 = \rho_2 = 0.01, 0.02, \ldots, 0.10$, the bottom of the line shows the smallest one, and the mean over the ten frequencies is shown by a symbol.

Figure 2 has four panels that differ in the number of clusters (G, H) and in the intra-cluster correlations of the regressor (ϕ_1, ϕ_2). One striking result in all four panels is that $t_{a,2}$ under-rejects very severely when there is no correlation of the disturbances in either one or both dimensions. This is predicted in case (ii) of Corollary 1, where $t_{a,2}$ asymptotically has a variance of $1/2$. This implies that tests based on it reject 0.56% of the time at the 5% level when they use the asymptotic critical value 1.96, and even less often when they use critical values from various t distributions. In fact, rejection frequencies for $t_{a,2}$ vary between 0.0074 and 0.0166 when $\rho_1 = \rho_2 = 0$ and between 0.0092 and 0.0266 when $\rho_1 = 0$ and $\rho_2 = 0.05$. As the asymptotic theory suggests, these rejection frequencies decline as either G or H increases.

In contrast, tests based on $t_{a,3}$ always over-reject, and they do so particularly severely when there is no or little correlation of the disturbances in either one or both dimensions. As must be the case, tests based on $t_{a,3}$ always reject more often than ones based on $t_{a,2}$. The difference is most pronounced when there is little correlation of the disturbances in either one or both dimensions, and it diminishes as either G or H increases. Tests based on $t_{a,2}$ also over-reject except for small values of the correlation coefficients. For both test statistics, the over-rejection diminishes in Panels (a)–(c) when we increase either $G = H$ or only H. The same is true for $t_{a,3}$ in Panel (d), when we increase H but not G. However, for the tests based on $t_{a,2}$ in Panel (d), increasing H when G is fixed at 10 actually causes over-rejection to increase.

Finite-sample distortions may arise from two main sources. The first is inaccuracy in the central limit theorem approximation to the sampling distribution of $\hat{\boldsymbol{\beta}}$. The second is the bias and sampling variability of the CRVE. Because the disturbances in our simulation DGP (27) are normal, although the regressors are lognormal, the CLT approximation is probably accurate. Therefore, the mediocre performance of the two-way t-tests in Figure 2 likely reflects the fact that, in many cases, neither $\hat{\boldsymbol{V}}_2$ nor $\hat{\boldsymbol{V}}_3$ provides a good estimate. As discussed previously, because the mistakes made in obtaining, say, $\hat{\boldsymbol{V}}_3^*$ for the bootstrap samples are

similar to the ones made in obtaining $\hat{\boldsymbol{V}}_3$ for the actual sample, it is of considerable interest to investigate wild cluster bootstrap tests. We study six of them: WCR_G, WCR_H, and WCR_I (wild cluster restricted with bootstrap clustering by G, by H, and by intersection), WCU_G, WCU_H, and WCU_I (wild cluster unrestricted with the same three forms of bootstrap clustering). Because the wild bootstraps (WR and WU) are expensive to compute and tend to yield results similar to WCR_I and WCU_I, we do not report results for them. Also, to keep the figures readable, we do not always report results for all six WCB tests.

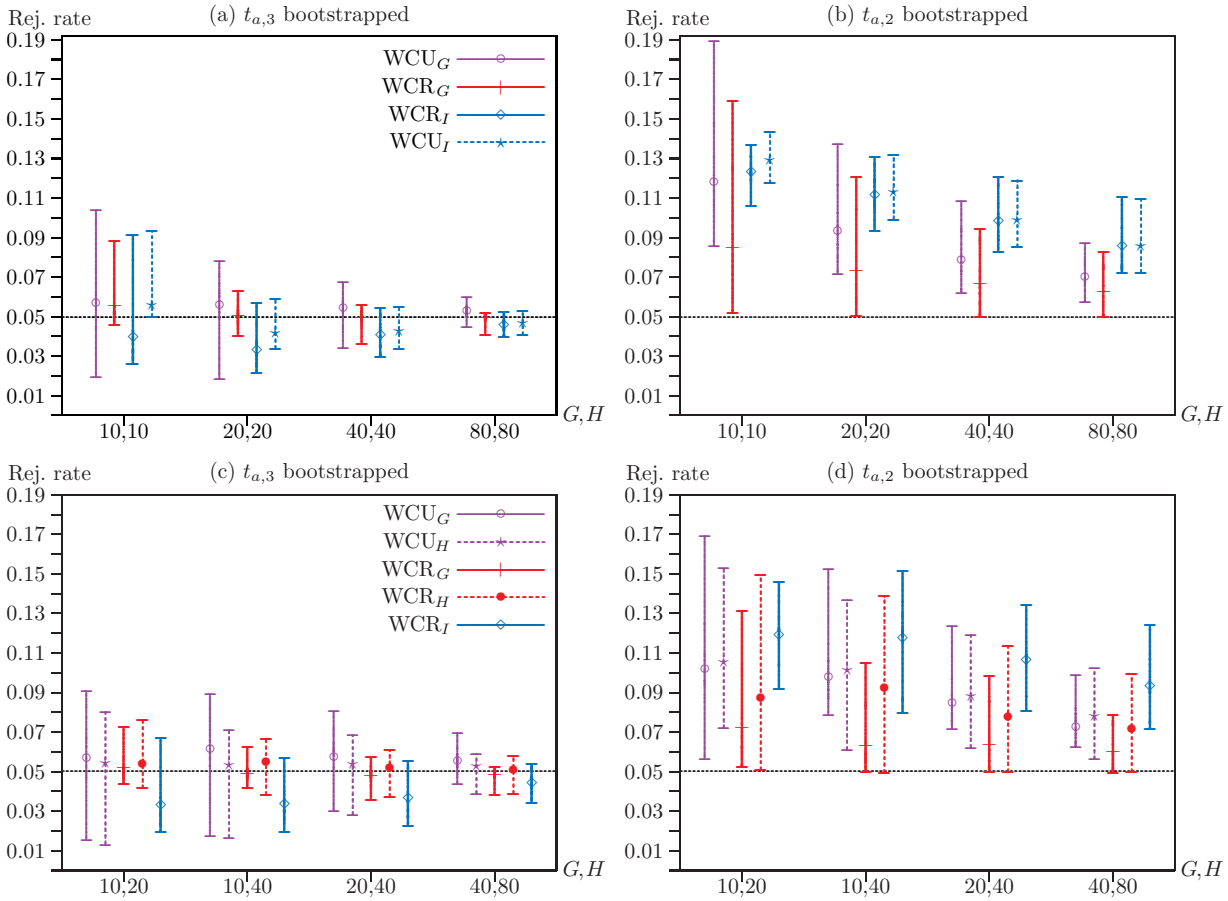
Figure 3 shows rejection frequencies for five wild cluster bootstrap tests for each of $t_{a,3}$ and $t_{a,2}$ in Panels (a) and (b), respectively. In Panel (a), we see that WCR_G and WCR_H based on $t_{a,3}$ always work extremely well. Note that, when $G = H$ in this figure, the only differences between the two reflect simulation randomness. The other four tests do not perform as well. WCU_G and WCU_H (not shown) always over-reject, which is expected for WCU tests; see Djogbenou et al. (2019). The two tests that cluster the bootstrap samples by intersection, WCR_I and WCU_I, do not perform well in most cases, except for WCU_I when $G = H = 10$. This case seems to be an accident. We conjecture that the tendency of all the WCU tests to over-reject just happens, in this case, to offset a tendency for bootstrapping by intersection to under-reject. Note that, because $N = 6400$, WR and WU are identical to WCR_I and WCU_I when $G = H = 80$.

Panel (b) of Figure 3, which deals with WCB tests based on $t_{a,2}$, looks very different from Panel (a). All of the bootstrap tests now over-reject in every case. Even the best of them, WCR_G, never performs particularly well. Interestingly, it performs best when $G < H$. The two tests that bootstrap by intersection, which almost always under-reject in Panel (a), now over-reject more severely than any of the other tests. Their performance is also very sensitive to the values of ρ_1 and ρ_2. They perform much better for small values of those coefficients than for large ones. The figure intentionally omits the case in which $\rho_1 = \rho_2 = 0$, where these tests reject very close to 5% of the time. This case will be discussed below in the context of Figure 5. For the cases studied here, it is rarely worthwhile to bootstrap $t_{a,2}$. Rejection frequencies for t-tests based on the t distribution are often closer to 0.05 than those for any of the bootstrap tests. We provide more evidence on this issue later in this section.

