Queen's Economics Department Working Paper No. 1386

Bootstrap and Asymptotic Inference with Multiway Clustering

James G. MacKinnon
Queen's University

Morten Ørregaard Nielsen
Queen's University

Matthew D. Webb
Carleton University

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

8-2017

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

August 18, 2017

Abstract

We study a cluster-robust variance estimator (CRVE) for regression models with clustering in two dimensions that was proposed in Cameron, Gelbach, and Miller (2011). We prove that this CRVE is consistent and yields valid inferences under precisely stated assumptions about moments and cluster sizes. We then propose several wild bootstrap procedures and prove that they are asymptotically valid. Simulations suggest that bootstrap inference tends to be much more accurate than inference based on the t distribution, especially when there are few clusters in at least one dimension. An empirical example confirms that bootstrap inferences can differ substantially from conventional ones.

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

JEL Codes: C15, C21, C23.

1 Introduction

The disturbances (error terms) in regression models often appear to be correlated within clusters. It is generally 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), Carter, Schnepel, and Steigerwald (2017), Pustejovsky and Tipton (2017), and Djogbenou, MacKinnon, and Nielsen (2017).

Although methods for one-way clustering are sufficient in many cases, it is often plausible that clustering should occur in two or more dimensions. For example, for panel data there may well be correlations both within jurisdictions across time periods and within time periods across

*We are grateful to Russell Davidson and seminar participants at the 2017 CEA Annual Meeting for comments. 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. Computer code for performing the bootstrap procedures proposed here may be found at http://qed.econ.queensu.ca/pub/faculty/mackinnon/two-way-boot/.

†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.

jurisdictions. Cameron, Gelbach, and Miller (2011) (CGM hereafter) proposes a method to calculate standard errors that are robust to multiway clustering, and this method has been widely used in empirical work. However, CGM does not state the conditions under which their "multiway CRVE" is asymptotically valid or provide a formal proof. Moreover, simulations in CGM suggest that using multiway cluster-robust standard errors does not always work well, especially when the number of clusters in either dimension is small.

In this paper, we obtain two important results. Firstly, we prove that the multiway CRVE is asymptotically valid for the case of two-dimensional clustering under precisely stated conditions. Variations of this CRVE can handle clustering in more than two dimensions, and it is clear that our proofs could be extended to handle such cases. However, we do not attempt to analyze higher-dimensional clustering, because the notation would be extremely tedious. To our knowledge, empirical work that uses the multiway CRVE very rarely goes beyond the two-dimensional case.

Secondly, we propose eight bootstrap methods and prove that all of them are asymptotically valid for the case of two-dimensional clustering. Two methods simply combine the multiway CRVE with the ordinary wild bootstrap, using either restricted or unrestricted estimates, and the other six combine it with variants of the wild cluster bootstrap. To our knowledge, these are the first bootstrap methods for this problem. Menzel (2017) develops a bootstrap procedure for multiway clustering, but that procedure is for comparisons of means and not for inference on regression coefficients. In every case that we have investigated, the proposed bootstrap methods (at least, the ones based on restricted estimates) yield much better inferences in finite samples than comparing t-statistics based on the multiway CRVE to the t distribution with degrees of freedom that depend on the numbers of clusters in each dimension.

In Section 2, we discuss the linear regression model with disturbances that are clustered in two dimensions and the two-way CRVE proposed in CGM. Then, in Section 3, we prove that inference based on the latter is asymptotically valid under a suitable set of assumptions. In Section 4, we discuss the wild bootstrap methods that we propose and prove that they too are asymptotically valid. In Section 5, we present the results from a number of simulation experiments which suggest that wild bootstrap inference is much more reliable than asymptotic inference, except perhaps in certain extreme cases. In Section 6, we illustrate our results by using 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. Mathematical proofs of our main results are presented in the appendix.

2 The Model

Consider a linear regression model with two-way clustered disturbances written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top) = \boldsymbol{\Omega}, \tag{1}$$

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, $\boldsymbol{\beta}$ is a $k \times 1$ parameter vector, and the $N \times N$ variance matrix $\boldsymbol{\Omega}$ has a particular structure based on two dimensions of clustering. The numbers of clusters in the two dimensions are G and H, respectively. We can rewrite (1) as

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

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 equation (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 excessively ugly algebra, we omit the superscripts. Thus we simply 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 equation (2), the vector $\boldsymbol{y}_g$ contains the subvectors $\boldsymbol{y}_{g1}$ through $\boldsymbol{y}_{gH}$. The variance matrices for $\boldsymbol{u}_g$, $\boldsymbol{u}_h$ and $\boldsymbol{u}_{gh}$ are denoted

$$\boldsymbol{\Omega}_g = \mathrm{E}(\boldsymbol{u}_g\boldsymbol{u}_g^\top \,|\, \boldsymbol{X}), \boldsymbol{\Omega}_h = \mathrm{E}(\boldsymbol{u}_h\boldsymbol{u}_h^\top \,|\, \boldsymbol{X}), \text{ and } \boldsymbol{\Omega}_{gh} = \mathrm{E}(\boldsymbol{u}_{gh}\boldsymbol{u}_{gh}^\top \,|\, \boldsymbol{X}), \tag{3}$$

respectively.

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.

Conditional on $\boldsymbol{X}$, the disturbances have mean zero and variance matrix $\boldsymbol{\Omega} = \mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top \,|\, \boldsymbol{X})$ with the structure

$$\mathrm{E}(\boldsymbol{u}_{g'h'}\boldsymbol{u}_{gh}^\top \,|\, \boldsymbol{X}) = \boldsymbol{0} \quad \text{if} \quad g' \neq g,\, h' \neq h \tag{4}$$

and arbitrary covariances if either $g = g'$ or $h = h'$. When $N_h = 1$ for all h, the model (1) reduces to the conventional one-way clustering model. When, in addition, $N_g = 1$ for all g, it reduces to the well-known linear regression model with heteroskedasticity of unknown form. Hence, as special cases, our results cover these models as well.

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

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}. \tag{5}$$

We let $\boldsymbol{Q}_N = N^{-1}\boldsymbol{X}^\top\boldsymbol{X}$ and $\boldsymbol{\Gamma}_N = N^{-2}\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X}$. With the structure in (4), we can write

$$\boldsymbol{\Gamma}_N = N^{-2}\sum_{g=1}^{G}\boldsymbol{X}_g^\top\boldsymbol{\Omega}_g\boldsymbol{X}_g + N^{-2}\sum_{h=1}^{H}\boldsymbol{X}_h^\top\boldsymbol{\Omega}_h\boldsymbol{X}_h - N^{-2}\sum_{g=1}^{G}\sum_{h=1}^{H}\boldsymbol{X}_{gh}^\top\boldsymbol{\Omega}_{gh}\boldsymbol{X}_{gh}. \tag{6}$$

The variance matrix of $\hat{\boldsymbol{\beta}}$, conditional on $\boldsymbol{X}$, is given by

$$\boldsymbol{V}_N = \boldsymbol{Q}_N^{-1}\boldsymbol{\Gamma}_N\boldsymbol{Q}_N^{-1}. \tag{7}$$

We then define the cluster-robust estimator of $\boldsymbol{V}_N$, i.e. the multiway CRVE, as

$$\hat{\boldsymbol{V}} = \boldsymbol{Q}_N^{-1}\hat{\boldsymbol{\Gamma}}\boldsymbol{Q}_N^{-1}. \tag{8}$$

Based on (6), the middle matrix here is defined as

$$\hat{\boldsymbol{\Gamma}} = N^{-2}\sum_{g=1}^{G}\boldsymbol{X}_g^\top\hat{\boldsymbol{u}}_g\hat{\boldsymbol{u}}_g^\top\boldsymbol{X}_g + N^{-2}\sum_{h=1}^{H}\boldsymbol{X}_h^\top\hat{\boldsymbol{u}}_h\hat{\boldsymbol{u}}_h^\top\boldsymbol{X}_h - N^{-2}\sum_{g=1}^{G}\sum_{h=1}^{H}\boldsymbol{X}_{gh}^\top\hat{\boldsymbol{u}}_{gh}\hat{\boldsymbol{u}}_{gh}^\top\boldsymbol{X}_{gh}; \tag{9}$$

see CGM. Here $\hat{\boldsymbol{u}}_g$, $\hat{\boldsymbol{u}}_h$, and $\hat{\boldsymbol{u}}_{gh}$ denote various subvectors of the vector of OLS residuals. In practice, the factors of N^{-2} in the three terms in (9) are almost always omitted, and $\boldsymbol{Q}_N$ is replaced by $\boldsymbol{X}^\top\boldsymbol{X}$. This leaves the value of $\hat{\boldsymbol{V}}$ unchanged. However, the three terms in (9) are usually multiplied by

$$\frac{G(N-1)}{(G-1)(N-k)}, \quad \frac{H(N-1)}{(H-1)(N-k)}, \quad \text{and} \quad \frac{GH(N-1)}{(GH-1)(N-k)}, \tag{10}$$

respectively, by analogy with the scalar factor that is conventionally employed with the one-way CRVE. We make use of the factors in (10) in our simulations, but for purposes of asymptotic theory we omit them without loss of generality.

One important practical issue is that the matrix $\hat{\boldsymbol{\Gamma}}$ defined in (9) is not necessarily positive definite in finite samples, which implies that the diagonal elements of $\hat{\boldsymbol{V}}$ may not all be positive. In fact, since the ranks of the three matrices in (9) cannot exceed G, H, and GH, respectively, it seems likely that $\hat{\boldsymbol{V}}$ will not be positive definite whenever the model contains significantly more than $\min(G, H)$ regressors. However, even when the number of regressors is small relative to G and H, there may very well be samples for which $\hat{\boldsymbol{V}}$ is not positive definite; see Section 5. For the models estimated in the empirical example of Section 6, $\hat{\boldsymbol{V}}$ is in fact not positive definite.

In order to deal with this problem, CGM suggests calculating the eigenvalues of $\hat{\boldsymbol{V}}$, say $\lambda_1, \ldots, \lambda_k$. When any of them is not positive, they then suggest that $\hat{\boldsymbol{V}}$ be replaced by the eigendecomposition $\hat{\boldsymbol{V}}^+ = \boldsymbol{U}\boldsymbol{\Lambda}^+\boldsymbol{U}^\top$, 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 $\hat{\boldsymbol{V}}^+$ is guaranteed only to be positive semidefinite, and it could have diagonal elements that equal 0. If the coefficients that correspond to those diagonal elements are of interest, then it is impossible to use $\hat{\boldsymbol{V}}^+$ for inference about them.

Even for the one-way CRVE, it is common to encounter singular variance matrices when there are fixed effects. This problem can be dealt with in various ways, most easily by projecting the regressand and all the regressors off the fixed effects before running the regression; see Pustejovsky and Tipton (2017). That trick could also be used with the two-way CRVE.

3 Asymptotic Theory

In this section, we derive the asymptotic limit theory for t-statistics based on the two-way CRVE $\hat{\boldsymbol{V}}$. We let $\boldsymbol{\beta}_0$ denote the true value of $\boldsymbol{\beta}$ and restrict our attention to the cluster-robust t-statistic

$$t_a = \frac{\boldsymbol{a}^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)}{\sqrt{\boldsymbol{a}^\top\hat{\boldsymbol{V}}\boldsymbol{a}}} \tag{11}$$

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. We impose the normalization that $\boldsymbol{a}^\top\boldsymbol{a} = 1$ to rule out degenerate cases, but it is much stronger than is actually needed.

In order to obtain our results, we need the following conditions, where, for any matrix $\boldsymbol{M}$, $\|\boldsymbol{M}\| = \left(\mathrm{Tr}(\boldsymbol{M}^\top\boldsymbol{M})\right)^{1/2}$ denotes the Euclidean (Frobenius) norm.

