Djogbenou, Antoine; MacKinnon, James G.; Nielsen, Morten Ørregaard

Working Paper

Validity of Wild Bootstrap Inference with Clustered Errors

Queen's Economics Department Working Paper, No. 1383

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

Suggested Citation: Djogbenou, Antoine; MacKinnon, James G.; Nielsen, Morten Ørregaard (2017) : Validity of Wild Bootstrap Inference with Clustered Errors, Queen's Economics Department Working Paper, No. 1383, Queen's University, Department of Economics, Kingston (Ontario)

This Version is available at:
https://hdl.handle.net/10419/188895

Queen's Economics Department Working Paper No. 1383

Validity of Wild Bootstrap Inference with Clustered Errors

Antoine Djogbenou
Queen's University

James G. MacKinnon
Queen's University

Morten Ørregaard Nielsen
Queen's University

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

6-2017

Validity of Wild Bootstrap Inference with Clustered Errors*

Antoine A. Djogbenou
Queen's University
antoined@econ.queensu.ca

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

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

June 7, 2017

Abstract

We study asymptotic inference based on cluster-robust variance estimators for regression models with clustered errors, focusing on the wild cluster bootstrap and the ordinary wild bootstrap. We state conditions under which both asymptotic and bootstrap tests and confidence intervals will be asymptotically valid. These conditions put limits on the rates at which the cluster sizes can increase as the number of clusters tends to infinity. To include power in the analysis, we allow the data to be generated under sequences of local alternatives. Simulation experiments illustrate the theoretical results and show that all methods can work poorly in certain cases.

Keywords: Clustered data, cluster-robust variance estimator, CRVE, inference, wild bootstrap, wild cluster bootstrap.

JEL Codes: C15, C21, C23.

1 Introduction

Many applications of the linear regression model in economics and other fields involve error terms that appear to be correlated within clusters. More generally, the use of clustering in fields such as survey design goes back many decades. In such cases, it is very common to use a cluster-robust variance estimator (CRVE) to calculate asymptotic t-statistics and Wald statistics, because neglecting the cluster structure can lead to severely biased standard errors and large size distortions (Moulton, 1986). Although CRVE-based t-statistics work well in many cases, this approach can fail (sometimes disastrously) when the number of clusters is small, cluster sizes vary a lot, or the variable(s) of interest take non-zero values for only a few clusters; see Cameron and Miller (2015) for a recent survey.

*We are grateful to Russell Davidson, Silvia Gonçalves, Bruce Hansen, and seminar participants at NY Camp Econometrics XII, the 2017 CEA Annual Meeting, and U. C. San Diego for comments. Nielsen thanks the Canada Research Chairs program, the Social Sciences and Humanities Research Council of Canada (SSHRC), and the Center for Research in Econometric Analysis of Time Series (CREATES, funded by the Danish National Research Foundation, DNRF78) for financial support. MacKinnon thanks the SSHRC for financial support. Some of the computations were performed at the Centre for Advanced Computing at Queen's University.

†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.

The wild cluster bootstrap (WCB) was proposed in Cameron, Gelbach, and Miller (2008) as a way to obtain more accurate inferences in finite samples than using cluster-robust t-statistics. Although it typically does provide more accurate inferences, it too can fail (sometimes to an extreme degree) in certain cases; see MacKinnon and Webb (2017b). Interestingly, MacKinnon and Webb (2017a) provides simulation evidence which shows that the ordinary wild bootstrap (WB) seems to work better than the wild cluster bootstrap in some of those cases. A formal treatment of the conditions under which the WCB (and the WB in a cluster context), yields asymptotically valid inferences is clearly needed.

In this paper, we provide an asymptotic analysis of cluster-robust inference with particular emphasis on the WCB and the WB. In particular, we first establish the asymptotic distribution of the least squares estimator and associated cluster-robust t-statistic when the error terms have a cluster structure. We then establish the asymptotic validity of the WCB and the WB. All our results are given under simple primitive assumptions and rate conditions on the heterogeneity of cluster sizes. Our results allow for heteroskedasticity of unknown form and do not restrict dependence within clusters.

We are not aware of any previous work on the asymptotic validity of wild bootstrap methods for clustered errors. Conditions for asymptotic validity of CRVE-based inference are given by White (1984, Chapter 6), Liang and Zeger (1986), Hansen (2007), and Carter, Schnepel, and Steigerwald (2017), among others. All but the last of these assume that clusters are equal-sized. Carter et al. (2017) allows clusters of unequal sizes and studies the effects of heterogeneity across clusters.

An obvious alternative to the wild cluster bootstrap is the pairs cluster bootstrap, in which the bootstrap samples are constructed by resampling $(\boldsymbol{X}_g, \boldsymbol{y}_g)$ pairs. Several variants of this procedure were studied in Cameron, Gelbach, and Miller (2008) using simulation methods. In almost all cases, the pairs bootstrap produced less reliable inferences than the wild cluster bootstrap. This might have been expected, because the ordinary pairs bootstrap generally yields less reliable inferences in regression models with heteroskedastic errors than does the ordinary wild bootstrap; see, among others, MacKinnon (2002) and Davidson and Flachaire (2008).

Simulation evidence from previous studies is not the only reason for not studying the pairs cluster bootstrap here. The fundamental problem with the pairs cluster bootstrap is that, unlike the WB or the WCB, it does not condition on $\boldsymbol{X}$, which makes it unattractive for two reasons. First, when cluster sizes are not equal across clusters, the sample size will vary across the bootstrap samples. Second, when any of the regressors is a dummy variable that varies at the cluster level, the number of treated clusters and treated observations will vary across the bootstrap samples. Indeed, when there are few treated clusters in the actual sample, there may be none at all in some of the bootstrap samples, which would cause the $\boldsymbol{X}^\top\boldsymbol{X}$ matrix to be singular.

The remainder of the paper is organized as follows. In Section 2, we present the model that we study and the associated asymptotic theory. In Section 3, we demonstrate the asymptotic validity of both the wild cluster bootstrap and the ordinary wild bootstrap. In Section 4, we present results of some simulation studies. Section 5 concludes. The proofs are relegated to the appendices.

2 The Model and Asymptotic Theory

Consider a linear regression model with clustered errors written as

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_G \end{bmatrix} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_G \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \\ \vdots \\ \boldsymbol{u}_G \end{bmatrix}, \tag{1}$$

where each cluster, indexed by g, has N_g observations. The total number of observations in the entire sample is $N = \sum_{g=1}^{G} N_g$, and the $N \times k$ matrix of covariates $\boldsymbol{X}$ contains k linearly independent columns. The vector $\boldsymbol{\beta}$ is a $k \times 1$ vector of unknown parameters. The variance matrix $\boldsymbol{\Omega}$ of $\boldsymbol{u}$, conditional on $\boldsymbol{X}$, is block-diagonal with G diagonal $N_g \times N_g$ block variance matrices

$$\boldsymbol{\Omega}_g = \mathrm{E}(\boldsymbol{u}_g \boldsymbol{u}_g^\top | \boldsymbol{X}_g), \quad g = 1, \ldots, G. \tag{2}$$

When $N_g = 1$ for all g, the model (1) reduces to the well-known linear regression model with heteroskedasticity of unknown form. Hence, as a special case, our results cover that model as well.

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

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \tag{3}$$

Letting $\boldsymbol{Q}_N = N^{-1} \boldsymbol{X}^\top \boldsymbol{X}$ and $\boldsymbol{\Gamma}_N = N^{-2} \sum_{g=1}^{G} \boldsymbol{X}_g^\top \boldsymbol{\Omega}_g \boldsymbol{X}_g = N^{-2} \boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X}$, the variance matrix of $\hat{\boldsymbol{\beta}}$, conditional on $\boldsymbol{X}$, is given by

$$\boldsymbol{V}_N = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g^\top \boldsymbol{\Omega}_g \boldsymbol{X}_g \right) (\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \boldsymbol{Q}_N^{-1} \boldsymbol{\Gamma}_N \boldsymbol{Q}_N^{-1}. \tag{4}$$

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

$$\hat{\boldsymbol{V}} = \boldsymbol{Q}_N^{-1} \hat{\boldsymbol{\Gamma}} \boldsymbol{Q}_N^{-1}, \tag{5}$$

where $\hat{\boldsymbol{\Gamma}} = N^{-2} \sum_{g=1}^{G} \boldsymbol{X}_g^\top \hat{\boldsymbol{u}}_g \hat{\boldsymbol{u}}_g^\top \boldsymbol{X}_g$.