In Figure 3, the values of ρ_1 and ρ_2 are always the same, and the values of ϕ_1 and ϕ_2 are always 0.40. In Figure 4, we relax both these constraints as detailed in the notes to the figure. Not surprisingly, the large set of parameter values across the 60 experiments means that the vertical lines are generally longer than their counterparts in Figure 3. In Panels (a) and (c) of Figure 4, we present results for $t_{a,3}$ bootstrapped in various ways. There are no results for bootstrapping by H in Panel (a) because, with $G = H$, these methods are equivalent to bootstrapping by G. There is not a lot to choose among the various procedures in Panel (a), except that the vertical line tends to be longer for WCU_G. The only surprising result is that WCU_G under-rejects in a number of cases, although it always over-rejects on average.

In Panel (c), we do show results for WCR_H and WCU_H, and we omit the ones for WCU_I to save space. The best methods are WCR_G and WCR_H, although the latter rejects a bit more often than the former, and it always has a larger range of outcomes. Thus, as in Figure 3, there appears to be modest evidence in favor of bootstrapping by the dimension with the smallest number of clusters. However, this might not be the right thing to do if,

Figure 4: Rejection frequencies for wild cluster bootstrap tests



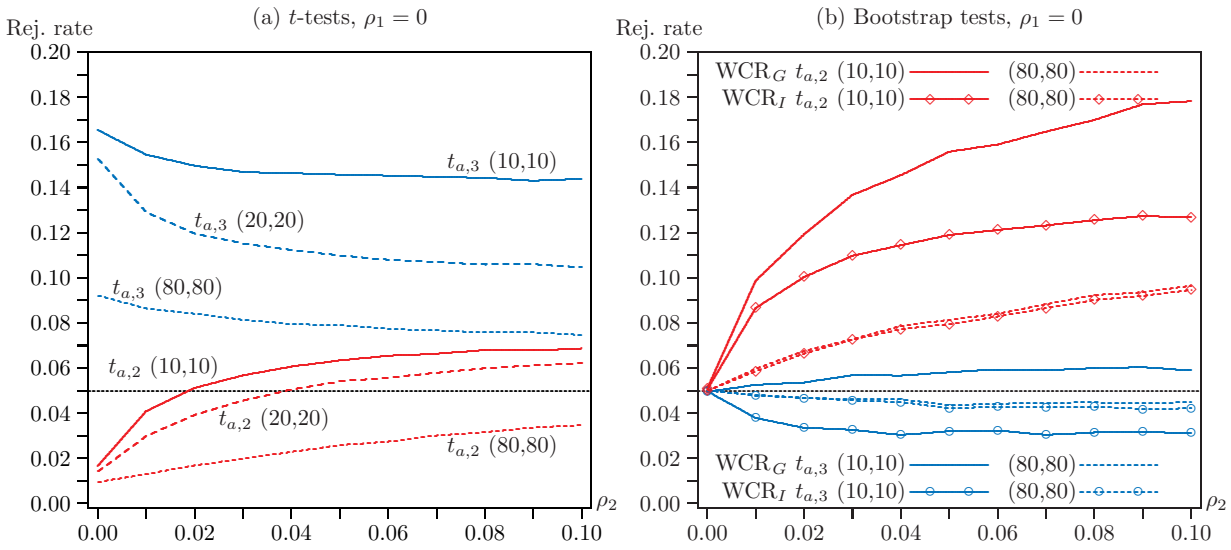
Notes: There are 100,000 replications, and $N = 6400$. All bootstrap tests use $B = 399$ and reject whenever $\hat{P}^*_S < 0.05$. There are 60 cases: for $i = 1, 2$ and $j = 2, 1$, $\rho_i = 0.05$, $\rho_j = 0.00, 0.02, \ldots, 0.10$, $\phi_i = 0.30$, and $\phi_j = 0.00, 0.15, \ldots, 0.60$. For each method and each pair of G, H values, the top of the vertical line shows the largest observed rejection frequency across the 60 cases, the bottom of the line shows the smallest one, and the mean over the 60 frequencies is shown by a symbol.

for example, the other dimension had substantially more intra-cluster correlation.

In Panels (b) and (d) of Figure 4, we see once again that bootstrapping $t_{a,2}$ almost always leads to over-rejection, which can be very severe. The cases in which WCR$_G$ and WCR$_H$ do not over-reject are typically ones in which there is no intra-cluster correlation for the dimension by which we do not bootstrap. In contrast, over-rejection tends to be very severe when there is a lot of intra-cluster correlation in the dimension by which we do not bootstrap.

In Figures 2 and 4, we saw that extreme results for $t_{a,2}$ and its bootstrapped counterparts tend to occur in the important special case in which there is no correlation in one dimension. In this case, a multiway CRVE has been applied even though it is in fact not needed. Figure 5 deals with this case in more detail. In both panels, $\rho_1 = 0$ and ρ_2 varies from 0.00 to 0.10. Thus, the leftmost point in each panel corresponds to case (iii) of Theorems 1 and 2, where the disturbances are independent. In Panel (a), we see that t-tests based on $t_{a,3}$ are prone to over-reject, while t-tests based on $t_{a,2}$ under-reject in this case, at least for large values of $G = H$.

Figure 5: Rejection frequencies for t-tests and wild cluster bootstrap tests



Notes: There are 100,000 replications, and $N = 6400$. All bootstrap tests use $B = 399$ and reject whenever $\hat{P}_S^* < 0.05$. In all cases, $\phi_1 = \phi_2 = 0.40$, and $\rho_1 = 0$.

Panel (b) of Figure 5 shows that using WCR$_G$ or WCR$_I$ with either t-statistic works perfectly when there is actually no correlation in either dimension. However, as soon as ρ_2 exceeds zero, the rejection frequencies for $t_{a,2}$ increase sharply. The rejection frequencies for $t_{a,3}$ also become worse as ρ_2 increases, but only very slightly. Thus, in this case (and all other cases, at least for WCR$_G$), bootstrapping $t_{a,3}$ yields reasonably accurate results, but bootstrapping $t_{a,2}$ yields very inaccurate ones.

The performance of the tests varies with the numbers of clusters in each dimension and the four parameters of the DGP (27). To provide an overall summary, Table 2 reports the average, over the 70 cases in Figures 3 and 4, of the absolute error in rejection percentage. For the tests based on $t_{a,2}$, the table reports results only for the t-test and the WCR$_G$ test, which is always the best of the bootstrap tests for $t_{a,2}$. In most cases, the former test outperforms the latter, although not always and not for larger values of G and H. In fact, the t-test based on $t_{a,2}$ performs worse for $(80, 80)$ than for any other values of G and H.

For the tests based on $t_{a,3}$, Table 2 reports results for the t-test and all six WCB tests. As in Figures 2 and 5, the t-test performs poorly and is clearly the worst test, often by a substantial margin. This is rather worrying in light of the fact that it is, at time of writing, by far the most commonly used procedure in empirical work. In contrast, at least some of the bootstrap tests always perform well. The best test in almost every case is WCR$_G$ based on $t_{a,3}$, which is identical to WCR$_H$ when $G = H$ and always outperforms it, at least by a little, when $G < H$. The two WCU tests and the two tests that bootstrap by intersection generally have little to recommend them, except for $G = H = 10$, where WCU$_I$ works slightly better than WCR$_G$. But, as explained above, this seems to be an aberration. Overall, based on the extensive simulation evidence summarized in Table 2, we comfortably recommend general use of WCR$_G$ based on $t_{a,3}$; that is, bootstrap clustering by the dimension with the smallest

Table 2: Average absolute error in rejection percentage for various tests

Method \ G, H	$10,10$	$10,20$	$10,40$	$20,20$	$20,40$	$40,40$	$40,80$	$80,80$
$t_{a,2}$ $t(R-1)$	1.83	1.31	1.29	1.29	1.21	1.28	1.62	2.04
$t_{a,2}$ WCR_G	3.35	2.10	1.23	2.26	1.28	1.65	0.95	1.26
$t_{a,3}$ $t(R-1)$	9.22	6.09	4.64	5.67	4.21	3.62	2.87	2.36
$t_{a,3}$ WCR_G	0.54	0.34	0.30	0.28	0.30	0.21	0.21	0.14
$t_{a,3}$ WCR_H	0.54	0.47	0.58	0.28	0.33	0.21	0.24	0.14
$t_{a,3}$ WCU_G	1.66	1.60	1.93	1.27	1.26	0.67	0.64	0.29
$t_{a,3}$ WCU_H	1.66	1.27	0.97	1.27	0.79	0.67	0.41	0.29
$t_{a,3}$ WCR_I	1.77	2.00	1.81	1.75	1.52	1.12	0.74	0.51
$t_{a,3}$ WCU_I	0.52	0.77	1.02	1.02	1.03	0.87	0.62	0.46

Notes: Results for each G, H pair are based on the 10 cases in Figure 3 and the 60 cases in Figure 4; see the notes to these figures for details. The reported numbers in the table are in percentages.

number of clusters, based on the three-term CRVE and restricted parameter estimates.