Assumption 1. The disturbances are such that $\boldsymbol{u}_{gh}$ is independent of $\boldsymbol{u}_{g'h'}$ if $g \neq g', h \neq h'$ and satisfy $\mathrm{E}(\boldsymbol{u}_{gh}|\boldsymbol{X}) = \boldsymbol{0}$ and (3) with $\boldsymbol{\Omega}_g$, $\boldsymbol{\Omega}_h$, and $\boldsymbol{\Omega}_{gh}$ positive definite. In addition, for some $\lambda > 0$,

$$\sup_{1\leq i\leq N_{gh},\, 1\leq g\leq G,\, 1\leq h\leq H} \mathrm{E}\left(|u_{gh,i}|^{2+\lambda} \,|\, \boldsymbol{X}\right) < \infty,$$

where $u_{gh,i}$ denotes the i^{th} element of $\boldsymbol{u}_{gh}$.

Assumption 2. The regressor matrix $\boldsymbol{X}$ satisfies $\boldsymbol{Q}_N \xrightarrow{P} \boldsymbol{Q}$, where $\boldsymbol{Q}$ is finite and positive definite, and

$$\sup_{1\leq i\leq N_{gh},\, 1\leq g\leq G,\, 1\leq h\leq H} \mathrm{E}\|\boldsymbol{X}_{gh,i}\|^{2+\lambda} < \infty,$$

where λ is the same as in Assumption 1 and $\boldsymbol{X}_{gh,i}$ denotes the i^{th} row of $\boldsymbol{X}_{gh}$.

Assumption 3. For λ defined in Assumption 1,

$$\frac{\sup_{1\leq g\leq G} N_g^{2+2/\lambda}}{N} + \frac{\sup_{1\leq h\leq H} N_h^{2+2/\lambda}}{N} \to 0.$$

Assumption 1 imposes the conditions that the disturbance vectors $\boldsymbol{u}_{gh}$ are independent across clusters with zero conditional means and constant, but possibly heterogeneous, conditional variance matrices. Conditions like Assumption 2 are standard in the asymptotic theory for linear regressions.

Assumption 3 restricts the extent of cluster size heterogeneity that is allowed in order to obtain asymptotic normality. Although the number of observations in each cluster can grow as the total number of observations, N, grows, the relative expansion rate of the clusters is controlled by the conditions in Assumption 3. More specifically, the cluster sizes in both dimensions can be either fixed constants or they can diverge. For example, we might have $N_g = c_g N^{\alpha_g}$, $N_h = c_h N^{\alpha_h}$, or $N_{gh} = c_{gh} N^{\alpha_{gh}}$, but α_g, α_h, and α_{gh} cannot be too large, and no cluster can be proportional to the entire sample.

In any case, the condition implies that the numbers of clusters in both dimensions, i.e. G and H, must diverge when N diverges. It is also clear that when more moments are assumed to exist in Assumption 1, i.e. when λ is higher, Assumption 3 is weaker. If, for example, four moments are assumed to exist ($\lambda = 2$) as in Djogbenou, MacKinnon, and Nielsen (2017), then the second condition of Assumption 3 is $\sup_g N_g^3/N + \sup_h N_h^3/N \to 0$, which is only slightly stronger than what is required in the one-way clustering setup in Djogbenou et al. (2017).

More generally, the conditions in Assumption 3 ensure that the information in the sample remains sufficiently spread out across clusters asymptotically, which is a critical requirement for the application of a central limit theorem. Therefore, the condition restricts the sizes of the largest clusters in each dimension, $\sup_g N_g$ and $\sup_h N_h$.

A substantial complication in the asymptotic theory for model (1) is that the stochastic order of magnitude of $\hat{\boldsymbol{\beta}}$ in (5) depends in a complex way on the intra-cluster correlation structure, the regressors, the relative cluster sizes, and interactions among these; see Carter, Schnepel, and Steigerwald (2017) and Djogbenou, MacKinnon, and Nielsen (2017). There are two extreme cases, with all other cases lying in between: (i) $\boldsymbol{\Omega}$ is diagonal with no intra-cluster correlation at all, and (ii) the $\boldsymbol{\Omega}_{gh}$ are dense matrices without restrictions, and the regressors are correlated within clusters. In case (i) we easily find that, under Assumption 2,

$$\|\boldsymbol{V}_N\| = O_P(N^{-1}). \tag{12}$$

Thus, in particular, $\hat{\boldsymbol{\beta}}$ clearly converges at rate $O_P(N^{-1/2})$, because $\boldsymbol{V}_N$ is the conditional variance matrix of $\hat{\boldsymbol{\beta}}$ under Assumption 1. On the other hand, in case (ii) for general $\boldsymbol{\Omega}_{gh}$ without restrictions, it holds that

$$\mathrm{E}(\boldsymbol{X}_{gh}^\top \boldsymbol{\Omega}_{gh} \boldsymbol{X}_{gh}) = \mathrm{E}\Bigg(\sum_{i,j=1}^{N_{gh}} \boldsymbol{X}_{gh,i}^\top \Omega_{gh,ij} \boldsymbol{X}_{gh,j} \Bigg) = O(N_{gh}^2), \tag{13}$$

where $\Omega_{gh,ij}$ is the $(i,j)^{\text{th}}$ element of $\boldsymbol{\Omega}_{gh}$. Similarly, $\mathrm{E}(\boldsymbol{X}_g^\top \boldsymbol{\Omega}_g \boldsymbol{X}_g) = O(N_g^2)$ and $\mathrm{E}(\boldsymbol{X}_h^\top \boldsymbol{\Omega}_h \boldsymbol{X}_h) =$

$O(N_h^2)$. It follows, using also (6), (7), and Assumption 2, that

$$\|\boldsymbol{V}_N\| = O_P(1)O_P\Big(N^{-2}\sum_{g=1}^{G} N_g^2 + N^{-2}\sum_{h=1}^{H} N_h^2 + N^{-2}\sum_{g=1}^{G}\sum_{h=1}^{H} N_{gh}^2\Big)$$
$$= O_P\Big(N^{-1}\sup_{1\leq g\leq G} N_g + N^{-1}\sup_{1\leq h\leq H} N_h\Big). \tag{14}$$

Note that $N^{-1}\sup_{g,h} N_{gh} \to 0$ is implied by $N^{-1}\sup_g N_g \to 0$ and $N^{-1}\sup_h N_h \to 0$. Therefore, in case (ii), $\hat{\boldsymbol{\beta}}$ converges at rate $O_P(N^{-1/2}\sup_g N_g^{1/2} + N^{-1/2}\sup_h N_h^{1/2})$. In general, it follows that, under Assumptions 1 and 2, the condition

$$\frac{\sup_{1\leq g\leq G} N_g}{N} + \frac{\sup_{1\leq h\leq H} N_h}{N} \to 0 \tag{15}$$

is sufficient for consistency of $\hat{\boldsymbol{\beta}}$ in the model (1).

Our first result in Theorem 1 below has several precursors in the literature on one-way clustering, although these are all obtained under assumptions that are very different from ours. In particular, White (1984, Chapter 6) assumes equal-sized, homogeneous (same variance) clusters, and Hansen (2007) assumes equal-sized, heterogeneous clusters. Thus both these papers assume that $N_g = N/G$ for all g. In contrast, the primitive moment and rate conditions in Assumptions 1 and 3 allow clusters to be heterogeneous.

More recently, Carter, Schnepel, and Steigerwald (2017) and Djogbenou, MacKinnon, and Nielsen (2017) obtain results for one-way clustering that allow clusters to be heterogeneous. The former paper invokes a primitive moment condition and makes some high-level assumptions about cluster-size heterogeneity and interactions between regressors and disturbances. The latter paper makes much weaker assumptions, similar to those in Assumptions 1–3, which allow for arbitrary dependence and correlation within each cluster. See Djogbenou et al. (2017) for a detailed comparison of the assumptions in the two papers.

Since we do not restrict the dependence within clusters in either dimension, and we wish to allow any structure for the intra-cluster variance matrices, we cannot normalize $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ in the usual way to obtain an asymptotic distribution. Instead, we follow Djogbenou, MacKinnon, and Nielsen (2017) and consider asymptotic limit theory for the studentized (self-normalized) quantities $(\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-1/2}\boldsymbol{a}^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, $(\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-1}\boldsymbol{a}^\top \hat{\boldsymbol{V}}\boldsymbol{a}$, and t_a. For related arguments, see Hansen (2007, Theorem 2) and Carter, Schnepel, and Steigerwald (2017).

The following result establishes the asymptotic normality of $\hat{\boldsymbol{\beta}}$ and t_a.

Theorem 1. *Suppose that Assumptions 1–3 are satisfied and the true value of $\boldsymbol{\beta}$ is given by $\boldsymbol{\beta}_0$. It then holds that*

$$\frac{\boldsymbol{a}^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)}{(\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{1/2}} \xrightarrow{d} \mathrm{N}(0,1), \tag{16}$$

and if also $\lambda \geq 2$ then

$$\frac{\boldsymbol{a}^\top \hat{\boldsymbol{V}}\boldsymbol{a}}{\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a}} \xrightarrow{P} 1 \text{ and} \tag{17}$$

$$t_a \xrightarrow{d} \mathrm{N}(0,1). \tag{18}$$

Equation (18) justifies the use of critical values and P values from a normal approximation to perform t-tests and construct confidence intervals. However, based on results in Bester, Conley,

and Hansen (2011), it will often be more accurate to use the $t(G-1)$ distribution in the one-way case; see also Cameron and Miller (2015) for a discussion of this issue. In the two-way case, CGM suggests using the $t(\min(G,H)-1)$ distribution, and we do this in Sections 5 and 6 below.

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 generally not the number of observations, N. This is seen clearly in the rate of convergence of the estimator in (16), which is $(\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{1/2}$ instead of $N^{-1/2}$, $G^{-1/2}$, or $H^{-1/2}$; see the discussion around (14).

4 Asymptotic Validity of the Wild (Cluster) Bootstrap

In this section, we consider the asymptotic validity of inference based on several variants of the wild bootstrap, or WB, and the wild cluster bootstrap, or WCB (Cameron, Gelbach, and Miller, 2008), as alternatives to the asymptotic inference justified in Theorem 1. Both the WB and WCB may be implemented using either restricted or unrestricted estimates in the bootstrap data-generating process. In general, it is desirable to impose restrictions on bootstrap DGPs (Davidson and MacKinnon, 1999), and there is compelling evidence in MacKinnon and Webb (2017) and Djogbenou, MacKinnon, and Nielsen (2017) that, for one-way clustering, the restricted versions of both the WB and WCB (henceforth WR and WCR) never perform much worse than the unrestricted versions (WU and WCU) and sometimes perform very much better. However, it is much easier to construct studentized bootstrap confidence intervals using WU and WCU than to construct confidence intervals based on WR and WCR; Hansen (1999) and MacKinnon (2015) discuss confidence intervals based on restricted bootstraps.

For the wild bootstrap, the bootstrap disturbance vectors $\boldsymbol{u}^*$ are obtained by multiplying each residual $\tilde{u}_{gh,i}$ (for WR) or $\hat{u}_{gh,i}$ (for WU) by a draw $v^*_{gh,i}$ from an auxiliary random variable v^* with mean 0 and variance 1. A popular choice is the Rademacher distribution, which takes on the values 1 and -1 with equal probabilities; see Davidson and Flachaire (2008). Thus it takes N draws from the auxiliary distribution to create each bootstrap sample.

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.[1] For each bootstrap sample, every residual within each cluster in the appropriate dimension is multiplied by the same draw from v^*.

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. However, when clustering is actually in two dimensions, the best the WCB can do is to preserve some of the intra-cluster correlations. In Theorem 2 below, we prove that both the WB and all suggested variants of the WCB are asymptotically valid. However, it is not clear which variant of the WCB is likely to perform best in any given case or whether any variant is likely to outperform the WB. This undoubtedly depends on G, H, the cluster sizes, $\boldsymbol{X}$, $\boldsymbol{\Omega}$, and so on.

We next describe the algorithm for all variants of the WB and WCB in some detail. We then prove the asymptotic validity of all variants. To describe the bootstrap algorithm and the properties of the bootstrap procedures, we introduce the notation $\ddot{\boldsymbol{u}}$ and $\ddot{\boldsymbol{\beta}}$, which will be taken to represent either restricted or unrestricted quantities as appropriate.

[1] Actually, if any of the possible intersections of the two dimensions were empty, the number of draws in the last case would be less than GH. For simplicity, we ignore this possibility.

Multiway Wild (Cluster) Bootstrap Algorithms.