When $N_g = 1$ for all g, so that $G = N$, the estimator $\hat{\boldsymbol{V}}$ reduces to the familiar heteroskedasticity-consistent covariance matrix estimator (HCCME) of Eicker (1963) and White (1980); see also Arellano (1987). Several variations of the CRVE have been proposed to reduce its finite-sample bias, in the same way that variations of the HCCME (e.g., MacKinnon and White, 1985) can reduce its bias; see, among others, Kauermann and Carroll (2001), Bell and McCaffrey (2002), Imbens and Kolesár (2016), and Pustejovsky and Tipton (2017). However, since our focus is on bootstrap inference, we maintain the version of the CRVE given in (5), which is simple to compute and analyze.[1]

We let $\boldsymbol{\beta}_0$ denote the true value of $\boldsymbol{\beta}$ and restrict our attention to the cluster-robust t-statistic

$$t_a = \frac{\boldsymbol{a}^\top (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)}{\sqrt{\boldsymbol{a}^\top \hat{\boldsymbol{V}} \boldsymbol{a}}} \tag{6}$$

for testing the null hypothesis H_0: $\boldsymbol{a}^\top \boldsymbol{\beta} = \boldsymbol{a}^\top \boldsymbol{\beta}_0$ with $\boldsymbol{a}^\top \boldsymbol{a} = 1$ (a normalization that rules out degenerate cases but is much stronger than really needed) against either a one-sided or two-sided alternative hypothesis.

We next derive the asymptotic limit theory for t_a. In order to obtain those 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 $\{\boldsymbol{u}_g\}$ are independent across $g = 1, \ldots, G$ and satisfy $\mathrm{E}(\boldsymbol{u}_g | \boldsymbol{X}) = \boldsymbol{0}$ and $\mathrm{E}(\boldsymbol{u}_g \boldsymbol{u}_g^\top | \boldsymbol{X}) = \boldsymbol{\Omega}_g$ for all $g = 1, \ldots, G$, where $\boldsymbol{\Omega}_g$ is positive definite. In addition, for some $\lambda \geq 0$,

$$\sup_{1 \leq i \leq N_g, 1 \leq g \leq G} \mathrm{E}\left(|u_{ig}|^{4+\lambda} | \boldsymbol{X} \right) < \infty.$$

[1] It is easy to see that $\hat{\boldsymbol{V}}$ is singular if $k > G$, since the rank of $\hat{\boldsymbol{V}}$ cannot exceed G. This occurs, for example, whenever there are cluster fixed effects. In that case, the diagonal elements of $\hat{\boldsymbol{V}}$ that correspond to the fixed effects are 0. However, this does not prevent us from using (5) to make inferences about the remaining elements of $\boldsymbol{\beta}$.

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

$$\sup_{1\leq i\leq N_g,\, 1\leq g\leq G} \mathrm{E}\|\boldsymbol{X}_{ig}\|^{4+\lambda} < \infty,$$

where λ is the same as in Assumption 1.

Assumption 3. For λ defined in Assumption 1 and μ_N denoting the smallest eigenvalue of $\boldsymbol{\Gamma}_N$,

$$G \to \infty \text{ and } \mu_N^{-\frac{4+\lambda}{6+2\lambda}} \sup_{1\leq g\leq G} \frac{N_g}{N} \xrightarrow{P} 0.$$

Assumption 1 imposes the conditions that the error vectors $\boldsymbol{u}_g$ 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.

A substantial complication in the asymptotic theory for the model (1) is that the stochastic order of magnitude of $\hat{\boldsymbol{\beta}}$ in (3) depends in a complex way on the intra-cluster correlation structure, the regressors, the relative cluster sizes, and interactions among these. 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) $\boldsymbol{\Omega}_g$ is a dense matrix without restrictions and the regressors are correlated. In case (i) we easily find that, under Assumption 2,

$$\|\boldsymbol{V}_N\| = O_P(N^{-1}). \tag{7}$$

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 a general $\boldsymbol{\Omega}_g$ without restrictions, it holds that

$$\mathrm{E}(\boldsymbol{X}_g^\top \boldsymbol{\Omega}_g \boldsymbol{X}_g) = \mathrm{E}\Big(\sum_{i,j=1}^{N_g} \boldsymbol{X}_{ig}^\top \Omega_{g,ij} \boldsymbol{X}_{jg}\Big) = O(N_g^2), \tag{8}$$

where $\Omega_{g,ij}$ is the $(i,j)^{\text{th}}$ element of $\boldsymbol{\Omega}_g$, and $\boldsymbol{X}_{ig}$ is the i^{th} row of $\boldsymbol{X}_g$. It follows by (8) and Assumption 2 that

$$\|\boldsymbol{V}_N\| = O_P(1) O_P\Big(N^{-2}\sum_{g=1}^{G} N_g^2\Big) = O_P\Big(N^{-1} \sup_{1\leq g\leq G} N_g\Big). \tag{9}$$

Therefore, in case (ii), $\hat{\boldsymbol{\beta}}$ converges at rate $O_P(N^{-1/2} \sup_{1\leq g\leq G} N_g^{1/2})$. In general, it follows that, under Assumptions 1 and 2,

$$G \to \infty \text{ and } \sup_{1\leq g\leq G} \frac{N_g}{N} \to 0 \tag{10}$$

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

Assumption 3 first requires the number of clusters G to diverge, which obviously implies that the total number of observations $N = \sum_{g=1}^{G} N_g$ also diverges. The second condition of Assumption 3 restricts the extent of heterogeneity of cluster sizes N_g that is allowed. This restriction is related to the order of magnitude of the matrix $\boldsymbol{\Gamma}_N$, and specifically its smallest eigenvalue. The reason for this can be seen from the relation

$$\|\boldsymbol{\Gamma}_N^{-1}\|^2 = \sum_{j=1}^{k} \omega_j^2(\boldsymbol{\Gamma}_N^{-1}) = \sum_{j=1}^{k} \omega_j^{-2}(\boldsymbol{\Gamma}_N) \leq k\omega_1^{-2}(\boldsymbol{\Gamma}_N) = k\mu_N^{-2},$$

where $\omega_j(\boldsymbol{A})$ is the function that returns the j^{th} eigenvalue of $\boldsymbol{A}$, sorted in ascending order. Because $\boldsymbol{\Gamma}_N$ is positive definite, for N sufficiently large, all its eigenvalues are positive, so that $\|\boldsymbol{\Gamma}_N^{-1}\| \leq C\mu_N^{-1}$. In view of Assumption 2, it follows from equation (4) that

$$\|\boldsymbol{V}_N^{-1}\| = O_P(\mu_N^{-1}). \tag{11}$$

Thus μ_N^{-1} can be interpreted as the rate at which information accumulates.

More generally, the second condition of Assumption 3 ensures 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 depends on the size of the largest cluster, $\sup_{1\leq g\leq G} N_g$, and on the information contained within each cluster, given by μ_N.

To analyze the role of μ_N, we investigate the two extreme cases described above, namely, (i) $\boldsymbol{\Omega}$ is diagonal, and (ii) the $\boldsymbol{\Omega}_g$ are dense. In case (i), it straightforwardly holds that

$$\mu_N^{-1} = O_P(N), \tag{12}$$

and in case (ii) we find (by the same arguments as in the proof of Lemma A.2) that

$$\mu_N^{-1} = O_P\Big(N\Big(\sup_{1\leq g\leq G} N_g\Big)^{-1}\Big). \tag{13}$$

Clearly, (12) implies a stronger condition in Assumption 3 than (13). Specifically, in the latter case (ii) where the $\boldsymbol{\Omega}_g$ are dense, Assumption 3 is in fact implied by (10), which is very simple and very weak. Thus, when there is a high degree of intra-cluster correlation, so that the effective cluster size (as measured by the information contained in each cluster) is smaller than the actual cluster size (N_g), more heterogeneity in N_g is allowed by the second condition of Assumption 3.

Note that the exponent on μ_N in Assumption 3 is increasing in λ and satisfies $-\frac{4+\lambda}{6+2\lambda} \in [-\frac{2}{3}, -\frac{1}{2})$. Because $\mu_N \xrightarrow{P} 0$, the second condition in Assumption 3 is weaker when more moments are assumed to exist, i.e. when λ is higher, cf. Assumption 1. In any case, a sufficient condition for Assumption 3 to hold is

$$G \to \infty \text{ and } \mu_N^{-2/3} \sup_{1\leq g\leq G} \frac{N_g}{N} \xrightarrow{P} 0. \tag{14}$$

If we use the strongest rate for μ_N in (12) and (13), then we can find a sufficient condition for Assumption 3 which does not depend on μ_N:

$$G \to \infty \text{ and } \sup_{1\leq g\leq G} \frac{N_g}{N^{1/3}} \to 0. \tag{15}$$

The second condition of Assumption 3, or either of the sufficient conditions (14) and (15), allows a variety of types of cluster-size heterogeneity. For example, the N_g can be fixed constants as $G \to \infty$, or the N_g can diverge as in, e.g., $N_g = c_g G^{\alpha}$, where c_g and α are fixed constants. In the former case, with the N_g being fixed constants, which could be considered a prototypical case, and where also $\boldsymbol{\Omega}_g$ are non-zero in general, then $\hat{\boldsymbol{\beta}}$ is $O_P(G^{-1/2})$ for general $\boldsymbol{\Omega}_g$.