6 Empirical Example

To illustrate the effects of using different methods of inference with multiway clustering, we consider an empirical example from Nunn and Wantchekon (2011), hereafter NW. This paper investigates whether current trust levels among different ethnic groups in several African countries are related to historical slave exports. NW studies the relationship between the volume of slave exports and current levels of trust between ethnicities using the equation

$$\text{trust}_{iedc} = \boldsymbol{\alpha}_c + \beta \text{exports}_e + \boldsymbol{X}_{iedc}^\top \boldsymbol{\phi}_1 + \boldsymbol{X}_d^\top \boldsymbol{\phi}_2 + \boldsymbol{X}_e^\top \boldsymbol{\phi}_3 + \varepsilon_{iedc}, \tag{28}$$

where, using NW's notation, i, e, d, and c indicate individual, ethnicity, district, and country, respectively. The outcome variable is trust_{iedc}, which is the level of trust an individual has towards their neighbors. We multiply the outcome variable by 1000 to avoid three leading zeros.

The coefficient of interest in (28) is β, which measures the extent to which historical slave exports of a given ethnicity affect trust levels for an individual of the same ethnicity today. On the right-hand side, $\boldsymbol{\alpha}_c$ is a vector of country-level fixed effects, $\boldsymbol{X}_{iedc}$ contains control variables such as age, gender, and education, $\boldsymbol{X}_d$ contains two district-level variables which may influence an ethnic group's current levels of trust, and $\boldsymbol{X}_e$ contains ethnicity-level variables to control for the degree of colonization and other historical differences. The trust variable comes from surveys for the Afrobarometer, conducted in 2005, which covered either 1200 or 2400 individuals in each of 17 countries. Survey respondents were asked to indicate the level of trust they had for their neighbors. For slave exports, NW use data from Nunn (2008) for the trans-Atlantic and Indian Ocean slave trades from 1400 to 1900. The final sample consists of $N = 20,027$ observations from 16 countries.

NW uses two-way clustered standard errors, where the clustering dimensions are geography at the district level and ethnicity. In the terminology of Abadie, Athey, Imbens, and Wooldridge (2017), NW takes a "model-based" approach rather than a "design-based" approach, implicitly treating (28) as a DGP that draws clusters at random from a meta-

Table 3: OLS estimates of the determinants of trust in neighbors

Dependent variable	slave exports		slave exports	
Trust of neighbors $\times 1000$	district		country	
$\hat{\beta}$	-0.6791		-0.6791	
$\hat{\boldsymbol{V}}_G$ (geo.) s.e. and $t(G-1)$ P value	0.0822	(.0000)	0.2051	(.0048)
$\hat{\boldsymbol{V}}_H$ (eth.) s.e. and $t(H-1)$ P value	0.1422	(.0000)	0.1422	(.0000)
$\hat{\boldsymbol{V}}_2$ (eth. & geo.) s.e. and $t(R-1)$ P value	0.1643	(.0001)	0.2496	(.0158)
$\hat{\boldsymbol{V}}_3$ (eth. & geo.) s.e. and $t(R-1)$ P value	0.1449	(.0000)	0.2078	(.0052)
	$t_{a,2}$	$t_{a,3}$	$t_{a,2}$	$t_{a,3}$
WCR_G (geography) bootstrap P values	.0000	.0006	.1190	.1394
WCR_H (ethnicity) bootstrap P values	.0020	.0019	.0115	.0741
WCR_I (ethnicity$\times$geography) bootstrap P values	.0000	.0002	.0120	.0681
WR (individual) bootstrap P values	.0000	.0005	.0084	.0246
Number of clusters, G (geography)	1257		16	
Number of clusters, H (ethnicity)	185		185	
Number of intersections, I (ethnicity$\times$geography)	3225		223	

Notes: This example is taken from Nunn and Wantchekon (2011, Table 1, column 1). All bootstrap P values are symmetric and based on the Rademacher distribution with $B = 9,999$. Stata `.do` files to replicate this table may be found at the authors' websites.

population. Because the independent variable of interest (exports) is invariant within ethnicities, it is impossible to use ethnicity fixed effects, even though it seems likely that social attitudes towards trust are correlated with ethnicity. Thus it surely makes sense to cluster by ethnicity (Moulton 1986).

The fact that many of the control variables are observed only at the district level provides justification for geographic clustering at that level. However, it seems plausible that the disturbances may be correlated at the country level in addition to the district level, for two reasons. First, the Afrobarometer survey frames some questions at the country level, and second, trust might well be influenced by country-level factors such as the rule of law or level of corruption. We therefore consider two levels of clustering (1257 districts or 16 countries) in the geographical dimension, G. The number of clusters in the ethnic clustering dimension, H, is always 185. There are either 3225 or 223 intersection-level clusters.

Table 3 reproduces and extends the results in NW's Table 1, column 1. NW uses three different variables for the key regressor. We focus on exports, but the results for exports/area and exports/(historical population) follow a broadly similar pattern. Table 3 presents the results from OLS estimation of (28). The first row of results presents the coefficient estimate. Following it, the top panel presents standard errors and P values clustered by three different one-way clustering variables, namely, geography at either the district or country levels, and ethnicity. All of these P values are extremely small.

The second panel of Table 3 presents two-way clustered standard errors and P values

based on the $t(R-1)$ distribution with clustering by both geography (either district or country) and ethnicity. As expected, the two-term CRVE standard errors are noticeably larger than the three-term ones, and therefore the P values as well. However, all the P values suggest that the null hypothesis can be rejected at the 2% level.

The third panel of Table 3 shows bootstrap P values for both $t_{a,2}$ and $t_{a,3}$. When clustering is by district in the geography dimension, all bootstrap tests are significant at the 1% level. However, when clustering is by country in that dimension, three of the ones based on $t_{a,3}$ are not significant at the 5% level. When the bootstrap clustering is also by country, which the experiments of Section 5 suggest may be the most reliable method, we do not reject even at the 10% level. Thus, based on the results for $t_{a,3}$, the evidence against the null hypothesis seems to be quite weak when geographic clustering is at the country level.

The bootstrap P values based on $t_{a,2}$ tell a different story. When we move from district-level to country-level clustering, they still increase, but the increases are far less dramatic, except for WCR_G. Moreover, again with that one exception, all the tests reject the null hypothesis at the 2% level. However, because bootstrap tests based on $t_{a,2}$ often over-reject severely in the experiments of Section 5, we are not inclined to believe these results.

7 Conclusion

We study variance estimation and bootstrap inference for regression models with two-way clustering. We consider two different cluster-robust variance estimators (CRVEs), one that involves three terms proposed in Cameron et al. (2011) and Thompson (2011), and one that involves two terms proposed in Davezies et al. (2018). In Section 3, we prove that t-tests based on both CRVE variants yield asymptotically valid inferences under precisely stated, but different, conditions. The two-term CRVE is consistent under less general conditions than the three-term one. The former is not consistent when the disturbances are independent or are clustered only at the level of the intersections of the two dimensions.

In Section 4, we propose several variants of the wild (cluster) bootstrap, each of which can be combined with either of the CRVEs. These appear to be the first such methods for least squares regression with multiway cluster-robust standard errors. The methods differ in the clustering imposed on the bootstrap disturbances, in the CRVE applied, and in using either restricted or unrestricted estimates. None of these bootstrap methods is capable of matching the two-dimensional nature of the clustered disturbances, and they do not all yield valid inferences in all cases. Nonetheless, we give precise conditions, that vary by bootstrap method, under which each method is valid or not valid.

In Section 5, we provide extensive simulation evidence. Using a one-way CRVE when there is actually two-way clustering can lead to extremely severe errors of inference, especially when the sample size is large. The conventional approach of comparing multiway cluster-robust t-statistics to quantiles from the t distribution can also lead to serious errors of inference, especially when the number of clusters in either dimension is small. Specifically, t-tests based on the three-term CRVE always seem to over-reject, while those based on the two-term CRVE may either under-reject or over-reject. In almost all the cases that we study, bootstrap methods based on t-statistics that use the three-term CRVE yield (much) more accurate inferences than the conventional approach of using the t distribution. In contrast,

bootstrapping t-statistics that use the two-term CRVE often yields inferences that are less accurate than comparing them to the t distribution.