1. Run an OLS regression of $\boldsymbol{y}$ on $\boldsymbol{X}$ to obtain $\hat{\boldsymbol{\beta}}$ and (multiway) $\hat{\boldsymbol{V}}$ defined in (5) and (8), respectively. Check that $\hat{\boldsymbol{V}}$ is positive semidefinite, and replace it by $\hat{\boldsymbol{V}}^{+}$ 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, given in (11), 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 the bootstrap (multiway) variance matrix estimate

   $$\hat{\boldsymbol{V}}^{*b} = (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}N^2\hat{\boldsymbol{\Gamma}}^{*b}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1} = \boldsymbol{Q}_N^{-1}\hat{\boldsymbol{\Gamma}}^{*b}\boldsymbol{Q}_N^{-1},$$

   where

   $$\begin{aligned} N^2\hat{\boldsymbol{\Gamma}}^{*b} =& \frac{G(N-1)}{(G-1)(N-k)}\sum_{g=1}^{G}\boldsymbol{X}_g^{\top}\hat{\boldsymbol{u}}_g^{*b}\hat{\boldsymbol{u}}_g^{*b\top}\boldsymbol{X}_g + \frac{H(N-1)}{(H-1)(N-k)}\sum_{h=1}^{H}\boldsymbol{X}_h^{\top}\hat{\boldsymbol{u}}_h^{*b}\hat{\boldsymbol{u}}_h^{*b\top}\boldsymbol{X}_h \\ & - \frac{GH(N-1)}{(GH-1)(N-k)}\sum_{g=1}^{G}\sum_{h=1}^{H}\boldsymbol{X}_{gh}^{\top}\hat{\boldsymbol{u}}_{gh}^{*b}\hat{\boldsymbol{u}}_{gh}^{*b\top}\boldsymbol{X}_{gh}; \end{aligned}$$

   If $\hat{\boldsymbol{V}}^{*b}$ is not positive semidefinite, replace it by $\hat{\boldsymbol{V}}^{*b+}$, which is the bootstrap analogue of the matrix $\hat{\boldsymbol{V}}^{+}$.

   (d) Calculate the bootstrap t-statistic

   $$t_a^{*b} = \frac{\boldsymbol{a}^{\top}(\hat{\boldsymbol{\beta}}^{*b} - \ddot{\boldsymbol{\beta}})}{\sqrt{\boldsymbol{a}^{\top}\hat{\boldsymbol{V}}^{*b}\boldsymbol{a}}}.$$

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

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

   where $\mathbb{I}(\cdot)$ denotes the indicator function. If the null hypothesis is H_2, but it is inappropriate to assume symmetry, then the symmetric P value $\hat{P}^*_S$ can be replaced by the equal-tail P value, which is simply $2\min(\hat{P}^*_{\mathrm{L}}, \hat{P}^*_{\mathrm{H}})$.

The above algorithm presents the steps needed to implement the WR, WCR, WU, and WCU bootstraps for testing the hypothesis H_0. If interest instead focuses on confidence intervals for the quantity $\boldsymbol{a}^{\top}\boldsymbol{\beta}$, studentized bootstrap confidence intervals based on WU and WCU can be constructed by calculating lower-tail and upper-tail quantiles of the t_a^{*b} instead of P values; see Davidson and MacKinnon (2004, Section 5.3).

A key feature of all the wild bootstrap methods we consider is that the variance matrix $\hat{\boldsymbol{V}}^{*b}$ obtained in step 3(c) does not in fact consistently estimate the limit of the true variance matrix $\boldsymbol{V}_N$; see (20) below. To accommodate this fact, we introduce the following notation. Let $\bar{\boldsymbol{\Omega}}$ denote the limiting variance matrix of the bootstrap disturbances. For the WB, this will be the matrix obtained by setting all the off-diagonal elements of $\boldsymbol{\Omega}$ to zero. For the WCB, it will be a block diagonal matrix, with either G, H, or GH nonzero blocks. We also define

$$\bar{\boldsymbol{\Gamma}}_N = N^{-2}\boldsymbol{X}^\top\bar{\boldsymbol{\Omega}}\boldsymbol{X} \quad \text{and} \quad \bar{\boldsymbol{V}}_N = \boldsymbol{Q}_N^{-1}\bar{\boldsymbol{\Gamma}}_N\boldsymbol{Q}_N^{-1};$$

see equations (6) and (7). The form of $\bar{\boldsymbol{\Gamma}}_N$ is much simpler than the form of $\boldsymbol{\Gamma}_N$, because the former involves at most one-way clustering while the latter involves two-way clustering. Notice that, except in very special cases, $\bar{\boldsymbol{V}}_N \neq \boldsymbol{V}_N$.

The following theorem is the bootstrap analogue of Theorem 1. It establishes the asymptotic normality of the WB and WCB estimators and t-statistics.

Theorem 2. *Suppose Assumptions 1–3 are satisfied with $\lambda \geq 2$, that the true value of $\boldsymbol{\beta}$ is $\boldsymbol{\beta}_0$, and that $\mathrm{E}^*|v^*|^{2+\lambda} < \infty$. It then holds that*

$$\frac{\boldsymbol{a}^\top(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})}{(\boldsymbol{a}^\top\bar{\boldsymbol{V}}_N\boldsymbol{a})^{1/2}} \xrightarrow{d^*} \mathrm{N}(0,1), \tag{19}$$

$$\frac{\boldsymbol{a}^\top\hat{\boldsymbol{V}}^*\boldsymbol{a}}{\boldsymbol{a}^\top\bar{\boldsymbol{V}}_N\boldsymbol{a}} \xrightarrow{P^*} 1, \text{ and} \tag{20}$$

$$t_a^* \xrightarrow{d^*} \mathrm{N}(0,1), \tag{21}$$

in probability.

Recall that $\ddot{\boldsymbol{\beta}}$ denotes whatever vector of estimates was used in step 3(b) above. For WR and WCR, this estimate satisfies the null hypothesis. For WU and WCU, it is the unrestricted OLS estimate of $\boldsymbol{\beta}$ in model (1). From (19), we see that the WB and WCB are unable to replicate the variance matrix of the vector $\hat{\boldsymbol{\beta}}$. The bootstrap estimator $\boldsymbol{a}^\top\hat{\boldsymbol{\beta}}^*$ asymptotically has variance $\boldsymbol{a}^\top\bar{\boldsymbol{V}}_N\boldsymbol{a}$, conditional on the original sample, whereas the actual estimator $\boldsymbol{a}^\top\hat{\boldsymbol{\beta}}$ asymptotically has variance $\boldsymbol{a}^\top\boldsymbol{V}_N\boldsymbol{a}$; compare (16) and (19).

More importantly, however, the distribution of the bootstrap t-statistic given in (21) replicates that of the original sample t-statistic. 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 $\boldsymbol{a}^\top\hat{\boldsymbol{\beta}}$, this has no effect on the asymptotic validity of the bootstrap, because t_a^* is based on a valid estimate of $\boldsymbol{a}^\top\bar{\boldsymbol{V}}_N\boldsymbol{a}$ by the result (20) and thus has the correct asymptotic distribution by the result (21).

Furthermore, all the results in Theorem 2 are conditional on the original sample, and hence also conditional on t_a. This implies that the results (19)–(21) hold for any possible realization of the original sample, and therefore also any possible realization of t_a, which is the crucial requirement for asymptotic validity of the bootstrap.

Let the cumulative distribution function (CDF) of t_a be denoted $P(t_a \leq x)$. Then the following result follows immediately from Theorems 1 and 2 by an application of the triangle inequality and Polya's Theorem.

Corollary 1. *Under the conditions of Theorem 2 and H_0,*

$$\sup_x |P^*(t_a^* \leq x) - P(t_a \leq x)| = o_P(1).$$

Corollary 1 implies that the P values computed in step 4 of the WB and WCB algorithms are asymptotically valid, as are studentized bootstrap confidence intervals. Intuitively, the bootstrap test must have the correct size asymptotically under the null hypothesis, because comparing t_a to the bootstrap distribution $P^*(t_a^* \leq x)$ is asymptotically equivalent to comparing it to $\mathrm{N}(0,1)$; c.f. the results (18) and (21).

5 Simulation experiments

In Theorems 1 and 2, we proved the asymptotic validity of inference based on t-statistics constructed using the multiway CRVE and on several variants of the wild bootstrap. However, we did not prove that the latter will outperform the former in finite samples, and we would very surprised if such a result could be proved. Nevertheless, it is reasonable to conjecture that bootstrap inference will typically be more reliable than inference based on the t distribution. In this section, we investigate this conjecture via simulation experiments. We also investigate how the performance of all methods depends on the number of clusters in each dimension and on other features of the data generating process.

In most of our experiments, the DGP is

$$y_{gh,i} = \beta_0 + \beta_1 X_{gh,i} + u_{gh,i}, \quad u_{gh,i} = \sigma_g v_g + \sigma_h v_h + \sigma_\epsilon \epsilon_{gh,i}, \tag{22}$$

where the v_g, v_h, and $\epsilon_{gh,i}$ are all independent standard normals. The values of σ_g, σ_h, 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_g (or ρ_h). This implies that the correlation is $\rho_g + \rho_h$ for disturbances that belong to the same cluster in both dimensions, and 0 for disturbances that do not belong to the same cluster in either dimension.

The regressor $X_{gh,i}$ is lognormally distributed. It is the exponential of a random variable that is generated in almost the same way as the $u_{gh,i}$, but with correlations ϕ_g and ϕ_h. We make $X_{gh,i}$ lognormal to avoid the risk that any of our results may be artifacts of an experimental design in which both the regressor and the disturbances are normally distributed.

Table 1 presents rejection frequencies for a large number of procedures for various values of G and H with balanced clusters. For example, when $N = 4000$ with $G = 10$ and $H = 20$, the clusters in the G dimension each contain 400 observations, the clusters in the H dimension each contain 200, and the clusters in the $G \times H$ dimension each contain 20. The details of the experiments are given in the notes to Table 1. Both ρ_g and ρ_h are equal to 0.05, and both ϕ_g and ϕ_h are equal to 0.40.[2] That is because, in our experience, intra-cluster correlations for residuals tend to be quite small, while intra-cluster correlations for at least some regressors can be large.

It is evident from Table 1 that using heteroskedasticity-robust standard errors, or clustering at the $G \times H$ level, always leads to very severe overrejection. Note that these procedures are identical when $N = GH$, as in the last three rows of the table. One-way clustering also always leads to overrejection, which can be very severe, especially when H is much larger than G and we cluster by H. Using t-statistics based on the two-way CRVE (denoted CV-M in the table), together with critical values from the $t(\min(G,H)-1)$ distribution, always results in overrejection, which is quite severe in some cases, but not nearly as severe as with one-way clustering. There is overrejection even when $G = H = 100$, which may be surprising. We would have obtained somewhat smaller rejection frequencies (but still greater than .05) if the regressor had been normally rather than lognormally distributed or if ρ_g and ρ_h had been larger.

[2] In contrast, CGM implicitly set $\rho_g = \rho_h = 1/3$ in many of their experiments. Their DGP has two regressors, each correlated in just one dimension, instead of one regressor that is correlated in both dimensions.