Because $\mu_N \xrightarrow{P} 0$, the second condition of Assumption 3 rules out the possibility that one cluster is proportional to the entire sample. However, it does allow one cluster, say $g = 1$, to be quite dominant, in the sense that $N_1 = N^{\alpha}$ satisfies the second condition of Assumption 3 for some $\alpha < 1$. Specifically, with allowance for any intra-cluster correlation structure, including independence, (15) shows that $\alpha < 1/3$ is allowed. However, in the case where the $\boldsymbol{\Omega}_g$ are dense, denoted case (ii) above, more heterogeneity of cluster sizes is allowed, and any $\alpha < 1$ satisfies (14). In this example, we note from (9) that the rate of convergence of $\hat{\boldsymbol{\beta}}$ can become very slow when α is close to one.

The possibility that the rate of convergence depends on a correlation structure is certainly not new. For example, Hansen (2007) showed that if both the time-series and cross-sectional dimensions in a panel setting diverge, then, in our notation, $\hat{\boldsymbol{\beta}}$ is either $\sqrt{N}$-convergent or $\sqrt{G}$-convergent depending on whether the degree of intra-cluster (time-series) correlation is strong or weak. Gonçalves (2011) extended Hansen (2007) to panels with both serial and cross-sectional dependence and found that the rate of convergence depended on a parameter, denoted ρ, that characterizes the degree of cross-sectional dependence.

Our first result in Theorem 1 below has several precursors in the literature, 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, which trivially satisfies our Assumption 3. In contrast, our primitive moment and rate conditions in Assumptions 1 and 3 allow heterogeneous clusters.

More recently, Carter, Schnepel, and Steigerwald (2017) obtains a result similar to Theorem 1 that allows clusters to be heterogeneous. That paper makes a primitive moment condition and some high-level assumptions to govern cluster-size heterogeneity, which are not trivially compared with our more primitive assumptions. In the moment condition in Assumption 1 of Carter et al. (2017) it is assumed, in addition to our Assumption 1, that $\boldsymbol{\Omega}_g^{-1/2}\boldsymbol{u}_g$ has the same moment structure as an independent sequence, up to the fourth order. This rules out any intra-cluster dependence in the third and fourth moments and allows a much sharper bound on $\mathrm{E}(\|\boldsymbol{X}_g^\top \boldsymbol{u}_g\|^4 \,|\, \boldsymbol{X})$, which becomes of order $O_P(N_g^2)$ instead of $O_P(N_g^4)$; c.f. our Lemma A.2. In contrast, we allow arbitrary dependence and correlation within each cluster in our Assumption 1, which is therefore substantially weaker than the corresponding moment condition in Assumption 1 of Carter et al. (2017).

Restrictions on cluster-size heterogeneity in Carter et al. (2017) are governed by their Assumptions 2(ii) and 2(iii), both of which are very high-level assumptions. While there is not much discussion of primitive sufficient conditions for their Assumption 2(iii), there is a rather detailed discussion of sufficient conditions for their Assumption 2(ii). Here, Carter et al. (2017) argue that (10) is a sufficient restriction on cluster size heterogeneity to satisfy their Assumption 2(ii). However, that argument relies on their statements that $\hat{\boldsymbol{\beta}} = O_P(N^{-1})$ and $\hat{\boldsymbol{\beta}}_g = O_P(N_g^{-1})$, where $\hat{\boldsymbol{\beta}}_g$ is the OLS estimator of $\boldsymbol{\beta}$ using only the observations in cluster g. Of course, both these statements would only be valid under quite restrictive further assumptions to limit the intra-cluster correlation; c.f. (8) and (9). Without restrictions on $\boldsymbol{\Omega}_g$, we find that their γ_g and $\bar{\gamma}$ satisfy $\gamma_g = O_P(N_g^2/N^2)$, $\bar{\gamma} = O_P(G^{-1}N^{-1}\sup_{1\leq g\leq G} N_g)$, and $\bar{\gamma}^{-1} = O_P(G\mu_N^{-1})$, respectively, such that their $G^{-1}\mathrm{E}(\Gamma) = O(GN^{-3}\mu_N^{-2}\sup_{1\leq g\leq G} N_g^3)$, and their Assumption 2(ii) requires this to tend to zero. In the case of dense $\boldsymbol{\Omega}_g$, this would require $GN^{-1}\sup_{1\leq g\leq G} N_g \to 0$, which is not possible because $G\sup_{1\leq g\leq G} N_g \geq N$. In contrast, our assumptions are primitive and straightforward to interpret.

Since we do not restrict the dependence within each cluster and wish to allow any structure for the intra-cluster variance matrices, $\boldsymbol{\Omega}_g$, we cannot normalize $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ in the usual way to obtain an asymptotic distribution. Instead, we 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. See, e.g., Hansen (2007, Theorem 2) or Carter, Schnepel, and Steigerwald (2017) for related arguments.

In order to analyze the asymptotic local power of asymptotic and bootstrap tests based on the cluster-robust t-statistic (6), we derive our results under the sequence of local alternatives

$$\boldsymbol{a}^\top(\boldsymbol{\beta}_G - \boldsymbol{\beta}_0) = (\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{1/2}\delta, \tag{16}$$

which is often referred as a "Pitman drift." Under (16), the DGP is characterized by a drifting sequence of true values of the parameter vector $\boldsymbol{\beta}$ indexed by G with drift parameter δ. When

$\delta = 0$, there is no drift, the null hypothesis H_0 is true, and the DGP is given by $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. In a more conventional setting, without clustering, the factor that multiplies δ would be $N^{-1/2}$.

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 (16). It then holds that*

$$\frac{\boldsymbol{a}^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_G)}{(\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{1/2}} \xrightarrow{d} \mathrm{N}(0,1), \tag{17}$$

$$\frac{\boldsymbol{a}^\top \hat{\boldsymbol{V}} \boldsymbol{a}}{\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a}} \xrightarrow{P} 1, \tag{18}$$

$$t_a \xrightarrow{d} \mathrm{N}(\delta, 1). \tag{19}$$

When the null hypothesis H_0 is true, the following corollary is an immediate consequence of Theorem 1.

Corollary 1. *Under the assumptions of Theorem 1 and H_0, it holds that $t_a \xrightarrow{d} \mathrm{N}(0,1)$.*

The result in Corollary 1 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; see also Cameron and Miller (2015) for a discussion of this issue.

An important consequence of the results in Theorem 1 and Corollary 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 (17), which is $(\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{1/2}$ instead of $N^{-1/2}$ or $G^{-1/2}$; see the discussion around (9).

The proof of Theorem 1 may be found in Appendix B. In this proof, we make use of the scalars $z_g = (\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-1/2} \boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} N^{-1} \boldsymbol{X}_g^\top \boldsymbol{u}_g$, which are indexed by cluster, and show that $\sum_{g=1}^G z_g$ converges in distribution. This makes it clear that, in an important sense, G rather than N is the relevant notion of sample size. Moreover, because we are summing over clusters, the clusters cannot be too heterogeneous. In particular, the information cannot be concentrated in one cluster (or a finite number of clusters), which is the reason why Assumption 3 imposes a restriction on $\sup N_g$.

Theorem 1, specifically the result (19), gives the asymptotic local power of the cluster-robust t-test as a function of δ. For example, for an α-level test against a left-sided alternative, the probability of rejecting the null hypothesis when the DGP is (16) is given by the asymptotic local power function

$$\Phi(z_\alpha - \delta), \tag{20}$$

where $\Phi(x)$ denotes the cumulative distribution function of the standard normal distribution, and z_α satisfies $\Phi(z_\alpha) = \alpha$. The asymptotic local power function (20) may seem to be too simple. However, the power of the t-test (or, equivalently, the asymptotic efficiency of the estimator) implicitly depends on G, the N_g, $\boldsymbol{X}$, and $\boldsymbol{\Omega}$ via the quantity $(\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{1/2}$ that appears in (16). The interpretation of δ implicitly changes whenever $(\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{1/2}$ changes.

Recalling the definition of $\boldsymbol{V}_N$ in (4), we see that individual cluster sizes, N_g, impact the power of the test in a way that depends heavily on the intra-cluster variance matrices, $\boldsymbol{\Omega}_g$, and is also confounded with the influence of the regressors $\boldsymbol{X}$. In general, the effects of the N_g, the $\boldsymbol{\Omega}_g$, and the regressors on the power of the t-test cannot be disentangled. They interact in a very complicated manner, so that the total number of observations cannot be relied upon as a notion of sample size. MacKinnon (2016) provides simulation evidence which illustrates this point.