Overall, the method of inference with the lowest error in rejection percentage throughout our extensive set of simulations is the restricted wild cluster bootstrap based on the three-term CRVE coupled with a bootstrap DGP that is clustered along the dimension with the smallest number of clusters. Such a DGP preserves the intra-cluster correlations for the dimension where the clusters are, on average, largest.

In Section 6, we illustrate several of our results using the data and one of the models of Nunn and Wantchekon (2011). We find that inferences can change substantially as the level of clustering in one of two dimensions changes. The P values, especially the bootstrap P values based on the three-term CRVE, become larger when the number of clusters in the geographical dimension is reduced, because clustering in that dimension is coarser. This is consistent with our simulation results, which suggest that it is particularly important to employ the wild cluster bootstrap when there are few clusters in either dimension.

Supplementary Appendix: Proofs of Main Results

A.1 Proof of Theorem 1

The result in (19) is an immediate consequence of Propositions 4.3 and 4.4 of Davezies et al. (2018), where (16) is assumed. Under (17), there is clustering only at the intersection level, and under (18) there is no clustering. Both of these are special cases of one-way clustering, so that (20) and (21) follow from Djogbenou et al. (2019) after noting that our assumptions imply that the cluster sizes N_{gh} are bounded almost surely.

A.2 Proof of Theorem 2

The results of the theorem follow directly from the definitions of $\hat{\boldsymbol{V}}_2$ and $\hat{\boldsymbol{V}}_3$ in (9) and (6), respectively, and application of Lemma A.1, which is proven in the next subsection. For example, under (16) it follows from this lemma that

$$R(\hat{\boldsymbol{V}}_2 - \hat{\boldsymbol{V}}_3) = R\hat{\boldsymbol{V}}_I = O_P\Big((GH)^{-1}R\Big) \xrightarrow{P} 0.$$

Lemma A.1. *Suppose Assumptions 1–6 are satisfied.*

(a) If (24) holds, so that the DGP is clustered along the first dimension, then

$$G\hat{\boldsymbol{V}}_G \xrightarrow{P} \boldsymbol{V}_G \quad \textit{and} \quad GH\hat{\boldsymbol{V}}_I \xrightarrow{P} \boldsymbol{V}_I.$$

(b) If (25) holds, so that the DGP is not clustered along the first dimension, but it is clustered along the second dimension, then

$$GH\boldsymbol{a}^\top\hat{\boldsymbol{V}}_G\boldsymbol{a} \xrightarrow{d} W_1^2 \quad \textit{and} \quad GH\,\hat{\boldsymbol{V}}_I \xrightarrow{P} \boldsymbol{V}_I,$$

where W_1^2 is a random variable satisfying $W_1^2 > 0$ almost surely.

(c) If (17) holds, so that the DGP is clustered by intersections, then

$$GH\,\hat{\boldsymbol{V}}_m \xrightarrow{P} \boldsymbol{V}_I \quad \textit{for } m \in \{G, H, I\}.$$

(d) If (18) holds, so that the DGP is not clustered, then

$$GH\hat{\boldsymbol{V}}_m \xrightarrow{P} \boldsymbol{V}_I \quad \text{for } m \in \{G, H, I\}.$$

A.3 Proof of Lemma A.1

Proof for cases (a) and (b): The results in case (a) and the second result in case (b) are given in Proposition 4.4 in Davezies et al. (2018). For the first result in case (b), we use the decomposition $\hat{\boldsymbol{u}}_{gh} = \boldsymbol{u}_{gh} - \boldsymbol{X}_{gh}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ such that $\sum_{h=1}^H \boldsymbol{X}_{gh}^\top \hat{\boldsymbol{u}}_{gh} = \sum_{h=1}^H \boldsymbol{X}_{gh}^\top \boldsymbol{u}_{gh} - (GH)^{-1} \sum_{h=1}^H \boldsymbol{X}_{gh}^\top \boldsymbol{X}_{gh} \boldsymbol{Q}^{-1} \boldsymbol{X}^\top \boldsymbol{u}$, where, under (25), $H^{-1} \sum_{h=1}^H \boldsymbol{X}_{gh}^\top \boldsymbol{X}_{gh} \xrightarrow{P} \boldsymbol{Q}_0$ and $\boldsymbol{Q} \xrightarrow{P} \boldsymbol{Q}_0$; see Assumption 2. Thus, for any arbitrary $\boldsymbol{\eta}$, we write

$$GH\boldsymbol{\eta}^\top \hat{\boldsymbol{\Gamma}}_H \boldsymbol{\eta} = \frac{1}{G} \sum_{g=1}^G \left(H^{-1/2} \sum_{h=1}^H \boldsymbol{\eta}^\top \boldsymbol{X}_{gh}^\top \hat{\boldsymbol{u}}_{gh} \right)^2 = \frac{1}{G} \sum_{g=1}^G \left(H^{-1/2} \sum_{h=1}^H z_{gh} \right)^2 + o_P(1),$$

where, for any (fixed) g, $z_{gh} = \boldsymbol{\eta}^\top \boldsymbol{X}_{gh}^\top \boldsymbol{u}_{gh} - G^{-1} \boldsymbol{\eta}^\top \boldsymbol{X}_h^\top \boldsymbol{u}_h$ is i.i.d. across h with mean zero and finite variance. For fixed G, it follows that the random vector $H^{-1/2} \sum_{h=1}^H (z_{1h}, \ldots, z_{Gh})^\top$ is asymptotically normal as $H \to \infty$ with mean zero and finite $G \times G$ variance matrix, say $\boldsymbol{J}_G$. Still for fixed G, it follows that $G^{-1} \sum_{g=1}^G (H^{-1/2} \sum_{h=1}^H z_{gh})^2 \xrightarrow{d} G^{-1} \sum_{m=1}^M \nu_m \|\boldsymbol{\mu}_m\|^2 Z_m^2$ as $H \to \infty$, where $(\nu_m, \boldsymbol{\mu}_m)$ denote the eigenvalues and eigenvectors of $\boldsymbol{J}_G$, $M \leq G$ is the number of non-zero eigenvalues, and Z_m denote i.i.d. standard normal random variables. Next, $G^{-1}\nu_m \|\mu_m\|^2 \to \omega_m^2 \in [0, \infty)$ for all $m \geq 1$, where $\omega_m > 0$ for at least one m. Hence, $G^{-1} \sum_{g=1}^G (H^{-1/2} \sum_{h=1}^H z_{gh})^2 \xrightarrow{d} \sum_{m=1}^\infty \omega_m^2 Z_m^2$, which is a (scaled) weighted sum of χ_1^2-distributions.

Proof for cases (c) and (d): Under (17), we can apply the results of Djogbenou et al. (2019), for the same reason as in the proof of (20), to conclude that each term in (7), multiplied by GH, converges in probability to $\boldsymbol{\Gamma}_I$ defined in (14). The convergence in probability of $\hat{\boldsymbol{V}}_2$ and $\hat{\boldsymbol{V}}_3$, normalized by GH, follows. Similarly, under (18), each term in (7), multiplied by GH, converges in probability to $\boldsymbol{\Gamma}_I$.

A.4 Proof of Theorem 3

To prove Theorem 3 we first present the bootstrap equivalents of Theorems 1 and 2. These are given in Theorems A.1 and A.2, the proofs of which are in the next subsections.

Theorem A.1. *Suppose Assumptions 1–7 are satisfied and that H_0 is true. Let $m \in \{G, H, I, NC\}$ denote bootstrap clustering by the first dimension, the second dimension, intersections, and individual observations, respectively; c.f. step 3(a). Then it holds that*

$$(\boldsymbol{a}^\top \ddot{\boldsymbol{V}}_m \boldsymbol{a})^{-1/2} \boldsymbol{a}^\top (\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}}) \xrightarrow{d^*} Z, \text{ in probability,}$$

where $Z \sim \mathrm{N}(0, 1)$.

Theorem A.2. *Suppose Assumptions 1–7 are satisfied and that H_0 is true.*

(i) Suppose the bootstrap DGP in step 3(a) is clustered along the first (G) dimension (results for bootstrap clustering along the second dimension are symmetric).

(a) If (24) holds, so that the DGP is clustered along the first dimension, then

$$G(\hat{\boldsymbol{V}}_2^* - \ddot{\boldsymbol{V}}_G) \xrightarrow{P^*} \mathbf{0} \quad \textit{and} \quad G(\hat{\boldsymbol{V}}_3^* - \ddot{\boldsymbol{V}}_G) \xrightarrow{P^*} \mathbf{0}, \ \textit{in probability.}$$

(b) If (25) holds, so that the DGP is not clustered along the first dimension, but it is clustered along the second dimension, then

$$GH\boldsymbol{a}^\top(\hat{\boldsymbol{V}}_2^* - \ddot{\boldsymbol{V}}_G - \ddot{\boldsymbol{V}}_I)\boldsymbol{a} \xrightarrow{d^*} W_0 \quad \textit{and} \quad GH\boldsymbol{a}^\top(\hat{\boldsymbol{V}}_3^* - \ddot{\boldsymbol{V}}_G)\boldsymbol{a} \xrightarrow{d^*} W_0, \ \textit{in probability,}$$

where W_0 is a zero mean random variable that is independent of Z in Theorem A.1.