Table 1: Rejection Frequencies for Balanced Clusters with Homoskedastic Errors

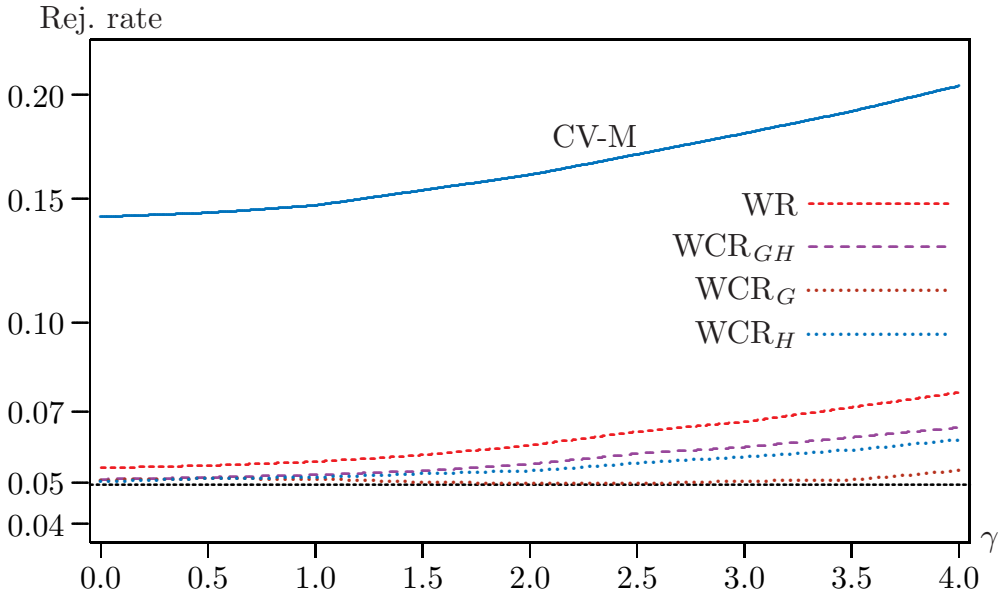
G	H	HC_1	CV_{GH}	CV_G	CV_H	CV-M	WR	WCR_{GH}	WCR_G	WCR_H	$\lambda_j \leq 0$
5	5	0.6865	0.3312	0.2322	0.2316	0.1934	0.1073	0.0954	0.0810	0.0811	0.1067
5	10	0.6378	0.3312	0.1977	0.2672	0.1300	0.0753	0.0682	0.0609	0.0579	0.0294
5	20	0.5902	0.4179	0.1657	0.3004	0.1004	0.0592	0.0540	0.0536	0.0533	0.0096
5	40	0.5495	0.4179	0.1375	0.3389	0.0876	0.0502	0.0465	0.0509	0.0538	0.0043
10	10	0.5814	0.3727	0.2277	0.2263	0.1427	0.0544	0.0514	0.0514	0.0515	0.0047
10	20	0.5251	0.2319	0.1867	0.2511	0.1151	0.0443	0.0422	0.0485	0.0517	0.0013
10	40	0.4762	0.3926	0.1551	0.2806	0.1026	0.0416	0.0399	0.0468	0.0544	0.0007
20	20	0.4519	0.3648	0.1976	0.1979	0.1115	0.0387	0.0371	0.0488	0.0488	0.0005
20	40	0.3933	0.3477	0.1575	0.2153	0.0976	0.0386	0.0374	0.0475	0.0516	0.0003
40	40	0.3218	0.3014	0.1620	0.1624	0.0919	0.0390	0.0385	0.0489	0.0485	0.0001
50	50	0.4521	0.4137	0.1783	0.1790	0.0832	0.0380	0.0375	0.0491	0.0496	0.0000
50	100	0.3920	0.3782	0.1351	0.2047	0.0775	0.0412	0.0410	0.0476	0.0519	0.0000
50	200	0.3533	0.3533	0.1099	0.2371	0.0761	0.0452	0.0452	0.0472	0.0538	0.0000
25	400	0.4422	0.4422	0.1051	0.3764	0.0899	0.0492	0.0492	0.0476	0.0595	0.0000
100	100	0.3169	0.3169	0.1456	0.1460	0.0718	0.0433	0.0433	0.0493	0.0492	0.0000

Notes:

All results are based on 400,000 replications, with $B = 399$ for the bootstrap methods.
All tests are at 5% nominal level.
The model contains a constant and one regressor, which follows a lognormal distribution. The disturbances are normally distributed.
When $G \leq 40$ and $H \leq 40$, $N = 4000$. Otherwise, $N = 10{,}000$.
In all cases, $\rho_g = \rho_h = 0.05$, and $\phi_g = \phi_h = 0.40$.
HC_1 uses heteroskedasticity-robust standard errors and $t(n-k)$ critical values.
CV_{GH} uses standard errors clustered at the $G \times H$ level and $t(GH-1)$ critical values.
CV_G uses standard errors clustered at the G level and $t(G-1)$ critical values.
CV_H uses standard errors clustered at the H level and $t(H-1)$ critical values.
CV-M uses multiway CRVE standard errors (using the eigendecomposition if necessary) and $t(\min(G,H)-1)$ critical values.
In most cases, the auxiliary random variable for the wild bootstrap is Rademacher. For WCR_G with $G = 5$ and WCR_H with $H = 5$, it follows Webb's 6-point distribution.
WR uses ordinary wild bootstrap symmetric P values.
WCR_{GH} uses wild cluster bootstrap symmetric P values, with clustering at the $G \times H$ level.
WCR_G uses wild cluster bootstrap symmetric P values, with clustering at the G level.
WCR_H uses wild cluster bootstrap symmetric P values, with clustering at the H level.
$\lambda_j \leq 0$ is the proportion of simulations for which any of the eigenvalues of $\hat{\boldsymbol{V}}$ is less than 10^{-8}.

In contrast, the bootstrap methods generally work very well and perform very similarly, except when $G = 5$ and $H \leq 10$. For the larger values of G and H, the WR and WCR_{GH} bootstraps actually underreject. This is evident even for $G = H = 100$. When the number of draws of the auxiliary random variable is very small (that is, when $G = 5$ for WCR_G and when $H = 5$ for WCR_H), we use the 6-point distribution proposed in Webb (2014) rather than the Rademacher distribution. For the latter, there would have been only $2^5 = 32$ distinct bootstrap samples. For the former, there are $6^5 = 7776$.

Figure 1: Rejection frequencies when cluster sizes vary (restricted bootstraps)



Notes:
There are 400,000 replications, with $N = 4000$ and $G = H = 10$.
All tests are at 5% nominal level.
CV-M rejection frequencies are based on the $t(9)$ distribution.
All bootstrap tests use the Rademacher distribution with symmetric P values and $B = 399$.

Interestingly, in every case, either the WCR$_G$ bootstrap, the WCR$_H$ bootstrap, or both of them, perform better than the WCR$_{GH}$ bootstrap. Except when $G = H = 5$, both these methods work remarkably well. When $G = H$, the WCR$_G$ and WCR$_H$ bootstraps are equivalent, although the rejection frequencies differ slightly because of simulation errors.

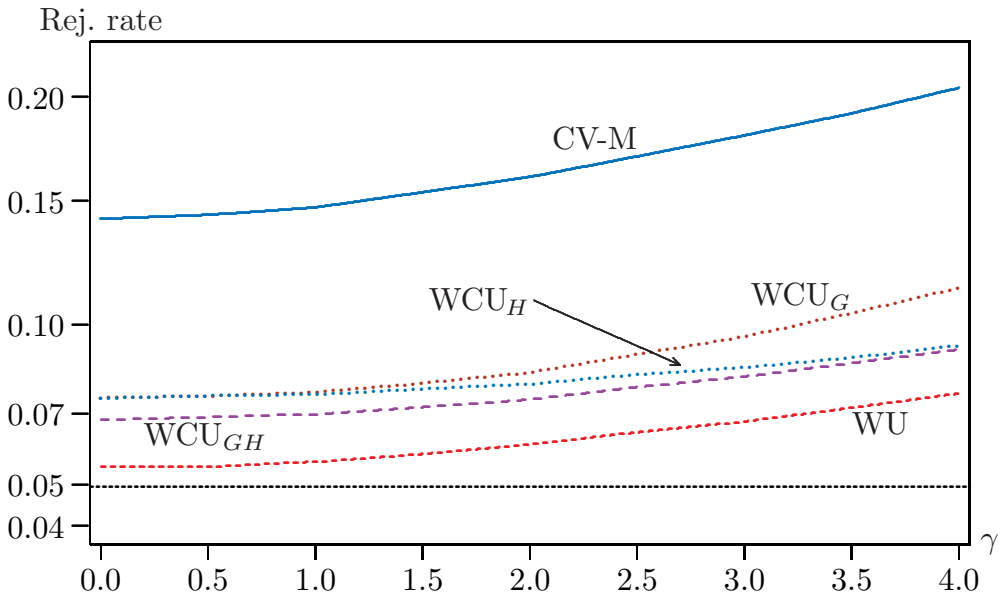
In the experiments of Table 1, all clusters are balanced. In the next set of experiments, with $G = H = 10$, we allow cluster sizes to vary in the G dimension, but not in the H dimension. In order to allow for unbalanced cluster sizes, N_g is determined by a parameter γ, as follows:

$$N_g = \left[N \frac{\exp(\gamma g/G)}{\sum_{j=1}^{G} \exp(\gamma j/G)} \right], \quad g = 1, \ldots, G-1, \tag{23}$$

where $[\cdot]$ denotes the integer part of its argument, and $N_G = N - \sum_{j=1}^{G-1} N_g$. When $\gamma = 0$, every N_g is equal to N/G. As γ increases, cluster sizes become increasingly unbalanced.

Figure 1 shows rejection frequencies for five tests as a function of the parameter γ, which varies between 0 and 4. When $\gamma = 0$, all clusters are of size 400 in the G dimension. When $\gamma = 4$, the smallest cluster in the G dimension has 36 observations and the largest has 1349. The vertical axis has been subjected to a square root transformation so that both large and small rejection frequencies can be shown legibly. It is evident that all the tests perform worse as cluster sizes become more variable. However, the bootstrap tests (all of them based on restricted estimates) always perform very much better than CV-M, that is, using multiway CRVE standard errors and $t(\min(G, H)-1))$ critical values. Among the bootstrap methods, the ordinary wild bootstrap (WR) always performs worst, and WCR$_G$ always performs best, especially when cluster sizes vary a lot.

Figure 2: Rejection frequencies when cluster sizes vary (unrestricted bootstraps)



Notes:
There are 400,000 replications, with $N = 4000$ and $G = H = 10$.
All tests are at 5% nominal level.
CV-M rejection frequencies are based on the $t(9)$ distribution.
All bootstrap tests use the Rademacher distribution with symmetric P values and $B = 399$.

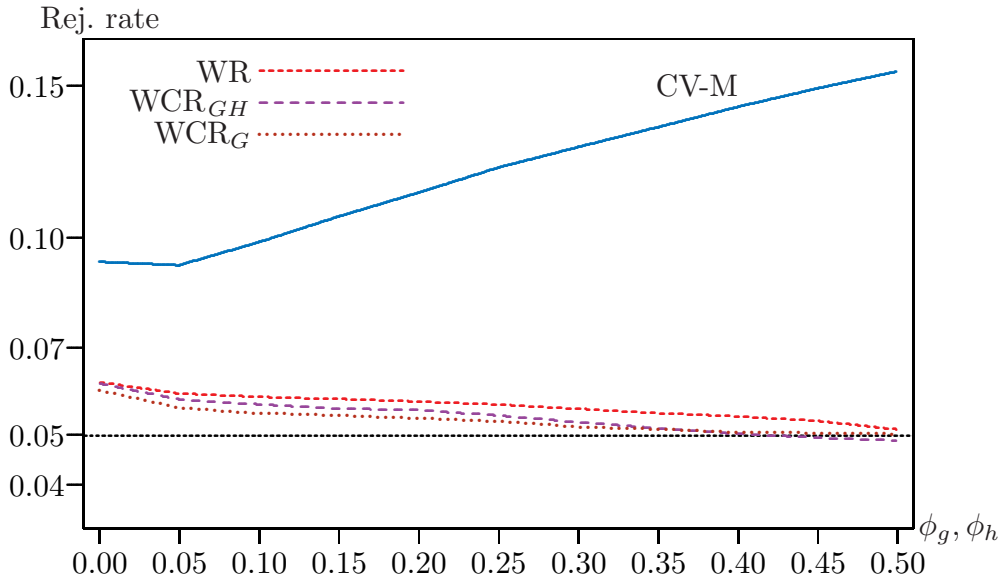
Figure 2 deals with the same experiments as Figure 1. The only difference is that the four bootstrap methods now use unrestricted estimates. The results for WU are almost the same as for WR, although the rejection frequencies are actually very slightly higher. For the other three methods, however, the rejection frequencies are substantially higher when we use unrestricted estimates. In the most extreme case, WCU$_G$ has a rejection frequency of 0.1141 when $\gamma = 4$.