3 Asymptotic Validity of the Wild (Cluster) Bootstrap

In this section, we consider the asymptotic validity of inference based on the wild cluster bootstrap (WCB) as an alternative to the asymptotic inference justified in Theorem 1. We consider two versions of the WCB. One of them (WCU) uses unrestricted estimates in the bootstrap data-generating process, and the other (WCR) uses estimates that satisfy the restriction H_0. The latter is the version proposed in Cameron, Gelbach, and Miller (2008), but it would be much easier to use the former to construct studentized bootstrap confidence intervals; see Davidson and MacKinnon (2004, Section 5.3). Cameron et al. (2008) provides no theoretical justification of the properties of the WCR bootstrap, nor any conditions under which it is valid or expected to work well.

The key feature of the wild cluster bootstrap DGP is the way in which the bootstrap error terms are generated. Let $v_1^*, v_2^*, \ldots, v_G^*$ denote IID realizations of an auxiliary random variable v^* with $\mathrm{E}^*(v^*) = 0$ and $\mathrm{Var}^*(v^*) = 1$. Here E^* and Var^* denote, respectively, the expectation and variance under the bootstrap measure P^*. The bootstrap error vectors $\boldsymbol{u}_g^*$, for $g = 1, \ldots, G$, are obtained by multiplying the residual vector $\hat{\boldsymbol{u}}_g$ (unrestricted) or $\tilde{\boldsymbol{u}}_g$ (restricted), for each cluster g, by the same draw v_g^* from the auxiliary distribution.

This may be contrasted with the ordinary wild bootstrap (WB) DGP, which was designed for regression models with independent, heteroskedastic errors but has recently been suggested in the context of cluster-robust inference by MacKinnon and Webb (2017a). For the WB, the bootstrap error vectors $\boldsymbol{u}_g^*$, for $g = 1, \ldots, G$, are obtained by multiplying each residual $\hat{u}_{ig}$ (unrestricted, WU) or $\tilde{u}_{ig}$ (restricted, WR), by a draw v_{ig}^* from the auxiliary distribution. We also analyze this bootstrap algorithm below.

We next describe the algorithm for the WCU and WCR bootstraps in some detail. We then prove the asymptotic validity of both versions. To describe the bootstrap algorithm and the properties of the bootstrap procedures, we introduce the notation $\ddot{\boldsymbol{u}}_g$ and $\ddot{\boldsymbol{\beta}}$, which will be taken to represent either restricted or unrestricted quantities, depending on which of WCR or WCU is being considered.

Wild Cluster Bootstrap Algorithm (WCU and WCR).

1. Estimate model (1) by OLS regression of $\boldsymbol{y}$ on $\boldsymbol{X}$ to obtain $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{V}}$ defined in (3) and (5), respectively. For 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, for H_0: $\boldsymbol{a}^\top \boldsymbol{\beta} = \boldsymbol{a}^\top \boldsymbol{\beta}_0$, given in (6).

3. For each of B bootstrap replications, indexed by b,

   (a) generate a new set of bootstrap errors given by $\boldsymbol{u}^{*b}$, where the subvector corresponding to cluster g is equal to $\boldsymbol{u}_g^{*b} = v_g^{*b} \ddot{\boldsymbol{u}}_g$, and v_g^{*b} denotes independent copies of the random variable v^* with $\mathrm{E}^*(v^*) = 0$ and $\mathrm{E}^*(v^{*2}) = 1$;

   (b) generate the bootstrap dependent variables according to $\boldsymbol{y}^{*b} = \boldsymbol{X}\ddot{\boldsymbol{\beta}} + \boldsymbol{u}^{*b}$;

   (c) obtain the bootstrap estimate $\hat{\boldsymbol{\beta}}^{*b} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}^{*b}$, the bootstrap residuals $\hat{\boldsymbol{u}}^{*b}$, and the bootstrap variance matrix estimate

$$\hat{\boldsymbol{V}}^{*b} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g^\top \hat{\boldsymbol{u}}_g^{*b} \hat{\boldsymbol{u}}_g^{*b\top} \boldsymbol{X}_g \right) (\boldsymbol{X}^\top \boldsymbol{X})^{-1};$$

(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 $\mathrm{H_L}$: $\boldsymbol{a}^\top\boldsymbol{\beta} < \boldsymbol{a}^\top\boldsymbol{\beta}_0$, $\mathrm{H_R}$: $\boldsymbol{a}^\top\boldsymbol{\beta} > \boldsymbol{a}^\top\boldsymbol{\beta}_0$, or $\mathrm{H_2}$: $\boldsymbol{a}^\top\boldsymbol{\beta} \neq \boldsymbol{a}^\top\boldsymbol{\beta}_0$, compute one of the following bootstrap P values:

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

where $\mathbb{I}(\cdot)$ denotes the indicator function. If the null hypothesis is $\mathrm{H_2}$, 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 WCU and WCR bootstraps for testing the hypothesis $\mathrm{H_0}$.[2] The following theorem is the bootstrap analogue of Theorem 1 and establishes the asymptotic normality of the WCB estimator and t-statistic.

Theorem 2. *Suppose Assumptions 1–3 are satisfied, that the true value of $\boldsymbol{\beta}$ is given by (16), and that $\mathrm{E}^*|v^*|^\eta < \infty$ for some $\eta > 2$. It then holds that*

$$\frac{\boldsymbol{a}^\top(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})}{(\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{1/2}} \xrightarrow{d^*} \mathrm{N}(0,1), \tag{21}$$

$$\frac{\boldsymbol{a}^\top \hat{\boldsymbol{V}}^* \boldsymbol{a}}{\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a}} \xrightarrow{P^*} 1, \tag{22}$$

$$t_a^* \xrightarrow{d^*} \mathrm{N}(0,1), \tag{23}$$

in probability.

The results in Theorem 2 are conditional on the original sample, and hence also conditional on t_a. This implies that the results (21)–(23) 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 2. *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 2 implies that the P values computed in step 4 of the WCU and WCR algorithms are asymptotically valid, as are studentized bootstrap confidence intervals. Intuitively, the WCB 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 $\mathrm{N}(0,1)$. Moreover, because the distribution (23) obtained for the bootstrap t-statistic, t_a^*, under the sequence

[2] With the WCU bootstrap, a slight modification of this algorithm can be used to construct studentized bootstrap confidence intervals 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).

of local alternatives (16) coincides with that of the original t-statistic, t_a, obtained under the null hypothesis H_0 in Corollary 1, Theorem 2 also implies that the WCB test has the same asymptotic local power function (20) as the asymptotic test based on t_a.

We next describe the algorithm for the ordinary (non-cluster) WU and WR bootstraps, and we then prove the asymptotic validity of both versions in the context of the cluster model (1).

Wild Bootstrap Algorithm (WU and WR).

All steps are identical to the corresponding steps in the WCU and WCR algorithms, except for step 3. (a), which is replaced by the following.

3. (a) generate a new set of bootstrap errors given by $\boldsymbol{u}^{*b}$, where $u_{ig}^{*b} = v_{ig}^{*b}\ddot{u}_{ig}$, and v_{ig}^{*b} denotes independent copies of the random variable v^* with $\mathrm{E}^*(v^*) = 0$ and $\mathrm{E}^*(v^{*2}) = 1$.

The following theorem is the wild bootstrap analogue of Theorem 2. It establishes the asymptotic normality of the WB estimator and t-statistic. To this end, let $\bar{\boldsymbol{\Omega}}$ denote the matrix obtained by setting the off-diagonal elements of $\boldsymbol{\Omega}$ to zero, $\bar{\boldsymbol{\Gamma}}_N = N^{-2}\boldsymbol{X}^\top\bar{\boldsymbol{\Omega}}\boldsymbol{X}$, and $\bar{\boldsymbol{V}}_N = \boldsymbol{Q}_N^{-1}\bar{\boldsymbol{\Gamma}}_N\boldsymbol{Q}_N^{-1}$; cf. (2), (4), and Assumption 2. Notice that, except in very special cases, $\bar{\boldsymbol{V}}_N \neq \boldsymbol{V}_N$.

Theorem 3. *Suppose that Assumptions 1 and 2 and condition (10) are satisfied, that the true value of $\boldsymbol{\beta}$ is given by (16), and that $\mathrm{E}^*|v^*|^\eta < \infty$ for some $\eta > 2$. It then holds that*

$$\frac{\boldsymbol{a}^\top(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})}{(\boldsymbol{a}^\top\bar{\boldsymbol{V}}_N\boldsymbol{a})^{1/2}} \xrightarrow{d^*} \mathrm{N}(0,1), \tag{24}$$

$$\frac{\boldsymbol{a}^\top\hat{\boldsymbol{V}}^*\boldsymbol{a}}{\boldsymbol{a}^\top\bar{\boldsymbol{V}}_N\boldsymbol{a}} \xrightarrow{P^*} 1, \tag{25}$$

$$t_a^* \xrightarrow{d^*} \mathrm{N}(0,1), \tag{26}$$

in probability.