(c) If (17) holds, so that the DGP is clustered by intersections, then

$$GH(\hat{\boldsymbol{V}}_2^* - \ddot{\boldsymbol{V}}_G - \ddot{\boldsymbol{V}}_I) \xrightarrow{P^*} \mathbf{0} \quad \textit{and} \quad GH(\hat{\boldsymbol{V}}_3^* - \ddot{\boldsymbol{V}}_G) \xrightarrow{P^*} \mathbf{0}, \ \textit{in probability.}$$

(d) If (18) holds, so that the DGP is not clustered, then

$$GH(\hat{\boldsymbol{V}}_2^* - \ddot{\boldsymbol{V}}_G - \ddot{\boldsymbol{V}}_I) \xrightarrow{P^*} \mathbf{0} \quad \textit{and} \quad GH(\hat{\boldsymbol{V}}_3^* - \ddot{\boldsymbol{V}}_G) \xrightarrow{P^*} \mathbf{0}, \ \textit{in probability.}$$

(ii) If the bootstrap DGP in step 3(a) is clustered by intersections, then

$$GH(\hat{\boldsymbol{V}}_2^* - 2\ddot{\boldsymbol{V}}_I) \xrightarrow{P^*} \mathbf{0} \quad \textit{and} \quad GH(\hat{\boldsymbol{V}}_3^* - \ddot{\boldsymbol{V}}_I) \xrightarrow{P^*} \mathbf{0}, \ \textit{in probability.}$$

(iii) If the bootstrap DGP in step 3(a) is the WB, then

$$GH(\hat{\boldsymbol{V}}_2^* - 2\ddot{\boldsymbol{V}}_I) \xrightarrow{P^*} \mathbf{0} \quad \textit{and} \quad GH(\hat{\boldsymbol{V}}_3^* - \ddot{\boldsymbol{V}}_I) \xrightarrow{P^*} \mathbf{0}, \ \textit{in probability.}$$

Let $m \in \{G, H, I, NC\}$ denote bootstrap clustering by the first dimension, the second dimension, intersections, and individual observations, respectively; c.f. step 3(a). We then decompose the bootstrap t-statistic as

$$t_{a,j}^* = \frac{\boldsymbol{a}^\top(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})}{(\boldsymbol{a}^\top \hat{\boldsymbol{V}}_j^* \boldsymbol{a})^{1/2}} = \left(\frac{\boldsymbol{a}^\top \ddot{\boldsymbol{V}}_m \boldsymbol{a}}{\boldsymbol{a}^\top \hat{\boldsymbol{V}}_j^* \boldsymbol{a}}\right)^{1/2} \frac{\boldsymbol{a}^\top(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})}{(\boldsymbol{a}^\top \ddot{\boldsymbol{V}}_m \boldsymbol{a})^{1/2}} = (A_{m,j}^*)^{1/2} B_m^*, \tag{A.1}$$

say. From Theorem A.1 we find that $B_m^* \xrightarrow{d^*} Z \sim \mathrm{N}(0,1)$, in probability, for all m.

For the first term on the right-hand side of (A.1), the result follows by direct application of Lemma A.1 and Theorem A.2. In particular, for cases (i)(a),(c), and (ii), $A_{m,j}^* \xrightarrow{P^*} q$, in probability, where $q = 1/2$ or $q = 1$ is the variance of the limit distribution of $t_{a,j}^*$. For case (i)(b), we write $A_{m,j}^* = 1 - (\hat{\boldsymbol{V}}_j^* - \ddot{\boldsymbol{V}}_G)/\hat{\boldsymbol{V}}_j^*$ and apply Lemma A.1 and Theorem A.2. Note that, because H_0 is true, the results of Lemma A.1 also apply to the variance estimators imposing the null, i.e. all $\hat{\boldsymbol{V}}$ in Lemma A.1 can be replaced by $\ddot{\boldsymbol{V}}$. The random variable W_1^2 may then be different, but since the explicit form of W_1^2 is not needed, that is not an issue. Finally, Z and W_0 are generated by the bootstrap measure and are both therefore independent of W_1^2.

A.5 Proof of Theorem A.1

We give the proof only for the case where the bootstrap is clustered along the first dimension; that is, $\boldsymbol{u}_g^* = \ddot{\boldsymbol{u}}_g v_g^*$. The proofs for the other cases are entirely analogous. First note that

$$(\boldsymbol{a}^\top \ddot{\boldsymbol{V}}_G \boldsymbol{a})^{-1/2} \boldsymbol{a}^\top (\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}}) = \sum_{g=1}^G z_g^*, \quad z_g^* = (\boldsymbol{a}^\top \ddot{\boldsymbol{V}}_G \boldsymbol{a})^{-1/2} \boldsymbol{a}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g v_g^*.$$

Because v_g^* is independent across g with mean zero and variance one, it follows that z_g^* is independent across g with $\mathrm{E}^*(z_g^*) = 0$ and $\mathrm{Var}^*(\sum_{g=1}^G z_g^*) = \sum_{g=1}^G \mathrm{Var}^*(z_g^*) = 1$. The Lyapunov condition is satisfied (with P-probability converging to one) because

$$\begin{aligned}\sum_{g=1}^G \mathrm{E}^*|z_g^*|^4 &\leq \mathrm{E}(v^{*4})(\boldsymbol{a}^\top \ddot{\boldsymbol{V}}_G \boldsymbol{a})^{-2} \|\boldsymbol{Q}^{-1}\|^4 \sum_{g=1}^G \Big\| \frac{1}{GH} \boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g \Big\|^4 \\ &= O_P(1)(\boldsymbol{a}^\top \ddot{\boldsymbol{V}}_G \boldsymbol{a})^{-2} \frac{1}{(GH)^4} \sum_{g=1}^G \Big\| \boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g \Big\|^4 \xrightarrow{P} 0,\end{aligned}$$

regardless of the clustering structure in the DGP. To see this, suppose first that (24) holds, in which case the DGP is clustered along the first (G) dimension. Then $\boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g = \sum_{h=1}^H \boldsymbol{X}_{gh}^\top \ddot{\boldsymbol{u}}_{gh}$ is of order $O_P(H)$ and $G\ddot{\boldsymbol{V}}_G \xrightarrow{P} \boldsymbol{V}_G > 0$; see Davezies et al. (2018) and Lemma A.1. However, if the DGP is not clustered along the first dimension (under (17), (18), or (25)), then $\boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g = O_P(H^{1/2})$ and $\ddot{\boldsymbol{V}}_G^{-1} = O_P(GH)$; see also Djogbenou et al. (2019) and Lemma A.1. In either case, $\sum_{g=1}^G \mathrm{E}^*|z_g^*|^4 = O_P(G^{-1})$.

A.6 Proof of Theorem A.2

In all cases, the factors $\boldsymbol{Q}^{-1}$ in the definitions of $\hat{\boldsymbol{V}}_j^*$ are functions only of the original data and satisfy $\boldsymbol{Q} \xrightarrow{P} \boldsymbol{Q}_0 > 0$. Hence, these factors have no impact on the proofs. We therefore prove most results for the corresponding $\hat{\boldsymbol{\Gamma}}_m^*$; see (7) and (22). Specifically, we prove the following lemma, which suffices for the theorem.

Lemma A.2. *Suppose Assumptions 1–7 are satisfied and that H_0 is true.*

(i) Suppose the bootstrap DGP in step 3(a) is clustered along the first (G) dimension (results for bootstrap clustering along the second dimension are symmetric).