There is nothing unique about Figures 1 and 2. In every case where we compare the performance of wild bootstrap methods that differ only in being based on either restricted or unrestricted estimates and residuals, the latter reject more often (often much more often) than the former. Only for the ordinary wild bootstrap (WR and WU) do the differences tend to be negligible. We conclude that, when bootstrapping models that use a multiway CRVE, it seems to be particularly important to impose the null hypothesis on the bootstrap DGP. In the remainder of the paper, we therefore do not report results for unrestricted bootstrap methods.

The DGP (22) evidently depends on the parameters ρ_g, ρ_h, ϕ_g, and ϕ_h. However, the results in Table 1 and Figures 1 and 2 are all for the same values of those parameters. We therefore performed a number of experiments to see how sensitive the results are to those values. Results for invalid methods like HC$_1$ and CV$_{GH}$ are, of course, quite sensitive to them. Results for CV-M are sometimes fairly sensitive, but we never found a case where results for the bootstrap methods change substantially.

As an example, consider Figure 3, which shows $\phi_g = \phi_h$ on the horizontal axis for balanced clusters with $G = H = 10$. As usual, the vertical axis shows rejection frequencies. The values of ρ_g and ρ_h are 0.05, as before. It is evident that the performance of CV-M deteriorates fairly sharply as the intra-cluster correlations of the regressor increase. In contrast, all of the bootstrap methods

Figure 3: Rejection frequencies as a function of the regressor's intra-cluster correlation



Notes:
There are 400,000 replications, with $N = 4000$ and $G = H = 10$.
All tests are at 5% nominal level.
CV-M rejection frequencies are based on the $t(9)$ distribution.
All bootstrap tests use the Rademacher distribution with symmetric P values and $B = 399$.

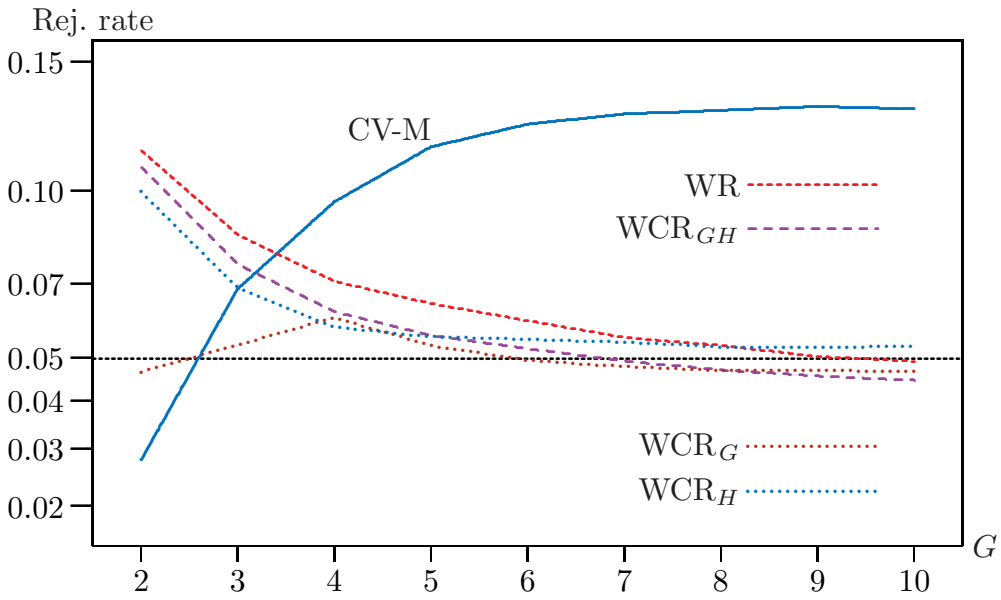
perform better as they increase, with rejection frequencies gradually decreasing from approximately 0.06 to approximately 0.05.

In empirical work, it is not uncommon for one of the two dimensions in which the data are clustered to be quite small. To investigate this situation, we perform a further set of experiments with $N = 4000$ in which $H = 20$ and G varies from 2 to 10. The value of γ is 2, so that the clusters in the G dimension vary in size. When $G = 2$, the smallest is 1075, and the largest is 2925. When $G = 10$, the smallest is 138, and the largest is 843. All clusters in the H dimension have 200 observations.

The results of these experiments, which are shown in Figure 4, may be surprising. The CV-M procedure actually underrejects for $G = 2$, perhaps because the t-statistic is being compared with the $t(1)$ distribution, for which the 5% critical value is 12.71. Results for this extreme case are probably also influenced by the fact that there is at least one negative eigenvalue 21.5% of the time when $G = 2$. The CV-M procedure overrejects more and more severely as G becomes larger, with the rejection frequency starting to drop only very slightly when $G = 10$.

For the various wild bootstrap procedures, we use the Rademacher distribution whenever the number of auxiliary random variables is at least 10 and the 6-point distribution otherwise. This means that we use the latter for WCR$_G$ whenever $G \leq 9$. Most of the bootstrap methods do not perform well when G is very small, although WCR$_G$ does perform remarkably well for $G = 2$ and $G = 3$. However, given that this method uses auxiliary random variables that can take on only 36 or 216 sets of values, these results should probably should not be taken too seriously. All the bootstrap methods perform surprisingly well for the larger values of G. In fact, they all reject between 4% and 6% of the time for $G \geq 6$, while CV-M rejects between 12.5% and 13.1% of the

Figure 4: Rejection frequencies for $H = 20$ and small values of G



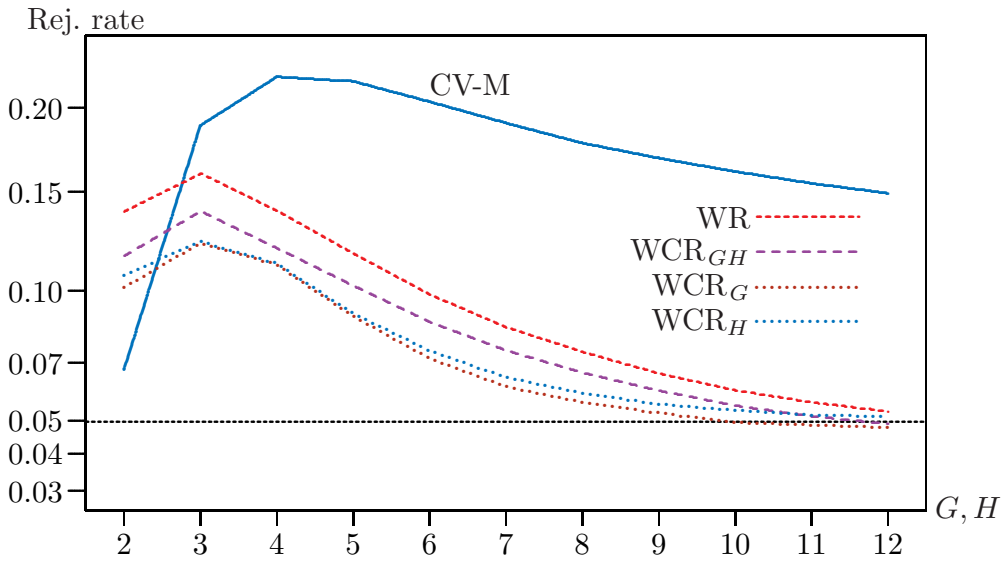
Notes:
There are 400,000 replications, with $N = 4000$, $H = 20$, and G varying from 2 to 10.
All tests are at 5% nominal level.
CV-M rejection frequencies are based on the $t(G-1)$ distribution.
Bootstrap tests use either the Rademacher distribution (if there are 10 or more draws of v^* per bootstrap sample) or the 6-point distribution (if not), with symmetric P values and $B = 399$.

time. It is difficult to say which method should be preferred, since the relative performance of different methods seems to vary with G.

The final set of experiments also involves small values of G and H, but this time $G = H$. Clusters are balanced in the H dimension, but $\gamma = 2$ for the G dimension. Results are shown in Figure 5. The CV-M procedure performs quite well for $G = H = 2$, which, as in the case with $G = 2, H = 20$ in Figure 4, is probably because the t-statistic is compared with the $t(1)$ distribution in this case. However, the CV-M procedure overrejects severely for all other values. The four bootstrap procedures overreject less severely (except for $G = H = 2$), but they improve rapidly once $G > 3$. For all but the largest values of $G = H$, the WCR$_G$ and WCR$_H$ bootstraps perform noticeably better than the WR and WCR$_{GH}$ ones. Note that WCR$_G$ and WCR$_H$ differ only because the cluster sizes are unbalanced in the G dimension and balanced in the H dimension.

Like all simulation results, the ones in this section must be interpreted with caution. The performance of all methods evidently depends on the N_g, the N_h, the $\mathbf{\Omega}$ matrix, and the $\boldsymbol{X}$ matrix. Since any or all of those could be very different from the ones in these experiments, we certainly cannot conclude that bootstrap methods will always work as well as they do here. For a case in which the wild cluster bootstrap works badly with one-way clustering, even when the number of clusters is large, see MacKinnon and Webb (2017, Section 6). Nevertheless, it does seem fairly safe to conclude that if CV-M and several bootstrap methods yield similar P values, the latter can probably be relied on. On the other hand, if CV-M yields a much smaller P value than the better bootstrap methods, the former is almost certainly not reliable.

Figure 5: Rejection frequencies for small values of G and H



Notes:

There are 400,000 replications, with $N = 4000$, and $G = H$ varying from 2 to 12.

All tests are at 5% nominal level.

Clusters in the G dimension vary in size, with $\gamma = 2$. For the H dimension, $N_h = N/H$.

CV-M rejection frequencies are based on the $t(G-1)$ distribution.

Bootstrap tests use either the Rademacher distribution (if there are 10 or more draws of v^* per bootstrap sample) or the 6-point distribution (if not), with symmetric P values and $B = 399$.

6 Empirical Example

To illustrate the implications of the simulation results in the previous section, we consider an empirical example from Nunn and Wantchekon (2011). This paper (NW hereafter) investigated whether current trust levels among different ethnic groups in several African countries are related to historical slave exports. NW studied the relationship between the volume of slave exports and current levels of trust between ethnicities using the following equation:

$$\text{trust}_{iedc} = \boldsymbol{\alpha}_c + \beta \text{exports}_e + \boldsymbol{X}'_{iedc}\boldsymbol{\phi}_1 + \boldsymbol{X}'_d\boldsymbol{\phi}_2 + \boldsymbol{X}'_e\boldsymbol{\phi}_3 + \varepsilon_{iedc}, \tag{24}$$

where i, e, d, and c indicate individual, ethnicity, district, and country, respectively. The subscript notation in (24) follows NW and therefore differs somewhat from what is used elsewhere in this paper. 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 many leading zeros and increase the number of significant digits in the reported coefficients and standard errors.

The principal coefficient of interest in equation (24) is β, which measures the extent to which historical slave exports of a certain ethnicity affect trust levels of 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, and $\boldsymbol{X}'_d$ contains two specific district-level variables which may influence an ethnic group's current levels of trust. These are the degree of ethnic fractionalization in the district, and the proportion of the district population that is of the same ethnic background as the survey respondent. Finally, $\boldsymbol{X}'_e$ contains ethnicity-level variables which

Table 2: OLS Estimates of the Determinants of Trust in Neighbors

Dependent Variable	slave exports		
Trust of Neighbors × 1000	district	country	region
$\hat{\beta}$	−0.6791	−0.6791	−0.6791
One-way CV_H: ethnicity, SE	0.1422	0.1422	0.1422
One-way CV_H: ethnicity, P value	0.0000	0.0000	0.0000
One-way CV_G: geography, SE	0.0822	0.2051	0.1814
One-way CV_G: geography, P value	0.0000	0.0475	0.1338
One-way CV_{GH}: ethnicity×geography, SE	0.0811	0.1426	0.1431
One-way CV_{GH}: ethnicity×geography, P value	0.0000	0.0000	0.0000
Two-way CV-M: ethnicity and geography, SE	0.1451	0.2070	0.1810
Two-way CV-M: ethnicity and geography, t-stat	−4.6809	−3.2809	−3.7527
Two-way CV-M: ethnicity and geography, P value	0.0000	0.0047	0.0133
Bootstrap WR: individual, P value	0.0005	0.0281	0.0384
Bootstrap WCR_{GH}: ethnicity×geography, P value	0.0006	0.0470	0.0538
Bootstrap WCR_H: ethnicity, P value	0.0012	0.0754	0.0832
Bootstrap WCR_G: geography, P value	0.0004	0.1329	0.2539
Number of clusters, G: geography	1257	17	6
Number of clusters, H: ethnicity	185	185	185

Notes:
This example is taken from Nunn and Wantchekon (2011), Table 1, column 1.
Geographical clustering is done at a different level in each column.
All bootstrap P values are symmetric and based on $B = 9999$.
All bootstrap procedures use the Rademacher distribution, except for the WCR_G bootstrap (geography) with clustering at the region level, which uses the 6-point distribution.
A Stata `.do` file to replicate this table may be found at
http://qed.econ.queensu.ca/pub/faculty/mackinnon/two-way-boot/.

aim to control for historical differences between ethnicities, including the degree of colonization.