The results in Theorem 3 differ in important ways from those in Theorem 2. From (24), we see that the WB is unable to replicate the intra-cluster correlation structure in $\boldsymbol{\Omega}_g$. This is expected, because the WB multiplies each residual by independent draws of the auxiliary random variable v^*, and hence the bootstrap DGP has independent (but possibly heteroskedastic) errors, even within clusters. In consequence, the wild bootstrap estimator $\boldsymbol{a}^\top\hat{\boldsymbol{\beta}}^*$ asymptotically has variance matrix $\boldsymbol{a}^\top\bar{\boldsymbol{V}}_N\boldsymbol{a}$, conditional on the original sample—see (24)—whereas both $\boldsymbol{a}^\top\hat{\boldsymbol{\beta}}$ and the corresponding WCB estimator asymptotically have variance matrix $\boldsymbol{a}^\top\boldsymbol{V}_N\boldsymbol{a}$ conditional on the original sample; see (17) and (21).

More importantly, however, the distribution of the WB t-statistic given in (26) replicates that of the original sample t-statistic under the null hypothesis as given in Corollary 1. The fact that $\hat{\boldsymbol{\beta}}^*$ has the wrong variance matrix has no effect on the asymptotic validity of the WB because t_a^* is based on an estimate of the correct variance matrix, $\boldsymbol{a}^\top\bar{\boldsymbol{V}}_N\boldsymbol{a}$,—see (25)—and thus has the correct asymptotic distribution. Moreover, as in Theorem 2, the results in Theorem 3 are conditional on the original sample. Hence Theorem 3 implies asymptotic validity of the WB. This is stated in the following corollary, which follows immediately from Theorems 1 and 3.

Corollary 3. *Under the conditions of Theorem 3, Assumption 3, and H_0,*

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

Like Corollary 2, this result implies that P values computed using the ordinary WB algorithms, WU and WR, as well as studentized bootstrap confidence intervals based on WU, are asymptotically valid. Moreover, since the result (26) is obtained under the sequence of local alternatives (16), it implies that the asymptotic local power functions of tests based on the WB coincide with those based on either the cluster-robust t-statistic (6) or the WCB. In other words, perhaps somewhat surprisingly, there is no loss of asymptotic efficiency or power from imposing independence within clusters in the bootstrap DGP.

However, because the normalization in (24) is in fact of order $N^{1/2}$ (see (B.11) in the proof of Theorem 3), it seems plausible that the distribution of t_a^* for the WB should approach the asymptotic $\mathrm{N}(0,1)$ distribution more rapidly than the distributions of either t_a^* for the WCB or t_a itself. This might well make it more difficult for the WB than for the WCB to mimic the distribution of t_a when μ_N^{-1} is small, e.g. when either G is small or the cluster sizes are heterogeneous and the $\boldsymbol{\Omega}_g$ are dense. We study this conjecture, and other aspects of the finite-sample performance of WB and WCB, in the next section.

4 Simulation experiments

In this section, we use Monte Carlo experiments to investigate the finite-sample performance of the procedures studied in Sections 2 and 3. Initially, we focus on cases in which cluster sizes vary, but not to an extreme extent. Later, we consider cases in which the rate condition given in Assumption 3 is either violated or close to being violated.

All of our experiments are based on the DGP

$$\boldsymbol{y}_g = \beta_1 + \beta_2 \boldsymbol{x}_g + \boldsymbol{u}_g, \quad \mathrm{E}(\boldsymbol{u}_g \boldsymbol{u}_g^\top) = \boldsymbol{\Omega}_g, \quad g = 1, \ldots, G, \tag{27}$$

where $\boldsymbol{\Omega}_g$ is an $N_g \times N_g$ matrix with every element on the principal diagonal equal to 1 and every off-diagonal element equal to ρ. Thus the error terms are equicorrelated with correlation coefficient ρ. In some of our simulations, the error terms are normally distributed. In others, they are generated by a normal mixture model with skewness of 1 and excess kurtosis of 3, in order to avoid the possibly excessive symmetry of normal errors. We obtained very similar results using both methods.[3] The null hypothesis is that $\beta_2 = 0$; this is equivalent to setting $\boldsymbol{a} = [0 \ 1]^\top$. Every experiment has 100,000 replications.

Since we have to impose conditions like Assumption 3 on the cluster sizes, we expect inference to be harder when cluster sizes are not all the same; for evidence on this point; see MacKinnon and Webb (2017b). In order to allow cluster sizes to vary systematically, we initially allocate N observations among G clusters using the equation

$$N_g = \left[\frac{N \exp(\gamma g / G)}{\sum_{j=1}^{G} \exp(\gamma j / G)} \right], \quad \text{for } g = 1, \ldots, G-1, \tag{28}$$

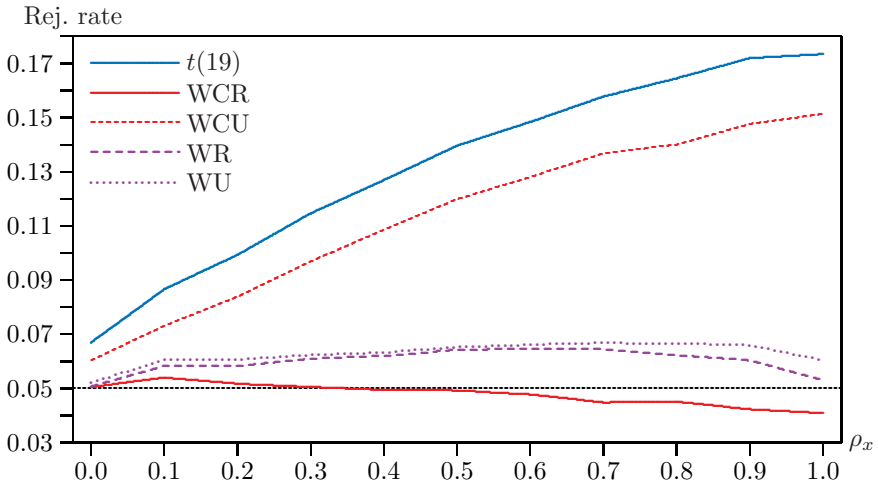
where $\gamma \geq 0$, $[\cdot]$ denotes the integer part of the argument, and $N_G = N - \sum_{g=1}^{G-1} N_g$. When $\gamma = 0$ and N/G is an integer, $N_g = N/G$ for all g. As γ increases, the cluster sizes become more and more unequal.

In the first set of experiments, the regressor is lognormally distributed and correlated within each cluster but uncorrelated across them, with correlation coefficient ρ_x, and the error terms are

[3] In a normal mixture model, the random variable u is equal to $u_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$ with probability p and $u_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$ with probability $1-p$. We used $p = 0.1967$, $\mu_1 = 0.7693$, $\mu_2 = -0.1884$, $\sigma_1 = 1.5734$, and $\sigma_2 = 0.6770$. To make the correlation within each cluster equal 0.1, the correlation between the components of the mixture was set to 0.2556.

generated by the normal mixture model described above.[4] Figure 1 shows rejection frequencies for five tests at the .05 level when $G = 20$, $N = 4000$, $\gamma = 3$, and $\rho = 0.1$. Because γ is quite large, cluster sizes vary from 33 to 598. For the t-test, we use critical values from the $t(G-1)$ distribution.[5] For the bootstrap tests, we report symmetric P values based on $B = 399$ bootstrap samples where the v^* come from the Rademacher distribution. Using alternative auxiliary distributions, such as the two-point Mammen distribution or the standard normal distribution, would inevitably change some of the results, in most cases not in a good way.

Figure 1: Rejection frequencies for continuous regressor, $G = 20$, $N = 4000$, $\gamma = 3$, $\rho = 0.10$



The horizontal axis shows ρ_x, which varies from 0.0 to 1.0 by increments of 0.1. We focus on ρ_x because past work, going back at least to Moulton (1986), has shown that the value of ρ_x is very important. When $\rho_x = 1$, the elements of $\boldsymbol{x}_g$ are constant within each cluster. It is evident that the cluster-robust t-test always overrejects, and it does so more and more severely as ρ_x increases. The WCR bootstrap does not work perfectly (except when $\rho_x = 0$), but it does perform quite well in all cases, tending to underreject for larger values of ρ_x. In contrast, the WCU bootstrap always overrejects, and for larger values of ρ_x it does so quite severely. The two ordinary wild bootstraps (WR and WU) perform almost perfectly when $\rho_x = 0$ and then deteriorate quite rapidly. However, they actually improve as ρ_x approaches 1.

We also performed a similar set of experiments with $\gamma = 0$, so that $N_g = 200$ for all G. The shapes of all the rejection frequency curves were essentially same as in Figure 1, but the size distortions were less than half as large. For example, the largest rejection frequency for the t-test was 0.0878 (size distortion 0.0378) instead of 0.1334 (size distortion 0.0834).

The results of the next set of experiments are shown in Figure 2. Here we fix ρ_x at 0.8 (which is roughly the worst value for the ordinary wild bootstrap tests) and vary G from 10 to 100 by 10 and then from 120 to 200 by 20. The value of γ is still 3, so cluster sizes change as G, and therefore N, increase. However, the way in which they vary is essentially the same as G increases. The largest

[4] We also ran some experiments in which the regressor was normally distributed. Most procedures worked a bit better, but the relations among them were largely unchanged.