(a) If (24) holds, so that the DGP is clustered along the first dimension, then

$$G(\hat{\boldsymbol{\Gamma}}_G^* - \ddot{\boldsymbol{\Gamma}}_G) \xrightarrow{P^*} \boldsymbol{0} \quad \text{and} \quad GH(\hat{\boldsymbol{\Gamma}}_I^* - \ddot{\boldsymbol{\Gamma}}_I) \xrightarrow{P^*} \boldsymbol{0}, \text{ in probability,}$$
$$\mathrm{Var}^*(GH(\hat{\boldsymbol{\Gamma}}_H^* - \ddot{\boldsymbol{\Gamma}}_I)) = O_P(1).$$

(b) If (25) holds, so that the DGP is not clustered along the first dimension, but it is clustered along the second dimension, then

$$GH(\hat{\boldsymbol{\Gamma}}_G^* - \ddot{\boldsymbol{\Gamma}}_G) \xrightarrow{P^*} \boldsymbol{0} \quad \text{and} \quad GH(\hat{\boldsymbol{\Gamma}}_I^* - \ddot{\boldsymbol{\Gamma}}_I) \xrightarrow{P^*} \boldsymbol{0}, \text{ in probability,}$$
$$GH\boldsymbol{a}^\top(\hat{\boldsymbol{V}}_H^* - \ddot{\boldsymbol{V}}_I)\boldsymbol{a} \xrightarrow{d^*} W_0, \text{ in probability,}$$

where W_0 is a zero mean random variable that is independent of Z in Theorem A.1.

(c) If (17) holds, so that the DGP is clustered by intersections, then

$$GH(\hat{\mathbf{\Gamma}}^*_G - \ddot{\mathbf{\Gamma}}_G) \xrightarrow{P^*} \mathbf{0} \quad \textit{and} \quad GH(\hat{\mathbf{\Gamma}}^*_m - \ddot{\mathbf{\Gamma}}_I) \xrightarrow{P^*} \mathbf{0}, \ \textit{in probability, for } m \in \{H, I\}.$$

(d) If (18) holds, so that the DGP is not clustered, then

$$GH(\hat{\mathbf{\Gamma}}^*_G - \ddot{\mathbf{\Gamma}}_G) \xrightarrow{P^*} \mathbf{0} \quad \textit{and} \quad GH(\hat{\mathbf{\Gamma}}^*_m - \ddot{\mathbf{\Gamma}}_I) \xrightarrow{P^*} \mathbf{0}, \ \textit{in probability, for } m \in \{H, I\}.$$

(ii) If the bootstrap DGP in step 3(a) is clustered by intersections, then

$$GH(\hat{\mathbf{\Gamma}}^*_m - \ddot{\mathbf{\Gamma}}_I) \xrightarrow{P^*} \mathbf{0}, \ \textit{in probability, for } m \in \{G, H, I\}.$$

(iii) If the bootstrap DGP in step 3(a) is the WB, then

$$GH(\hat{\mathbf{\Gamma}}^*_m - \ddot{\mathbf{\Gamma}}_I) \xrightarrow{P^*} \mathbf{0}, \ \textit{in probability, for } m \in \{G, H, I\}.$$

A.7 Proof of Lemma A.2

We prove convergence in mean square. That is, we show that the second moment (conditional on the sample) converges to zero (in P-probability). Let $\boldsymbol{\eta}$ be an arbitrary conforming vector.

Proof for case (i): First, using the decomposition $\hat{\boldsymbol{u}}^*_g = \ddot{\boldsymbol{u}}_g v^*_g - \boldsymbol{X}_g(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})$, we find that

$$\boldsymbol{\eta}^\top(\hat{\mathbf{\Gamma}}^*_G - \ddot{\mathbf{\Gamma}}_G)\boldsymbol{\eta} = \frac{1}{(GH)^2}\sum_{g=1}^{G} \boldsymbol{\eta}^\top \boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g \ddot{\boldsymbol{u}}_g^\top \boldsymbol{X}_g \boldsymbol{\eta}(v^{*2}_g - 1) \tag{A.2}$$

$$- \frac{2}{(GH)^2}\sum_{g=1}^{G} \boldsymbol{\eta}^\top \boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g (\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})^\top \boldsymbol{X}_g^\top \boldsymbol{X}_g \boldsymbol{\eta} v^*_g \tag{A.3}$$

$$+ \frac{1}{(GH)^2}\sum_{g=1}^{G} \left(\boldsymbol{\eta}^\top \boldsymbol{X}_g^\top \boldsymbol{X}_g (\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})\right)^2. \tag{A.4}$$

Because $(v^{*2}_g - 1)$ is independent and identically distributed across g with mean zero and finite variance, the conditional second moment of (A.2) is

$$\mathrm{E}^*((\text{A.2})^2) = \frac{1}{(GH)^4}\mathrm{E}^*\left((v^{*2} - 1)^2\right)\sum_{g=1}^{G}(\boldsymbol{\eta}^\top \boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g \ddot{\boldsymbol{u}}_g^\top \boldsymbol{X}_g \boldsymbol{\eta})^2.$$

Under (24), where the DGP is clustered along the first (G) dimension, $\boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g = \sum_{h=1}^{H} \boldsymbol{X}_{gh}^\top \ddot{\boldsymbol{u}}_{gh}$ is of order $O_P(H)$. However, if the DGP is not clustered along the first dimension (under (17), (18), or (25)), then $\boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g = O_P(H^{1/2})$. This shows the results for (A.2) for case (i). The conditional second moment of (A.3) is

$$\mathrm{E}^*((\text{A.3})^2) = \frac{4}{(GH)^4}\mathrm{E}^*\Bigg(\sum_{g_1,g_2=1}^{G} \boldsymbol{\eta}^\top \boldsymbol{X}_{g_1}^\top \ddot{\boldsymbol{u}}_{g_1} \ddot{\boldsymbol{u}}_{g_2}^\top \boldsymbol{X}_{g_2} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}_{g_1}^\top \boldsymbol{X}_{g_1} \boldsymbol{\eta} v^*_{g_1} v^*_{g_2}\Bigg)^2,$$

where we note that expanding the square results in four summations, but two of these are eliminated because v^*_g is independent across g, so that the summation indexes must be equal

in pairs. Using this together with the aforementioned orders of magnitude of $\boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g$ and the facts that $(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = O_P((GH)^{-1})$ and $\boldsymbol{X}_g^\top \boldsymbol{X}_g = O_P(H)$ (Davezies et al. 2018) yields the desired results for (A.3) for case (i). Finally, (A.4) is a non-negative random variable, and noting that $\mathrm{Var}^*(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}}) = \ddot{\boldsymbol{V}}_G$, its conditional mean is

$$\mathrm{E}^*((\text{A.4})) = \frac{1}{(GH)^2} \sum_{g=1}^{G} \boldsymbol{\eta}^\top \boldsymbol{X}_g^\top \boldsymbol{X}_g \ddot{\boldsymbol{V}}_G \boldsymbol{X}_g^\top \boldsymbol{X}_g \boldsymbol{\eta} = O_P(G^{-1} \|\ddot{\boldsymbol{V}}_G\|).$$

The results for (A.4) for case (i) then follow by application of Lemma A.1.

Next, we find that

$$GH\boldsymbol{\eta}^\top(\hat{\boldsymbol{\Gamma}}_I^* - \ddot{\boldsymbol{\Gamma}}_I)\boldsymbol{\eta} = \frac{1}{GH} \sum_{g=1}^{G} \sum_{h=1}^{H} \boldsymbol{\eta}^\top \boldsymbol{X}_{gh}^\top \ddot{\boldsymbol{u}}_{gh} \ddot{\boldsymbol{u}}_{gh}^\top \boldsymbol{X}_{gh} \boldsymbol{\eta} (v_g^{*2} - 1) \tag{A.5}$$

$$- \frac{2}{GH} \sum_{g=1}^{G} \sum_{h=1}^{H} \boldsymbol{\eta}^\top \boldsymbol{X}_{gh}^\top \ddot{\boldsymbol{u}}_{gh} (\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})^\top \boldsymbol{X}_{gh}^\top \boldsymbol{X}_{gh} \boldsymbol{\eta} v_g^* \tag{A.6}$$

$$+ \frac{1}{GH} \sum_{g=1}^{G} \sum_{h=1}^{H} \left(\boldsymbol{\eta}^\top \boldsymbol{X}_{gh}^\top \boldsymbol{X}_{gh} (\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}}) \right)^2. \tag{A.7}$$

The proofs for each of the terms (A.5)–(A.7) are nearly identical to those for (A.2)–(A.4), and they are therefore omitted.