The trust variable comes from surveys for the Afrobarometer. These surveys were conducted in 2005 and 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. Data on slave exports were obtained from Nunn (2008) and include the number of slave exports from each country, as well as information about the ethnicities of the slaves. The data cover the four major African slave trades from 1400 to 1900, the trans-Atlantic, Indian Ocean, Red Sea, and trans-Saharan, although only data from the trans-Atlantic and Indian Ocean slave trades were used in NW. After cleaning the data, the final sample consists of $N = 21,702$ observations.

Table 2 reproduces and extends the results from NW's Table 1 (column 1). NW used three different variables as their key regressor. We focus on exports, but the results for exports/area and exports/(historical population) follow a broadly similar pattern. Table 2 presents the results from the OLS regression specified in equation (24). To illustrate the differences between CV-M and the bootstrap procedures, we consider different levels of clustering in the geographic dimension, where

NW simply clustered by district. We do not mean to imply that this is incorrect. However, it is interesting to see what would happen if the geographic clustering dimension were chosen differently. We choose it in two alternative ways, once clustering by country and once clustering by investment region.[3] Thus, the number of clusters in the geographic dimension (which can be thought of as G) is either 1257 districts, 17 countries, or 6 regions. The number of clusters in the ethnic clustering dimension (which can be thought of as H) is always 185 ethnicities.

The first row of results in Table 2 presents the coefficient estimate. Following it, the top panel presents standard errors and P values clustered by three different one-way clustering variables, specifically, ethnicity, geography, and the intersection of ethnicity and geography. The majority of these P values are quite small, with only the P values for geographical clustering at either the country or region level being insignificant at the 1% level.

The second panel of Table 2 presents two-way clustered standard errors and P values based on the $t(\min(G,H)-1)$ distribution with clustering by both ethnicity and geography. The first column in this panel reports results similar to those of NW, since they used district as their measure of geography.[4] The second and third columns report standard errors and P values using country and region as the measure of geography. As the geographical clustering variable becomes coarser, the multiway clustering P value becomes larger, although all three P values are quite small. It may seem odd that the P value in the third column is larger than the one in the second column, even though the t-statistic is larger in absolute value (-3.7527 versus -3.2809). That is because the test in the second column uses the $t(16)$ distribution and the one in the third column uses $t(5)$.

The most interesting part of Table 2 is the third panel, which shows the various bootstrap P values. In every case, the bootstrap P value increases with the coarseness of the geographical clustering variable. All of the bootstrap P values are significant at the 1% level when clustering by district in the geography dimension as in NW. However, when clustering instead by country or region in the geography dimension, several of the bootstrap P values are not significant at the 5% or even the 10% level. In fact, all of the bootstrap P values are larger than the associated CV-M P values, even when there is a large number of clusters in both dimensions.

The WCR_G (geography) P value increases dramatically across columns, from 0.0004 when clustering by district to 0.2539 when clustering by region. The WCR_H (ethnicity) P value also increases across columns, but less dramatically. Given the often severe overrejection by the CV-M procedure that is evident in the simulations of Section 5, coupled with the generally good performance of the bootstrap methods, especially WCR_G and WCR_H, it seems that the evidence against the null hypothesis is not at all strong when geographic clustering is by country or region.

7 Conclusion

In this paper, we obtain two important results. In Section 3, we prove that the multiway cluster-robust variance estimator (CRVE) is asymptotically valid for the case of two-dimensional clustering under precisely stated conditions which limit the extent of cluster size heterogeneity and the rates

[3] The 17 countries were further grouped into 6 investment regions based on their proximity to each other as well as their shared economic and political ties. The aggregation of countries to investment regions is as follows: Francophone West Africa: *Benin, Mali, Senegal*; Nigeria (region): *Nigeria*; East Africa: *Kenya, Tanzania, Uganda*; Southern Africa (excluding South Africa): *Botswana, Madagascar, Mozambique, Malawi, Namibia, Zambia, Zimbabwe*; South Africa (region): *Lesotho, South Africa*; Other West Africa: *Ghana*. Further details about the methodology can be found at http://www.riscura.com/brightafrica/segmenting-africa/.

[4] NW did not report a P value for their estimate. Care should be taken using the `cgmreg` Stata command, as it will calculate a P value using the $t(n-k)$ distribution instead of the $t(\min(G,H)-1)$ distribution which is used throughout this paper.

at which cluster sizes can grow with the sample size. Our conditions also imply that the numbers of clusters in both dimensions must increase as the sample size increases, and they restrict the sizes of the largest clusters in each dimension.

In Section 4, we propose eight bootstrap methods, which appear to be the first ones for least squares regression with multiway cluster-robust standard errors. Two of these methods simply combine the multiway CRVE with the ordinary wild bootstrap, using either restricted or unrestricted estimates, and the other six combine it with the wild cluster bootstrap with the residuals clustered according either to one of the two dimensions or to their intersection. None of these bootstraps is capable of matching the two-dimensional nature of the clustered disturbances. Despite this, we prove that they all yield valid inferences asymptotically. This happens because cluster-robust t-statistics are asymptotically pivotal.

In Section 5, we provide extensive simulation evidence to show that the conventional approach of comparing multiway cluster-robust t statistics to the t distribution with degrees of freedom equal to one less than the minimum of the numbers of clusters in each dimension can lead to serious overrejection, especially when the number of clusters in either dimension is small or cluster sizes vary substantially. In almost all the cases that we study, bootstrap methods based on restricted estimates yield more accurate inferences than this conventional approach. They are generally much more accurate when the conventional approach works poorly. Bootstrap methods based on unrestricted estimates generally also outperform the conventional approach, but they can be substantially inferior to ones based on restricted estimates.

The best method often seems to be to use a multiway CRVE variant of the one-dimensional wild cluster bootstrap for the dimension with the smallest number of clusters. This ensures that the bootstrap DGP preserves the within-cluster correlation for the dimension with the clusters that are, on average, largest. It is essential that the same procedure for calculating multiway CRVE standard errors be used for both the original data and the bootstrap samples. However, since the performance of all methods evidently depends on the features of the model and dataset, it would be premature to conclude that any one of the restricted bootstrap methods should necessarily be chosen over the others. In most cases, all bootstrap methods tend to yield similar inferences.

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 results can change substantially as the level of clustering in one of two dimensions changes. Moreover, the P values, especially the bootstrap P values, tend to become larger as the number of clusters in one of the dimensions is reduced because clustering in that dimension is coarser. This strongly suggests that the conventional results may be unreliable, especially when there are few clusters in either dimension.

Appendix: Proofs of main results

The next two subsections contain the proofs of our two main results. Throughout, C denotes a generic finite constant, which may take different values in different places.

A.1 Proof of Theorem 1

As usual, we give the proof conditional on $\boldsymbol{X}$, which is sufficient because the limits do not depend on $\boldsymbol{X}$. Thus, we may treat $\boldsymbol{X}$ as if it were non-random.

Proof of (16). We write the left-hand side of equation (16) as $(\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-1/2} \sum_{i=1}^N z_i$, where $z_i = N^{-1} \boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} \boldsymbol{X}_i^\top u_i$ (ignoring the subscript N for the triangular array notation), and apply the

central limit theorem of Romano and Wolf (2000) for m-dependent processes. Recall that z_i is m-dependent for some $m \geq 0$, that may depend on the sample size, if the vector $(z_1, \dots, z_j)$ is independent of $(z_{j+n}, z_{j+n+1}, \dots)$ whenever $n > m$. We note that m is a characteristic of the data and is not necessarily equal to the cluster size chosen by the econometrician. Specifically, it could be the case that $m = 0$ even though $N_{gh} \neq 1$, as would happen if choosing to cluster when the data are in fact independent. However, by Assumption 1, we have the upper bound $m \leq \sup_g N_g + \sup_h N_h$.

Thus, we verify the conditions of Romano and Wolf (2000), which involve the quantities

$$B_{N,j}^2 = \text{Var}\left(\sum_{i=1}^{j} z_i\right) \quad \text{and} \quad B_N^2 = B_{N,N}^2 = \text{Var}\left(\sum_{i=1}^{N} z_i\right) = \boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a}.$$

Specifically, setting their $\gamma = 0$, we verify that (i) $\sup_i \mathrm{E}|z_i|^{2+\lambda} \leq \Delta_N$, (ii) $B_{N,j}^2/j \leq K_N$ for all $j \geq m$, (iii) $B_N^2/N \geq L_N$, (iv) $K_N/L_N = O(1)$, (v) $\Delta_N/L_N^{1+\lambda/2} = O(1)$, and (vi) $m^{2+2/\lambda}/N \to 0$.

First, note that if $m = 0$ all the conditions are easily seen to be satisfied, so we assume that $m \geq 1$. By Assumptions 1 and 2, condition (i) holds with $\Delta_N = CN^{-2-\lambda}$, and

$$B_{N,j}^2 = N^{-2}\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} j^2 \boldsymbol{\Gamma}_j \boldsymbol{Q}_N^{-1} \boldsymbol{a} = j^2 N^{-2} \boldsymbol{a}^\top \boldsymbol{V}_j \boldsymbol{a}(1 + o(1)).$$

We also note that $\boldsymbol{a}^\top \boldsymbol{V}_j \boldsymbol{a} = C_j m j^{-1}$ for some constant C_j that depends on j, but is bounded and bounded away from zero, uniformly in j. We therefore find that condition (ii) is satisfied with $K_N = CmN^{-2}$ and condition (iii) with $L_N = CmN^{-2}$. It follows that condition (iv) is trivially satisfied and that condition (v) holds if $m^{-1-\lambda/2} = O(1)$, which is satisfied because $m \geq 1$. Finally, because $m \leq \sup_g N_g + \sup_h N_h$, condition (vi) holds by Assumption 3 and the c_r-inequality. Thus, we have verified the conditions of Romano and Wolf (2000), which then proves (16).