[5] The CRVE used in these experiments is slightly more complicated than the one in equation (5). We multiply the latter by the factor $G(N-1)/\big((G-1)(N-k)\big)$, because that is what popular programs do. Without this factor, or if we had used the standard normal distribution instead of the $t(19)$ distribution, the overrejection would have been even more severe.

Figure 2: Rejection frequencies for continuous regressor, $\rho_x = 0.8$, $\gamma = 3$, $\rho = 0.10$

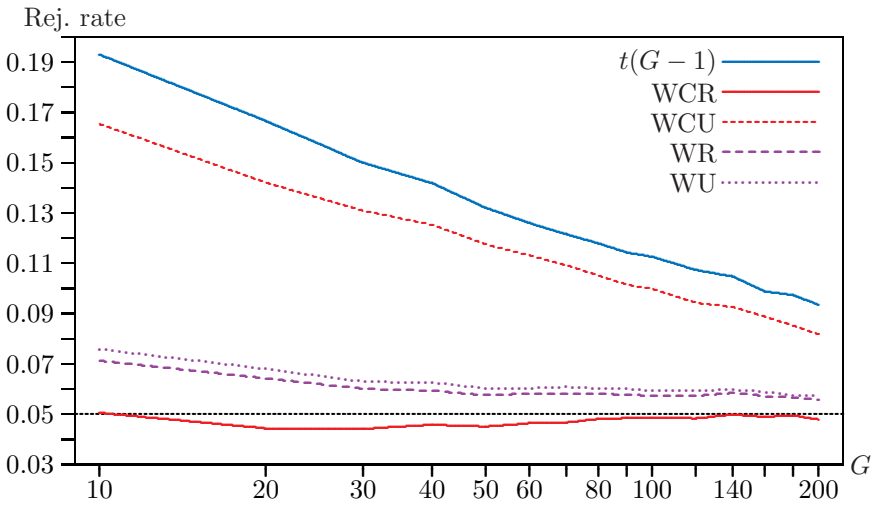
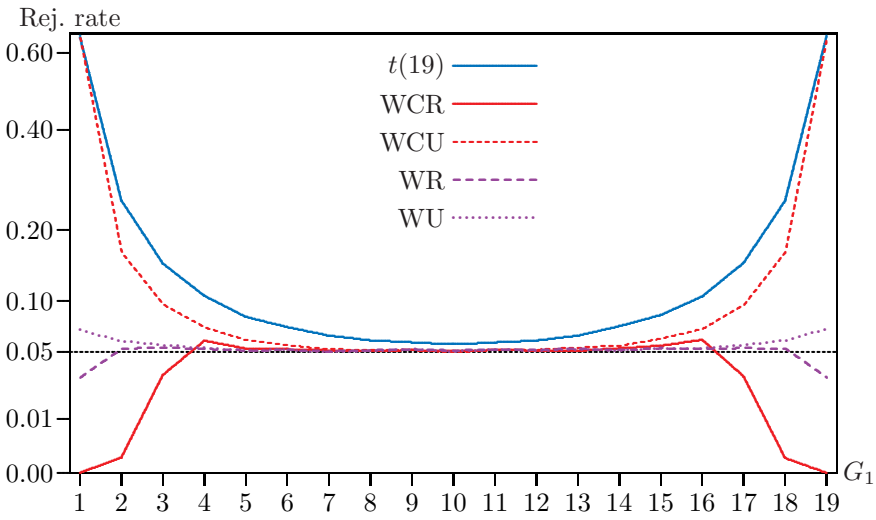


Figure 3: Rejection frequencies for treatment dummy, $G = 20$, $N = 4000$, $\gamma = 0$, $\rho = 0.10$

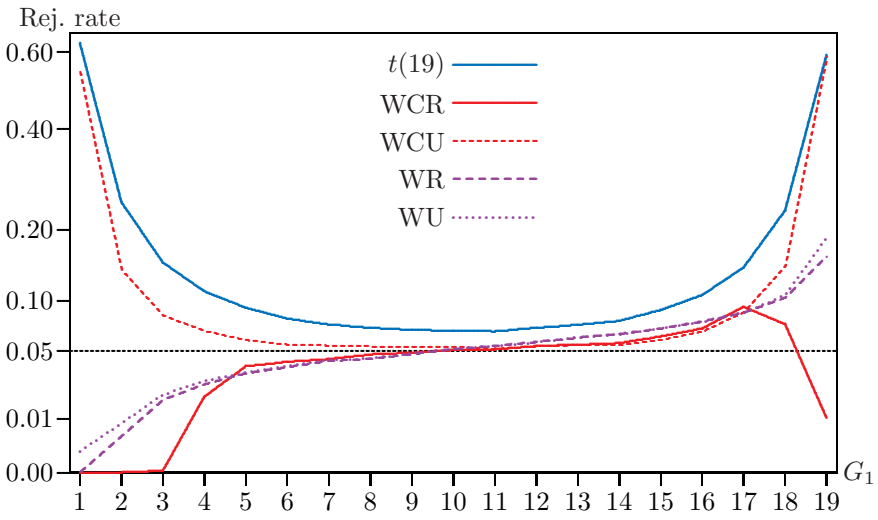


sample size is $N = 40,000$.

There are three striking results in Figure 2. The first is that all the bootstrap tests work better than the t-test. The second is that WCR performs very much better than WCU. It actually underrejects for most values of G.[6] This probably reflects the fact that the bootstrap DGP is estimated more efficiently when the model is estimated subject to restrictions; see Davidson and MacKinnon (1999). In particular, the unrestricted residuals be may worse estimators of the error terms than the restricted ones, especially for high-leverage observations where the regressor happens to be particularly large. The third result is that the two ordinary wild bootstrap tests perform very similarly, with WR always overrejecting a bit less than WU. It also looks as if WR and WU are improving less rapidly than WCU as G increases.

[6] This did not happen when the regressor was normally distributed.

Figure 4: Rejection frequencies for treatment dummy, $G = 20$, $N = 4000$, $\gamma = 3$, $\rho = 0.10$



In the next three experiments, a typical element of the test regressor in (27) is a dummy variable that equals 1 for some clusters and 0 for others; it can be thought of as a cluster-level treatment dummy. Many applications of cluster-robust inference involve this type of variable, and it is well-known that inference can be problematical when the number of treated, or untreated, clusters is small; see MacKinnon and Webb (2017b). We only study the pure treatment model here, but difference-in-differences (DiD) regressions are similar. In the DiD context, there are additional regressors, and the treatment variable is typically equal to 1 only for some observations within the treated clusters. When there are few treated clusters, exactly the same problems for inference arise.

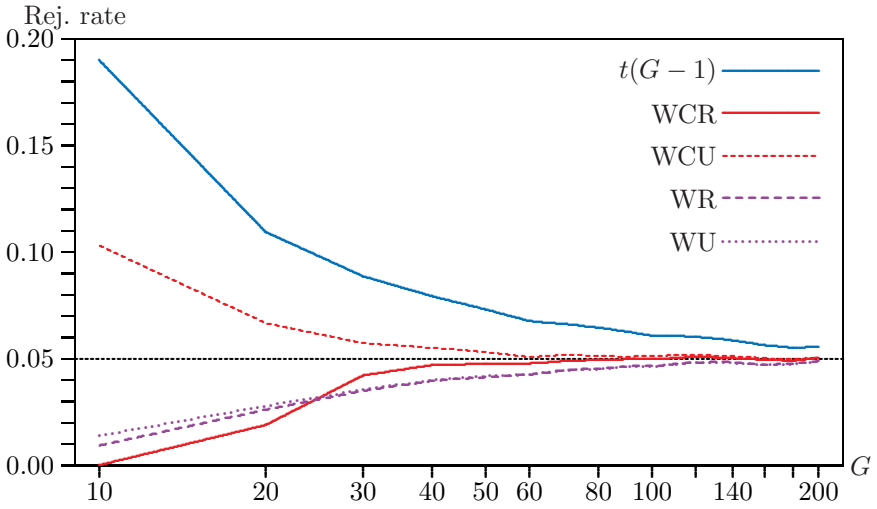
Figure 3 shows rejection frequencies for the same five tests when the regressor is a treatment dummy that equals 1 for G_1 out of $G = 20$ equal-sized clusters with $N = 4000$. Once again, the error terms are drawn from a normal mixture model. The vertical axis has been subjected to a square root transformation so that both very large and very small rejection frequencies can be shown on the same graph. This is essential, because the cluster-robust t-tests and the WCU bootstrap both reject more than 60% of the time when $G_1 = 1$ and $G_1 = 19$, and the WCR bootstrap never rejects in the same cases. These extreme overrejections and underrejections are precisely what the theory of MacKinnon and Webb (2017b) predicts for this model. However, all the bootstrap methods work very well for $6 \leq G_1 \leq 14$.