For $\hat{\boldsymbol{\Gamma}}_H^*$ we find that

$$\begin{aligned} GH\boldsymbol{\eta}^\top \hat{\boldsymbol{\Gamma}}_H^* \boldsymbol{\eta} &= \frac{1}{GH} \sum_{h=1}^{H} (\boldsymbol{\eta}^\top \boldsymbol{X}_h^\top \hat{\boldsymbol{u}}_h^*)^2 = \frac{1}{GH} \sum_{h=1}^{H} \left(\boldsymbol{\eta}^\top \boldsymbol{X}_h^\top \boldsymbol{u}_h^* - \boldsymbol{X}_h (\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}}) \right)^2 \\ &= \frac{1}{GH} \sum_{h=1}^{H} (\boldsymbol{\eta}^\top \boldsymbol{X}_h^\top \boldsymbol{u}_h^*)^2 + A^*, \end{aligned} \tag{A.8}$$

where $\mathrm{E}^*|A^*| = O_P(G\|\ddot{\boldsymbol{V}}_G\|)$, in probability, by application of Lemma A.1 and the Cauchy-Schwarz inequality, because

$$\mathrm{E}^*\left(\frac{1}{GH} \sum_{h=1}^{H} (\boldsymbol{\eta}^\top \boldsymbol{X}_h^\top \boldsymbol{X}_h (\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}}) \right)^2 = \frac{1}{GH} \sum_{h=1}^{H} \boldsymbol{\eta}^\top \boldsymbol{X}_h^\top \boldsymbol{X}_h \ddot{\boldsymbol{V}}_G \boldsymbol{X}_h^\top \boldsymbol{X}_h \boldsymbol{\eta} = O_P(G\|\ddot{\boldsymbol{V}}_G\|).$$

This shows that A^* is of the required order of magnitude in part (a) and is negligible in parts (b)–(d). Noting that $\boldsymbol{X}_h^\top \boldsymbol{u}_h^* = \sum_{g=1}^{G} \boldsymbol{X}_{gh}^\top \ddot{\boldsymbol{u}}_{gh} v_g^*$, the main term in (A.8) satisfies

$$\frac{1}{GH} \sum_{h=1}^{H} (\boldsymbol{\eta}^\top \boldsymbol{X}_h^\top \boldsymbol{u}_h^*)^2 - GH\boldsymbol{\eta}^\top \ddot{\boldsymbol{\Gamma}}_I \boldsymbol{\eta} = \frac{1}{GH} \sum_{h=1}^{H} \sum_{g=1}^{G} \boldsymbol{\eta}^\top \boldsymbol{X}_{gh}^\top \ddot{\boldsymbol{u}}_{gh} \ddot{\boldsymbol{u}}_{gh}^\top \boldsymbol{X}_{gh} \boldsymbol{\eta} (v_g^{*2} - 1) \tag{A.9}$$

$$+ \frac{1}{GH} \sum_{h=1}^{H} \sum_{g_1 \neq g_2}^{G} \boldsymbol{\eta}^\top \boldsymbol{X}_{g_1 h}^\top \ddot{\boldsymbol{u}}_{g_1 h} \ddot{\boldsymbol{u}}_{g_2 h}^\top \boldsymbol{X}_{g_2 h} \boldsymbol{\eta} v_{g_1}^* v_{g_2}^*. \tag{A.10}$$

By independence of $(v_g^{*2} - 1)$ across g, it is easily seen that $\mathrm{E}^*((\text{A.9})^2) = O_P(G^{-1})$, showing the results for (A.9) for case (i). Similarly,

$$\mathrm{E}^*((\text{A.10})^2) = \frac{2}{(GH)^2} \sum_{g_1 \neq g_2}^{G} \left(\sum_{h=1}^{H} \boldsymbol{\eta}^\top \boldsymbol{X}_{g_1 h}^\top \ddot{\boldsymbol{u}}_{g_1 h} \ddot{\boldsymbol{u}}_{g_2 h}^\top \boldsymbol{X}_{g_2 h} \boldsymbol{\eta} \right)^2,$$

which is $O_P(1)$ in part (a) and $o_P(1)$ in parts (c) and (d), showing the results for (A.10) for those parts. Thus, only part (b) remains for (A.10). For any fixed h, as $G \to \infty$,

$$\frac{1}{G^{1/2}}\boldsymbol{\eta}^\top \boldsymbol{X}_h^\top \boldsymbol{u}_h^* = \frac{1}{G^{1/2}}\sum_{g=1}^{G} \boldsymbol{\eta}^\top \boldsymbol{X}_{gh}^\top \ddot{\boldsymbol{u}}_{gh} v_g^* \xrightarrow{d^*} \mathrm{N}\Big(0, \operatorname*{plim}_{G\to\infty}\frac{1}{G}\sum_{g=1}^{G}(\boldsymbol{\eta}^\top \boldsymbol{X}_{gh}^\top \ddot{\boldsymbol{u}}_{gh})^2\Big), \tag{A.11}$$

in probability. Moreover, for fixed $h_1 \neq h_2$, as $G \to \infty$,

$$\mathrm{E}^*\Big(\frac{1}{G}\boldsymbol{\eta}^\top \boldsymbol{X}_{h_1}^\top \boldsymbol{u}_{h_1}^* \boldsymbol{\eta}^\top \boldsymbol{X}_{h_2}^\top \boldsymbol{u}_{h_2}^*\Big) = \frac{1}{G}\sum_{g=1}^{G} \boldsymbol{\eta}^\top \boldsymbol{X}_{gh_1}^\top \ddot{\boldsymbol{u}}_{gh_1} \ddot{\boldsymbol{u}}_{gh_2}^\top \boldsymbol{X}_{gh_2}\boldsymbol{\eta} \xrightarrow{P} 0. \tag{A.12}$$

It follows from (A.11), (A.12), and the continuous mapping theorem that, for fixed H, as $G \to \infty$,

$$\frac{1}{GH}\sum_{h=1}^{H}(\boldsymbol{\eta}^\top \boldsymbol{X}_h^\top \boldsymbol{u}_h^*)^2 \xrightarrow{d^*} \frac{1}{H}\sum_{h=1}^{H}\Big(\operatorname*{plim}_{G\to\infty}\frac{1}{G}\sum_{g=1}^{G}(\boldsymbol{\eta}^\top \boldsymbol{X}_{gh}^\top \ddot{\boldsymbol{u}}_{gh})^2\Big) Z_h^2, \text{ in probability,}$$

where $Z_h \sim i.i.d.\mathrm{N}(0,1)$ for $h = 1, \ldots, H$. Because $(GH)^{-1}\sum_{g=1}^{G}\sum_{h=1}^{H}(\boldsymbol{\eta}^\top \boldsymbol{X}_{gh}^\top \ddot{\boldsymbol{u}}_{gh})^2 \xrightarrow{P} \boldsymbol{\eta}^\top \boldsymbol{\Gamma}_I \boldsymbol{\eta} < \infty$ by Lemma A.1, it follows that

$$\frac{1}{GH}\sum_{h=1}^{H}(\boldsymbol{\eta}^\top \boldsymbol{X}_h^\top \boldsymbol{u}_h^*)^2 \xrightarrow{d^*} \sum_{m=1}^{\infty} \upsilon_m^2 Z_m^2, \text{ in probability,} \tag{A.13}$$

where $Z_m \sim i.i.d.\mathrm{N}(0,1)$ for $m = 1, 2, \ldots$. The right-hand side of (A.13) is a weighted sum of χ_1^2-distributions, where the weights satisfy $\sum_{m=1}^{\infty}\upsilon_m^2 = \boldsymbol{\eta}^\top \boldsymbol{\Gamma}_I \boldsymbol{\eta}$. Hence, using $\boldsymbol{Q} \xrightarrow{P} \boldsymbol{Q}_0$ and combining (A.8), (A.9), (A.10), and (A.13), we find for part (b) that

$$GH\boldsymbol{a}^\top(\hat{\boldsymbol{V}}_H^* - \ddot{\boldsymbol{V}}_I)\boldsymbol{a} \xrightarrow{d^*} \sum_{m=1}^{\infty}\tau_m^2(Z_m^2 - 1) = W_0, \text{ in probability,}$$

where the weights τ_m are derived from υ_m by setting $\boldsymbol{\eta} = \boldsymbol{Q}_0^{-1}\boldsymbol{a}$ in the latter and the τ_m thus satisfy $\sum_{m=1}^{\infty}\tau_m^2 = \boldsymbol{a}^\top \boldsymbol{V}_I \boldsymbol{a}$. Finally, W_0 is independent of Z because

$$\mathrm{E}^*\Big((\boldsymbol{a}^\top \ddot{\boldsymbol{V}}_G \boldsymbol{a})^{-1/2}\boldsymbol{a}^\top \boldsymbol{Q}^{-1}\frac{1}{GH}\sum_{g_1=1}^{G}\boldsymbol{X}_{g_1}^\top \ddot{\boldsymbol{u}}_{g_1} v_{g_1}^* \frac{1}{G}\Big(\sum_{g_2=1}^{G}\boldsymbol{\eta}^\top \boldsymbol{X}_{g_2 h}^\top \ddot{\boldsymbol{u}}_{g_2 h} v_{g_2}^*\Big)^2\Big) = O_P((GH)^{-1/2})$$

using Lemma A.1, independence of v_g^* across g (to eliminate the summation over g_2), and the fact that $\boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g = O_P(H^{1/2})$ under (25).