Proof of (17). We start with the decomposition

$$\frac{\boldsymbol{a}^\top \hat{\boldsymbol{V}} \boldsymbol{a}}{\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a}} - 1 = (\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-1} \boldsymbol{a}^\top (\hat{\boldsymbol{V}} - \boldsymbol{V}_N)\boldsymbol{a} = (\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-1} \boldsymbol{a}^\top \Big(\boldsymbol{A}_{1N} - \boldsymbol{A}_{2N} - \boldsymbol{A}_{2N}^\top + \boldsymbol{A}_{3N}\Big)\boldsymbol{a},$$

where

$$\begin{aligned}
\boldsymbol{A}_{1N} =& \frac{1}{N^2} \boldsymbol{Q}_N^{-1} \Bigg(\sum_{g=1}^{G} \boldsymbol{X}_g^\top \boldsymbol{u}_g \boldsymbol{u}_g^\top \boldsymbol{X}_g + \sum_{h=1}^{H} \boldsymbol{X}_h^\top \boldsymbol{u}_h \boldsymbol{u}_h^\top \boldsymbol{X}_h - \sum_{g=1}^{G} \sum_{h=1}^{H} \boldsymbol{X}_{gh}^\top \boldsymbol{u}_{gh} \boldsymbol{u}_{gh}^\top \boldsymbol{X}_{gh} \Bigg) \boldsymbol{Q}_N^{-1} - \boldsymbol{V}_N, \\
\boldsymbol{A}_{2N} =& \frac{1}{N^2} \boldsymbol{Q}_N^{-1} \Bigg(\sum_{g=1}^{G} \boldsymbol{X}_g^\top \boldsymbol{u}_g (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \boldsymbol{X}_g^\top \boldsymbol{X}_g + \sum_{h=1}^{H} \boldsymbol{X}_h^\top \boldsymbol{u}_h (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \boldsymbol{X}_h^\top \boldsymbol{X}_h \\
& - \sum_{g=1}^{G} \sum_{h=1}^{H} \boldsymbol{X}_{gh}^\top \boldsymbol{u}_{gh} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \boldsymbol{X}_{gh}^\top \boldsymbol{X}_{gh} \Bigg) \boldsymbol{Q}_N^{-1} = \boldsymbol{A}_{2N,1} + \boldsymbol{A}_{2N,2} + \boldsymbol{A}_{2N,3}, \text{ and} \\
\boldsymbol{A}_{3N} =& \frac{1}{N^2} \boldsymbol{Q}_N^{-1} \Bigg(\sum_{g=1}^{G} \boldsymbol{X}_g^\top \boldsymbol{X}_g (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \boldsymbol{X}_g^\top \boldsymbol{X}_g + \sum_{h=1}^{H} \boldsymbol{X}_h^\top \boldsymbol{X}_h (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \boldsymbol{X}_h^\top \boldsymbol{X}_h \\
& - \sum_{g=1}^{G} \sum_{h=1}^{H} \boldsymbol{X}_{gh}^\top \boldsymbol{X}_{gh} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \boldsymbol{X}_{gh}^\top \boldsymbol{X}_{gh} \Bigg) \boldsymbol{Q}_N^{-1} = \boldsymbol{A}_{3N,1} + \boldsymbol{A}_{3N,2} + \boldsymbol{A}_{3N,3}.
\end{aligned}$$

Thus, we need to show that $(\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-1} \boldsymbol{a}^\top \boldsymbol{A}_{mN} \boldsymbol{a} \xrightarrow{P} 0$ for $m = 1, 2, 3$, or, since $(\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-1}$ is at most of order $O(N)$, we show that $N\boldsymbol{a}^\top \boldsymbol{A}_{mN} \boldsymbol{a} \xrightarrow{P} 0$ for $m = 1, 2, 3$

To prove the result for $m = 1$, we note that $\mathrm{E}(\boldsymbol{A}_{1N}) = 0$ and prove convergence in mean-square. Let C_2 denote all the pairs (i_1, i_2) such that u_{i_1} and u_{i_2} have at least one cluster in common and let C_4 denote all quadruplets (i_1, i_2, i_3, i_4) such that u_{i_1}, u_{i_2}, u_{i_3}, and u_{i_4} all have at least one cluster in common. Using the notation $\boldsymbol{P}_{12} = N^{-2}\boldsymbol{Q}_N^{-1}\boldsymbol{X}_{i_1}^\top\boldsymbol{X}_{i_2}\boldsymbol{Q}_N^{-1}$ and $\boldsymbol{P}_{1234} = N^{-4}\boldsymbol{Q}_N^{-1}\boldsymbol{X}_{i_1}^\top\boldsymbol{X}_{i_2}\boldsymbol{Q}_N^{-2}\boldsymbol{X}_{i_3}^\top\boldsymbol{X}_{i_4}\boldsymbol{Q}_N^{-1}$, we then find, by independence of errors that do not have at least one cluster in common, that

$$\begin{aligned}
\mathrm{E}(\boldsymbol{A}_{1N}^2) &= \sum_{i_1,i_2,i_3,i_4 \in C_2 \times C_2} \mathrm{E}(u_{i_1}u_{i_2}u_{i_3}u_{i_4})\boldsymbol{P}_{1234} - \Bigg(\sum_{i_1,i_2 \in C_2} \mathrm{E}(u_{i_1}u_{i_2})\boldsymbol{P}_{12}\Bigg)^2 \\
&= \sum_{i_1,i_2,i_3,i_4 \in C_4} \mathrm{E}(u_{i_1}u_{i_2}u_{i_3}u_{i_4})\boldsymbol{P}_{1234} + \sum_{i_1,i_2,i_3,i_4 \in C_2 \times C_2 \setminus C_4} \mathrm{E}(u_{i_1}u_{i_2})\mathrm{E}(u_{i_3}u_{i_4})\boldsymbol{P}_{1234} \\
&\quad - \Bigg(\sum_{i_1,i_2 \in C_2} \mathrm{E}(u_{i_1}u_{i_2})\boldsymbol{P}_{12}\Bigg)^2 \\
&= \sum_{i_1,i_2,i_3,i_4 \in C_4} \mathrm{E}(u_{i_1}u_{i_2}u_{i_3}u_{i_4})\boldsymbol{P}_{1234} - \sum_{i_1,i_2,i_3,i_4 \in C_4} \mathrm{E}(u_{i_1}u_{i_2})\mathrm{E}(u_{i_3}u_{i_4})\boldsymbol{P}_{1234}.
\end{aligned}$$

Using Assumptions 1 and 2, the right-hand side is $O(N^{-4}(\sup_g N_g + \sup_h N_h)^4)$, so that $N\boldsymbol{a}^\top\boldsymbol{A}_{1N}\boldsymbol{a} = O_P(N^{-1}(\sup_g N_g + \sup_h N_h)^2) = o_P(1)$ by Assumption 3.

Next, for $m = 2$, the proofs for the three terms are identical, so we give only the first one. Here we use the fact that $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top\boldsymbol{X}_g^\top\boldsymbol{X}_g\boldsymbol{Q}_N^{-1}\boldsymbol{a}$ is a scalar and find that

$$\boldsymbol{a}^\top\boldsymbol{A}_{2N,1}\boldsymbol{a} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top\frac{1}{N^2}\sum_{g=1}^{G}\boldsymbol{X}_g^\top\boldsymbol{X}_g\boldsymbol{Q}_N^{-1}\boldsymbol{a}\boldsymbol{a}^\top\boldsymbol{Q}_N^{-1}\boldsymbol{X}_g^\top\boldsymbol{u}_g.$$

We first note that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(\|\boldsymbol{V}_N\|^{1/2}) = O_P(N^{-1/2}(\sup_g N_g + \sup_h N_h)^{1/2})$. Next, by the c_r-inequality,

$$\sup_{1\le g\le G} N_g^{-4}\mathrm{E}(\|\boldsymbol{X}_g^\top\boldsymbol{u}_g\|^4) \le \sup_{1\le g\le G} N_g^{-1}\sum_{i=1}^{N_g}\|\boldsymbol{X}_{g,i}\|^4\mathrm{E}|u_{g,i}|^4 \le \sup_{1\le i\le N_g, 1\le g\le G}\|\boldsymbol{X}_{g,i}\|^4\mathrm{E}|u_{g,i}|^4 \le C \tag{A.1}$$

by Assumptions 1 and 2 because $\lambda \ge 2$, and similarly

$$\sup_{1\le g\le G} N_g^{-4}\|\boldsymbol{X}_g^\top\boldsymbol{X}_g\|^4 \le C \tag{A.2}$$

by Assumption 2. Then it follows that

$$\mathrm{E}\left(\left\|\frac{1}{N^2}\sum_{g=1}^{G}\boldsymbol{X}_g^\top\boldsymbol{X}_g\boldsymbol{Q}_N^{-1}\boldsymbol{a}\boldsymbol{a}^\top\boldsymbol{Q}_N^{-1}\boldsymbol{X}_g^\top\boldsymbol{u}_g\right\|^2\right) \le \frac{1}{N^4}\|\boldsymbol{Q}_N^{-1}\|^4\sum_{g=1}^{G}\|\boldsymbol{X}_g^\top\boldsymbol{X}_g\|^2\mathrm{E}(\|\boldsymbol{X}_g^\top\boldsymbol{u}_g\|^2),$$

which is $O(N^{-3}\sup_g N_g^3)$ using also Assumption 2 and the Cauchy-Schwarz inequality. This yields the bound $N\boldsymbol{a}^\top\boldsymbol{A}_{2N,1}\boldsymbol{a} = O_P(N^{-1}(\sup_g N_g + \sup_h N_h)^{1/2}\sup_g N_g^{3/2}) = o_P(1)$ under Assumption 3. Finally, the proof for $m = 3$ is nearly identical to that for $m = 2$, using, for the first term, the bound

$$\begin{aligned}
\|N\boldsymbol{a}^\top\boldsymbol{A}_{3N,1}\boldsymbol{a}\| &\le \frac{1}{N}\|\boldsymbol{Q}_N^{-1}\|^2\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2\sum_{g=1}^{G}\|\boldsymbol{X}_g^\top\boldsymbol{X}_g\|^2 \\
&= O_P\left(\left(\frac{\sup_{1\le g\le G} N_g(\sup_{1\le g\le G} N_g + \sup_{1\le h\le H} N_h)}{N}\right)\right) = o_P(1).
\end{aligned}$$

Proof of (18). Follows from (16), (17), and Slutsky's Theorem.

A.2 Proof of Theorem 2

For the proof of Theorem 2 we define the quantities

$$\ddot{\boldsymbol{V}}_N = \boldsymbol{Q}_N^{-1}\ddot{\boldsymbol{\Gamma}}_N\boldsymbol{Q}_N^{-1} \quad \text{and} \quad \ddot{\boldsymbol{\Gamma}}_N = N^{-2}\boldsymbol{X}^\top \mathrm{E}^*(\boldsymbol{u}^*\boldsymbol{u}^{*\top})\boldsymbol{X},$$

which are interpreted as the bootstrap true values, see also (6) and (7). We note that, by identical steps to those in the proofs of (16) and (17), under the assumptions of Theorem 2 it holds that

$$\frac{\boldsymbol{a}^\top(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)}{(\boldsymbol{a}^\top\boldsymbol{V}_N\boldsymbol{a})^{1/2}} = O_P(1) \quad \text{and} \quad \frac{\boldsymbol{a}^\top\ddot{\boldsymbol{V}}_N\boldsymbol{a}}{\boldsymbol{a}^\top\bar{\boldsymbol{V}}_N\boldsymbol{a}} \xrightarrow{P} 1. \tag{A.3}$$

Proof of (19). Proceeding as in the proof of (16) we write the left-hand side of equation (19) as $(\boldsymbol{a}^\top\bar{\boldsymbol{V}}_N\boldsymbol{a})^{-1/2}\sum_{i=1}^N z_i^*$, where $z_i^* = N^{-1}\boldsymbol{a}^\top\boldsymbol{Q}_N^{-1}\boldsymbol{X}_i^\top u_i^*$, and apply the central limit theorem of Romano and Wolf (2000). We note that m under the bootstrap measure is no longer the same as in the original data, although it is bounded from above by the m for the original data. For example, if the bootstrap data is constructed using the WB, then $m = 0$ under the bootstrap measure. In any case, we will continue to use the notation m since this detail is not important for the proof. We need to verify that the conditions of Romano and Wolf (2000) are satisfied under the bootstrap measure with probability converging to one. To this end, we define

$$B_{N,j}^2 = \mathrm{Var}^*\left(\sum_{i=1}^j z_i^*\right) \quad \text{and} \quad B_N^2 = \mathrm{Var}^*\left(\sum_{i=1}^N z_i^*\right),$$

which are now random variables. We then verify that (i) $\sup_i \mathrm{E}^*|z_i|^{2+\lambda} = O_P(\Delta_N)$, (ii) $B_{N,j}^2/j = O_P(K_N)$ for all $j \geq m$, (iii) $(B_N^2/N)^{-1} = O_P(L_N)$, (iv) $K_N/L_N = O(1)$, (v) $\Delta_N/L_N^{1+\lambda/2} = O(1)$, and (vi) $m^{2+2/\lambda}/N \to 0$.