Perhaps surprisingly, the ordinary wild bootstrap works very much better than the wild cluster bootstrap for small and large values of G_1. This result is predicted in MacKinnon and Webb (2017a) for cases in which all clusters are the same size. Since all methods tend to work relatively well when clusters are the same size and G_1 is not too small, we need to investigate other cases.

Figure 4 shows rejection frequencies for a case in which $\gamma = 3$ and clusters are treated from smallest to largest.[7] Although there are a few exceptions for particular methods and particular values of G_1, all methods clearly work less well when $\gamma = 3$ than when $\gamma = 0$. The ordinary wild bootstrap works very much worse than before, underrejecting for small values of G_1 and overrejecting for large ones, as predicted by MacKinnon and Webb (2017a). WCU generally overrejects more

[7] If the error terms had been symmetric, treating the G_1 smallest clusters would have been equivalent to treating the $G_0 = G - G_1$ largest ones. Since the asymmetry here seems to have a very modest impact, it is safe to look at, say, the results for $G_1 = 18$ and use them to infer the results for treating the two largest clusters.

Figure 5: Rejection frequencies for treatment dummy, $N = 200G$, $G_1 = 0.2G$, $\gamma = 3$, $\rho = 0.10$



severely than before. WCR underrejects more severely for small values of G_1 and less severely for $G_1 = 19$, and it actually overrejects for $10 \leq G_1 \leq 18$.

The situation depicted in Figure 4 is rather extreme. In practice, it is unlikely that only the very smallest or very largest clusters would be treated. Thus, with highly variable cluster sizes and, say, just 3 or 4 treated clusters out of 20, we would expect all methods to perform better than they do in Figure 4 but not as well as they do in Figure 3.

Next, we turn our attention to what happens as G increases, with $\gamma = 3$ and the fraction of treated clusters held constant. Figure 5 shows rejection frequencies for various values of G that range from 10 to 200, with $G_1/G = 0.2$; as in Figure 2, the actual values are $10, 20, \ldots, 100, 120, \ldots, 200$. The rejection frequencies for $G = 20$ correspond to the ones for $G_1 = 4$ in Figure 4, although they differ slightly due to simulation randomness. As the results of Section 3 suggest, all methods improve steadily as G increases. However, the two wild cluster bootstrap methods evidently improve faster than the two ordinary wild bootstrap ones. For $G \geq 30$, the best methods are clearly WCR and WCU. These results are consistent with those in Figure 2, although WCR no longer seems to have a clear advantage over WCU.

In Figures 2 and 5, the largest cluster constitutes 27.5% of the sample for $G = 10$ but only 1.8% for $G = 200$. In the remaining experiments, we investigate cases where one large cluster dominates all the others, because this is a situation that is ruled out by the second condition of Assumption 3. The regressor is lognormally distributed and correlated within clusters with $\rho_x = 0.8$, and the error terms are normally distributed with $\rho = 0.1$. We set $N = 200(G-1)$ and $N_1 = 1000(N/2000)^\alpha$ for $\alpha \leq 1$ and then divide the remaining observations as evenly as possible among the remaining clusters. The values of G are $11, 21, \ldots, 101$ and $121, 141, \ldots, 201$. When $\alpha = 1$, exactly half the observations are always in the first cluster. When $\alpha < 1$, this is still true for $G = 11$, but the fraction of observations in the first cluster declines steadily as G increases. For example, when $\alpha = 0.9$, $N_1/N = 0.371$, and when $\alpha = 0.5$, $N_1/N = 0.112$.

Figure 6 shows rejection frequencies for CRVE t-tests for various values of α. Since our experimental design violates the rate condition given in Assumption 3 when $\alpha = 1$, it is not surprising that the rejection frequency increases steadily with G. This is also true when $\alpha = 0.95$. There appears to be no systematic change in rejection frequencies when $\alpha = 0.9$, but for smaller values

Figure 6: Rejection frequencies for CRVE t-tests, continuous regressor with one big cluster

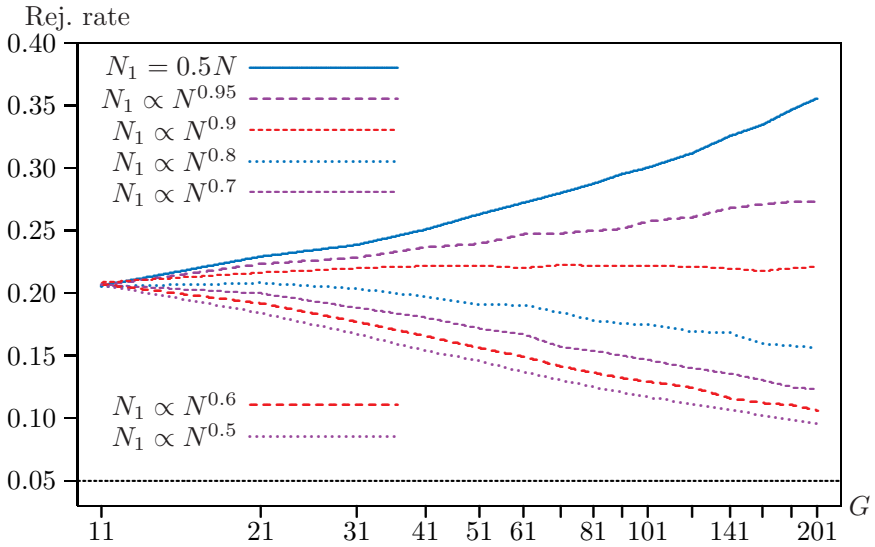
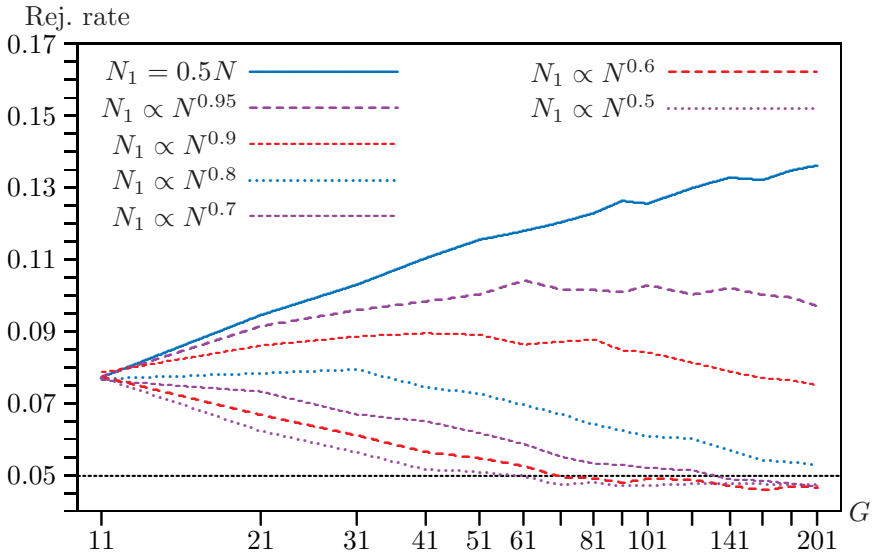


Figure 7: Rejection frequencies for WCR tests, continuous regressor with one big cluster



they clearly drop as G increases. However, even for the smallest values of α, G would evidently have to be very large for t-tests to yield reliable inferences.

Figure 7 shows rejection frequencies for the WCR bootstrap for the same set of experiments. These are much smaller than the ones in Figure 6. They still increase with G when $\alpha = 1$, but they eventually start to decrease for $\alpha = 0.95$ and $\alpha = 0.9$, and they decrease rapidly for smaller values of α. In quite a few cases, the procedure actually underrejects.