Proof for case (ii): First, we find that

$$GH\boldsymbol{\eta}^\top(\hat{\boldsymbol{\Gamma}}_G^* - \ddot{\boldsymbol{\Gamma}}_I)\boldsymbol{\eta} = \frac{1}{GH}\sum_{g=1}^{G}\sum_{h=1}^{H}\boldsymbol{\eta}^\top \boldsymbol{X}_{gh}^\top \ddot{\boldsymbol{u}}_{gh}\ddot{\boldsymbol{u}}_{gh}^\top \boldsymbol{X}_{gh}\boldsymbol{\eta}(v_{gh}^{*2} - 1) \tag{A.14}$$

$$+ \frac{1}{GH}\sum_{g=1}^{G}\sum_{h_1\neq h_2}^{H}\boldsymbol{\eta}^\top \boldsymbol{X}_{gh_1}^\top \ddot{\boldsymbol{u}}_{gh_1}\ddot{\boldsymbol{u}}_{gh_2}^\top \boldsymbol{X}_{gh_2}\boldsymbol{\eta} v_{gh_1}^* v_{gh_2}^* \tag{A.15}$$

$$- \frac{2}{GH}\sum_{g=1}^{G}\sum_{h_1,h_2=1}^{H}\boldsymbol{\eta}^\top \boldsymbol{X}_{gh_1}^\top \ddot{\boldsymbol{u}}_{gh_1}(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})^\top \boldsymbol{X}_{gh_2}^\top \boldsymbol{X}_{gh_2}\boldsymbol{\eta} v_{gh_1}^* v_{gh_2}^* \tag{A.16}$$

$$+ \frac{1}{GH}\sum_{g=1}^{G}\sum_{h_1,h_2=1}^{H}\boldsymbol{\eta}^\top \boldsymbol{X}_{gh_1}^\top \boldsymbol{X}_{gh_1}(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})^\top \boldsymbol{X}_{gh_2}^\top \boldsymbol{X}_{gh_2}\boldsymbol{\eta} v_{gh_1}^* v_{gh_2}^*. \tag{A.17}$$

The proofs for (A.14), (A.16), and (A.17) are nearly identical to those for (A.2)–(A.4), and are therefore omitted. For (A.15) we find

$$\mathrm{E}^*((\text{A.15})^2) = \frac{1}{(GH)^2} \sum_{g_1,g_2}^{G} \sum_{h_1 \neq h_1'}^{H} \sum_{h_2 \neq h_2'}^{H} t(g_1, g_2, h_1, h_1', h_2, h_2') \mathrm{E}^*(v^*_{g_1 h_1} v^*_{g_1 h_1'} v^*_{g_2 h_2} v^*_{g_2 h_2'}), \quad \text{(A.18)}$$

where $t(g_1, g_2, h_1, h_1', h_2, h_2') = \boldsymbol{\eta}^\top \boldsymbol{X}^\top_{g_1 h_1} \ddot{\boldsymbol{u}}_{g_1 h_1} \ddot{\boldsymbol{u}}^\top_{g_1 h_1'} \boldsymbol{X}_{g_1 h_1'} \boldsymbol{\eta} \boldsymbol{\eta}^\top \boldsymbol{X}^\top_{g_2 h_2} \ddot{\boldsymbol{u}}_{g_2 h_2} \ddot{\boldsymbol{u}}^\top_{g_2 h_2'} \boldsymbol{X}_{g_2 h_2'} \boldsymbol{\eta}$ is a function only of the original data and is $O_P(1)$. By independence of v^*_{gh} across both g and h, the right-hand side of (A.18) is non-zero only if $g_1 = g_2$ and either $h_1 = h_2, h_1' = h_2'$ or $h_1 = h_2', h_1' = h_2$. In either situation, one summation over g and two summations over h are eliminated, so that (A.18) is at most $O_P(G^{-1})$, which proves the result for $\hat{\boldsymbol{\Gamma}}^*_G$.

The proof for $\hat{\boldsymbol{\Gamma}}^*_H$ is identical to that for $\hat{\boldsymbol{\Gamma}}^*_G$ after interchanging the g and h subscripts throughout. Finally, $GH\boldsymbol{\eta}^\top(\hat{\boldsymbol{\Gamma}}^*_I - \ddot{\boldsymbol{\Gamma}}_I)\boldsymbol{\eta}$ is equal to the sum of (A.14), (A.16), and (A.17), with $h_1 = h_2$ in the latter two, so we have already proven the required result for this term.

Proof for case (iii): The proofs for case (iii) are nearly identical to those for case (ii) and are therefore omitted.

Acknowledgements

We are grateful to an associate editor, two referees, Brendan Beare, Colin Cameron, Russell Davidson, Silvia Gonçalves, SeoJeong (Jay) Lee, and Konrad Menzel for helpful comments. We also thank participants at the 2017 and 2019 Canadian Economics Association Annual Meetings, the 2017 Canadian Econometric Study Group Meeting, the 2017 Southern Economics Association Meeting, the 2018 Society of Labor Economists Meeting, the 2018 North American Summer Meeting of the Econometric Society, the 2018 International Association for Applied Econometrics Annual Conference, the 2018 Joint Statistical Meetings, the CIREQ Conference on Recent Advances in Bootstrap Methods (May 2019), and the 2019 Annual Meeting of the Statistical Society of Canada. In addition, we thank those who attended presentations at the Canada Mortgage and Housing Corporation, the Vancouver School of Economics, the Montreal Econometric Workshop (at McGill University), the University of Melbourne, the University of New South Wales, and the University of Sydney. We thank Scott McNeil and Christopher Cheng for research assistance. MacKinnon and Webb thank the Social Sciences and Humanities Research Council of Canada (SSHRC) for financial support. Nielsen thanks the Canada Research Chairs program, the SSHRC, and the Center for Research in Econometric Analysis of Time Series (CREATES, funded by the Danish National Research Foundation, DNRF78) for financial support. Some of the computations were performed at the Centre for Advanced Computing at Queen's University.

References

Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge (2017). When should you adjust standard errors for clustering? NBER Working Papers 24003.

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

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

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

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

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

Davezies, L., X. D'Haultfœuille, and Y. Guyonvarch (2018). Asymptotic results under multiway clustering. ArXiv e-prints, CREST.

Davidson, R. and E. Flachaire (2008). The wild bootstrap, tamed at last. *Journal of Econometrics 146*, 162–169.

Davidson, R. and J. G. MacKinnon (1999). The size distortion of bootstrap tests. *Econometric Theory 15*, 361–376.

Davidson, R. and J. G. MacKinnon (2004). *Econometric Theory and Methods*. New York: Oxford University Press.

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.

Gonçalves, S. and T. J. Vogelsang (2011). Block bootstrap HAC robust tests: The sophistication of the naive bootstrap. *Econometric Theory 27*, 745–791.

Hansen, B. E. (1999). The grid bootstrap and the autoregressive model. *Review of Economics and Statistics 81*, 594–607.

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

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

Ibragimov, R. and U. K. Müller (2016). Inference with few heterogeneous clusters. *Review of Economics and Statistics 98*, 83–96.

Imbens, G. W. and M. Kolesár (2016). Robust standard errors in small samples: Some practical advice. *Review of Economics and Statistics 98*, 701–712.

MacKinnon, J. G. (2015). Wild cluster bootstrap confidence intervals. *L'Actualité Économique 91*, 11–33.

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

MacKinnon, J. G. and M. D. Webb (2018). The wild bootstrap for few (treated) clusters.

Econometrics Journal 21, 114–135.

MacKinnon, J. G. and H. White (1985). Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics 29*, 305–325.

Menzel, K. (2018). Bootstrap with cluster-dependence in two or more dimensions. ArXiv e-prints, New York University.

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

Nunn, N. (2008). The long-term effects of Africa's slave trades. *Quarterly Journal of Economics 123*, 139–176.

Nunn, N. and L. Wantchekon (2011). The slave trade and the origins of mistrust in Africa. *American Economic Review 101*, 3221–3252.

Pustejovsky, J. E. and E. Tipton (2018). Small sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics 36*, 672–683.

Roodman, D., J. G. MacKinnon, M. Ø. Nielsen, and M. D. Webb (2019). Fast and wild: Bootstrap inference in Stata using boottest. *Stata Journal 19*, 4–60.

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

Webb, M. D. (2014). Reworking wild bootstrap based inference for clustered errors. QED Working Paper 1315, Queen's University.

White, H. (1984). *Asymptotic Theory for Econometricians*. San Diego: Academic Press.