First, because $\mathrm{E}^*|v^*|^{2+\lambda} < \infty$, condition (i) holds as in the proof of (16). Next, we find that

$$B_{N,j}^2 = N^{-2}\boldsymbol{a}^\top\boldsymbol{Q}_N^{-1}j^2\ddot{\boldsymbol{\Gamma}}_j\boldsymbol{Q}_N^{-1}\boldsymbol{a} = j^2N^{-2}\boldsymbol{a}^\top\bar{\boldsymbol{V}}_j\boldsymbol{a}(1 + o_P(1))$$

using Assumption 2 and (A.3). As before, $\boldsymbol{a}^\top\bar{\boldsymbol{V}}_j\boldsymbol{a} = C_j m j^{-1}$ for some constant C_j that depends on j, but is bounded and bounded away from zero, uniformly in j. Therefore, conditions (ii)–(vi) follow in exactly the same way as in the proof of (16), which proves (19).

Proof of (20). We note that $\boldsymbol{X}_g^\top\hat{\boldsymbol{u}}_g^* = \boldsymbol{X}_g^\top\boldsymbol{u}_g^* - \boldsymbol{X}_g^\top\boldsymbol{X}_g(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})$, and similarly for $\boldsymbol{X}_h^\top\hat{\boldsymbol{u}}_h^*$ and $\boldsymbol{X}_{gh}^\top\hat{\boldsymbol{u}}_{gh}^*$, which implies the decomposition

$$(\boldsymbol{a}^\top\ddot{\boldsymbol{V}}_N\boldsymbol{a})^{-1}\boldsymbol{a}^\top(\hat{\boldsymbol{V}}^* - \ddot{\boldsymbol{V}}_N)\boldsymbol{a} = (\boldsymbol{a}^\top\ddot{\boldsymbol{V}}_N\boldsymbol{a})^{-1}\boldsymbol{a}^\top\Big(\boldsymbol{B}_{1N}^* - \boldsymbol{B}_{2N}^* - \boldsymbol{B}_{2N}^{*\top} + \boldsymbol{B}_{3N}^*\Big)\boldsymbol{a},$$

where

$$\boldsymbol{B}^*_{1N} = \frac{1}{N^2}\boldsymbol{Q}_N^{-1}\Big(\sum_{g=1}^{G}\boldsymbol{X}_g^\top \boldsymbol{u}_g^*\boldsymbol{u}_g^{*\top}\boldsymbol{X}_g + \sum_{h=1}^{H}\boldsymbol{X}_h^\top \boldsymbol{u}_h^*\boldsymbol{u}_h^{*\top}\boldsymbol{X}_h - \sum_{g=1}^{G}\sum_{h=1}^{H}\boldsymbol{X}_{gh}^\top \boldsymbol{u}_{gh}^*\boldsymbol{u}_{gh}^{*\top}\boldsymbol{X}_{gh}\Big)\boldsymbol{Q}_N^{-1} - \ddot{\boldsymbol{V}}_N,$$

$$\begin{aligned}\boldsymbol{B}^*_{2N} =& \frac{1}{N^2}\boldsymbol{Q}_N^{-1}\Big(\sum_{g=1}^{G}\boldsymbol{X}_g^\top \boldsymbol{u}_g^*(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})^\top\boldsymbol{X}_g^\top\boldsymbol{X}_g + \sum_{h=1}^{H}\boldsymbol{X}_h^\top \boldsymbol{u}_h^*(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})^\top\boldsymbol{X}_h^\top\boldsymbol{X}_h \\ &- \sum_{g=1}^{G}\sum_{h=1}^{H}\boldsymbol{X}_{gh}^\top \boldsymbol{u}_{gh}^*(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})^\top\boldsymbol{X}_{gh}^\top\boldsymbol{X}_{gh}\Big)\boldsymbol{Q}_N^{-1} = \boldsymbol{B}^*_{2N,1} + \boldsymbol{B}^*_{2N,2} + \boldsymbol{B}^*_{2N,3}, \text{ and}\end{aligned}$$

$$\begin{aligned}\boldsymbol{B}^*_{3N} =& \frac{1}{N^2}\boldsymbol{Q}_N^{-1}\Big(\sum_{g=1}^{G}\boldsymbol{X}_g^\top \boldsymbol{X}_g(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})^\top\boldsymbol{X}_g^\top\boldsymbol{X}_g + \sum_{h=1}^{H}\boldsymbol{X}_h^\top \boldsymbol{X}_h(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})^\top\boldsymbol{X}_h^\top\boldsymbol{X}_h \\ &- \sum_{g=1}^{G}\sum_{h=1}^{H}\boldsymbol{X}_{gh}^\top \boldsymbol{X}_{gh}(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})^\top\boldsymbol{X}_{gh}^\top\boldsymbol{X}_{gh}\Big)\boldsymbol{Q}_N^{-1} = \boldsymbol{B}^*_{3N,1} + \boldsymbol{B}^*_{3N,2} + \boldsymbol{B}^*_{3N,3}.\end{aligned}$$

Using this decomposition together with (A.3) and the fact that $(\boldsymbol{a}^\top\ddot{\boldsymbol{V}}\boldsymbol{a})^{-1}$ is at most of order $O_P(N)$, it is sufficient to prove that $N\boldsymbol{a}^\top\boldsymbol{B}^*_{mN}\boldsymbol{a} = o_{P^*}(1)$, in probability, for $m = 1, 2, 3$. The proofs for each term roughly follow those for the corresponding term in the proof of (17).

To prove the result for $m = 1$, we note that $\mathrm{E}^*(\boldsymbol{B}^*_{1N}) = 0$ by definition of $\ddot{\boldsymbol{V}}_N$, and prove convergence in mean-square. Let C_2 denote all the pairs (i_1, i_2) such that $u^*_{i_1}$ and $u^*_{i_2}$ have at least one cluster in common under the bootstrap measure (for example, under the WB measure, $C_2 = \{(i_1, i_2) : i_1 = i_2\}$) and let C_4 denote all quadruplets (i_1, i_2, i_3, i_4) such that $u^*_{i_1}$, $u^*_{i_2}$, $u^*_{i_3}$, and $u^*_{i_4}$ all have at least one cluster in common under the bootstrap measure. Recalling the notation $\boldsymbol{P}_{12}$ and $\boldsymbol{P}_{1234}$, we then find that, by independence of the bootstrap auxiliary draws for bootstrap errors that do not have at least one cluster in common under the bootstrap measure,

$$\begin{aligned}\mathrm{E}^*(\boldsymbol{B}^{*2}_{1N}) &= \sum_{i_1,i_2,i_3,i_4\in C_4}\mathrm{E}^*(u^*_{i_1}u^*_{i_2}u^*_{i_3}u^*_{i_4})\boldsymbol{P}_{1234} - \sum_{i_1,i_2,i_3,i_4\in C_4}\mathrm{E}^*(u^*_{i_1}u^*_{i_2})\mathrm{E}^*(u^*_{i_3}u^*_{i_4})\boldsymbol{P}_{1234} \\ &= \sum_{i_1,i_2,i_3,i_4\in C_4}(\eta_4 - 1)\ddot{u}_{i_1}\ddot{u}_{i_2}\ddot{u}_{i_3}\ddot{u}_{i_4}\boldsymbol{P}_{1234},\end{aligned}$$

where $\eta_4 = \mathrm{E}^*(v^{*4}) < \infty$ because $\lambda \geq 2$. Using the decomposition $\ddot{u}_i = u_i + \boldsymbol{X}_i^\top(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ and noting that the summation over $(i_1, i_2, i_3, i_4) \in C_4$ contains at most $(\sup_g N_g + \sup_h N_h)^4$ elements, it follows that the right-hand side is $O_P(N^{-4}(\sup_g N_g + \sup_h N_h)^4)$, so that $N\boldsymbol{a}^\top\boldsymbol{B}^*_{1N}\boldsymbol{a} = O_{P^*}(N^{-1}(\sup_g N_g + \sup_h N_h)^2) = o_{P^*}(1)$, in probability, by Assumption 3.

Next, for $m = 2$, the proofs for the three terms are again identical, so we give only the first one. Using the fact that $(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})^\top\boldsymbol{X}_g^\top\boldsymbol{X}_g\boldsymbol{Q}_N^{-1}\boldsymbol{a}$ is a scalar, we find that

$$\boldsymbol{a}^\top\boldsymbol{B}^*_{2N,1}\boldsymbol{a} = (\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})^\top\frac{1}{N^2}\sum_{g=1}^{G}\boldsymbol{X}_g^\top\boldsymbol{X}_g\boldsymbol{Q}_N^{-1}\boldsymbol{a}\boldsymbol{a}^\top\boldsymbol{Q}_N^{-1}\boldsymbol{X}_g^\top\boldsymbol{u}_g^*,$$

where, by (19), $\|\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}}\| = O_{P^*}(\|\bar{\boldsymbol{V}}_N\|^{1/2}) = O_{P^*}(N^{-1/2}(\sup_g N_g + \sup_h N_h)^{1/2})$, in probability. We then use $\ddot{\boldsymbol{u}}_g = \boldsymbol{u}_g + \boldsymbol{X}_g^\top(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ and the c_r-inequality to find that

$$\sup_{1\leq g\leq G}N_g^{-4}\|\boldsymbol{X}_g^\top\ddot{\boldsymbol{u}}_g\|^4 \leq 2^3\sup_{1\leq g\leq G}N_g^{-4}\|\boldsymbol{X}_g^\top\boldsymbol{u}_g\|^4 + 2^3\|\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^4\sup_{1\leq g\leq G}N_g^{-4}\|\boldsymbol{X}_g^\top\boldsymbol{X}_g\|^4 = O_P(1)$$

using (A.1), (A.2), and (A.3), and hence

$$\sup_{1\leq g\leq G}N_g^{-4}\mathrm{E}^*\|\boldsymbol{X}_g^\top\boldsymbol{u}_g^*\|^4 \leq \sup_{1\leq g\leq G}N_g^{-4}\|\boldsymbol{X}_g^\top\ddot{\boldsymbol{u}}_g\|^4\mathrm{E}^*|v^*|^4 = O_P(1). \tag{A.4}$$

It follows that

$$\mathrm{E}^*\left(\left\|\frac{1}{N^2}\sum_{g=1}^{G}\boldsymbol{X}_g^\top\boldsymbol{X}_g\boldsymbol{Q}_N^{-1}\boldsymbol{a}\boldsymbol{a}^\top\boldsymbol{Q}_N^{-1}\boldsymbol{X}_g^\top\boldsymbol{u}_g^*\right\|^2\right) \leq \frac{1}{N^4}\|\boldsymbol{Q}_N^{-1}\|^4\sum_{g=1}^{G}\|\boldsymbol{X}_g^\top\boldsymbol{X}_g\|^2\mathrm{E}^*(\|\boldsymbol{X}_g^\top\boldsymbol{u}_g^*\|^2)$$
$$= O_P\left(N^{-3}\sup_{1\leq g\leq G}N_g^3\right)$$

using the Cauchy-Schwarz inequality, (A.2), and (A.4). This yields the bound $N\boldsymbol{a}^\top\boldsymbol{B}_{2N,1}^*\boldsymbol{a} = O_{P^*}(N^{-1}(\sup_g N_g + \sup_h N_h)^{1/2}\sup_g N_g^{3/2}) = o_{P^*}(1)$, in probability, under Assumption 3. Finally, the proof for $m = 3$ is nearly identical to that for $m = 2$, using, for the first term, the bound

$$\|N\boldsymbol{a}^\top\boldsymbol{B}_{3N,1}^*\boldsymbol{a}\| \leq \frac{1}{N}\|\boldsymbol{Q}_N^{-1}\|^2\|\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}}\|^2\sum_{g=1}^{G}\|\boldsymbol{X}_g^\top\boldsymbol{X}_g\|^2$$
$$= O_{P^*}\left(\left(\frac{\sup_{1\leq g\leq G}N_g(\sup_{1\leq g\leq G}N_g + \sup_{1\leq h\leq H}N_h)}{N}\right)\right) = o_{P^*}(1).$$

Proof of (21). Follows from (19), (20), and Slutsky's Theorem.

References

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*, to appear.

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 (2017). Validity of wild bootstrap inference with clustered errors. QED Working Paper 1383, Queen's University, Department of Economics.

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

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.

Menzel, K. (2017). Bootstrap with clustering in two or more dimensions. ArXiv e-prints, New York University.

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 (2017). Small sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics 35*, to appear.

Romano, J. P. and M. Wolf (2000). A more general central limit theorem for m-dependent random variables with unbounded m. *Statistics & Probability Letters 47*, 115–124.

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

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