In contrast, we see from Figure 8 that rejection frequencies for the WCU bootstrap are quite high when $G = 11$ but decrease with G for all values of α except $\alpha = 1$, where they appear to increase very slowly. This procedure always works at least somewhat better than the CRVE t-test,

Figure 8: Rejection frequencies for WCU tests, continuous regressor with one big cluster

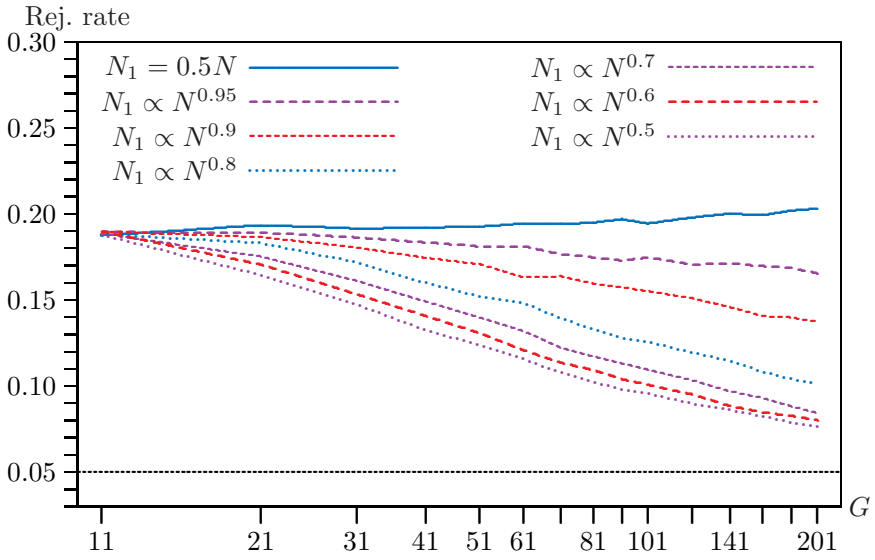
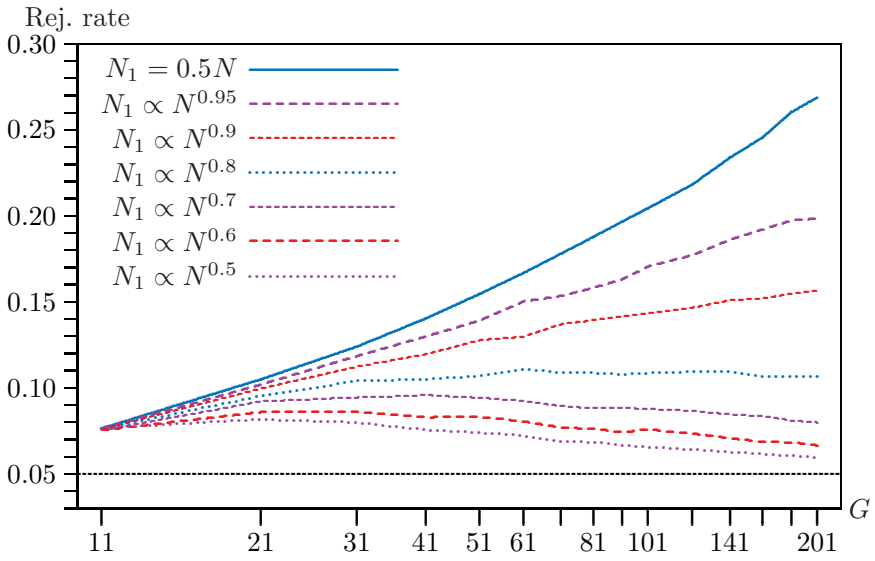


Figure 9: Rejection frequencies for WR tests, continuous regressor with one big cluster



especially for larger values of G.

Finally, we see from Figure 9 that the ordinary wild bootstrap (WR in this case, but WU is very similar) works quite well when G is small, but it then overrejects more severely for all values of α. It continues to overreject more and more severely even for large values of G when $\alpha \geq 0.9$. Only for the smallest values of α does WR clearly improve as G increases

It can be dangerous to draw firm conclusions from simulation experiments, and we are therefore reluctant to do so. Nevertheless, the results presented here, plus others that we do not present, strongly suggest that Theorems 2 and 3 are relevant for moderate numbers of clusters such as may often be encountered in practice, provided no single cluster is unduly large. They also suggest

that the ordinary wild bootstrap does not provide an asymptotic refinement (that is, a rate of improvement with G that is faster than the rate for the t-test), but that the wild cluster bootstrap may do so, at least in certain cases. Nevertheless, the former can sometimes perform better than the latter when G is small, even though it fails to mimic a crucial feature of the DGP.

5 Conclusion

In this paper, we have provided a formal analysis of the asymptotic properties of CRVE t-tests, the wild cluster bootstrap, and the ordinary wild bootstrap for linear regression models with clustered errors. The analysis makes quite weak assumptions about how the number of clusters and their sizes change as the sample size increases. This requires that, in the three key theorems of the paper, we use a self-normalizing rate of convergence that depends on the structure of the regressors and the variance matrix of the error terms. It would be impossible to obtain conventional rates of convergence for the least squares estimator $\hat{\boldsymbol{\beta}}$ without making much stronger assumptions.

The principal results of the paper are grouped into three theorems. Theorem 1 provides a theoretical foundation for asymptotic inference based on cluster-robust t-tests and cluster-robust confidence intervals. It differs from previous work in that it uses primitive assumptions which are straightforward to interpret. Theorem 2 provides a similar foundation for the wild cluster bootstrap (WCB), in both its restricted (WCR) and unrestricted (WCU) versions. Although the former generally performs better, the latter can perform well and is much easier to use for constructing confidence intervals. Finally, Theorem 3 shows that it is valid to base inferences on the ordinary wild bootstrap, combined with cluster-robust standard errors, even though the distribution of the bootstrap parameter estimates does not converge to the asymptotic distribution of the actual parameter estimates. Simulation evidence suggests that, in consequence, the WB tends to improve less rapidly than the WCB as the number of clusters becomes larger.

Appendix A: Preliminary lemmas

To prove our main results, we use the following preliminary lemmas. Throughout, C denotes a generic finite constant, which may take different values in different places.

Lemma A.1. *Let $w_g, g = 1, \ldots, G$, be independent random variables satisfying $\sum_{g=1}^{G} \mathrm{E}(w_g) = 0$ and $\sup_{1\leq g\leq G} \mathrm{E}|w_g|^\theta < \infty$ for some $\theta \geq 1$. Then $\sum_{g=1}^{G} w_g = O_P(G^{\max\{1/\theta, 1/2\}})$.*

Proof. Defining the de-meaned random variables $\tilde{w}_g = w_g - \mathrm{E}(w_g)$, it is evident that $\sum_{g=1}^{G} w_g - \sum_{g=1}^{G} \tilde{w}_g = 0$, and we can therefore assume $\mathrm{E}(w_g) = 0$ in the remainder of the proof. First suppose $1 \leq \theta \leq 2$. Let $\epsilon > 0$ be arbitrary, and choose K such that $K^\theta = \epsilon^{-1} \sup_g \mathrm{E}|w_g|^\theta$. By Markov's inequality and the von Bahr-Esseen inequality,

$$P\left(\sum_{g=1}^{G} w_g > KG^{1/\theta}\right) \leq \frac{\mathrm{E}\left|\sum_{g=1}^{G} w_g\right|^\theta}{K^\theta G} \leq \frac{\sum_{g=1}^{G} \mathrm{E}|w_g|^\theta}{K^\theta G} \leq \frac{\sup_g \mathrm{E}|w_g|^\theta}{K^\theta} = \epsilon.$$

If $\theta \geq 2$, then we apply the same proof setting $\theta = 2$. □

Lemma A.2. *Suppose Assumptions 1 and 2 are satisfied. Then*

$$\sup_{1\leq g\leq G} N_g^{-2-\xi} \mathrm{E}(\|\boldsymbol{X}_g^\top \boldsymbol{u}_g\|^{2+\xi} | \boldsymbol{X}) = O_P(1) \quad \text{for } 0 \leq \xi \leq 2 + \lambda,$$

$$\sup_{1\leq g\leq G} N_g^{-2-\xi} \|\boldsymbol{X}_g^\top \boldsymbol{X}_g\|^{2+\xi} = O_P(1) \quad \text{for } 0 \leq \xi \leq \lambda/2.$$

Proof. By the c_r inequality, the left-hand side of the first equation is bounded as

$$\sup_{1\leq g\leq G} N_g^{-2-\xi} N_g^{1+\xi} \sum_{i=1}^{N_g} \mathrm{E}(\|\boldsymbol{X}_{ig}^\top u_{ig}\|^{2+\xi}|\boldsymbol{X}) = \sup_{1\leq g\leq G} N_g^{-1} \sum_{i=1}^{N_g} \|\boldsymbol{X}_{ig}\|^{2+\xi} \mathrm{E}|u_{ig}|^{2+\xi}$$
$$\leq \sup_{1\leq i\leq N_g, 1\leq g\leq G} \|\boldsymbol{X}_{ig}\|^{2+\xi} \mathrm{E}|u_{ig}|^{2+\xi} = O_P(1)$$

by Assumptions 1 and 2 because $\xi < 2+\lambda$. Similarly, the left-hand side of the second equation is bounded by

$$\sup_{1\leq g\leq G} N_g^{-2-\xi} N_g^{1+\xi} \sum_{i=1}^{N_g} \|\boldsymbol{X}_{ig}^\top \boldsymbol{X}_{ig}\|^{2+\xi} = \sup_{1\leq g\leq G} N_g^{-1} \sum_{i=1}^{N_g} \|\boldsymbol{X}_{ig}\|^{4+2\xi} \leq \sup_{1\leq i\leq N_g, 1\leq g\leq G} \|\boldsymbol{X}_{ig}\|^{4+2\xi} = O_P(1)$$

by Assumption 2 because $\xi < \lambda/2$. □

Appendix B: Proofs of main results

B.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 (17). Because $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_G = \boldsymbol{Q}_N^{-1} N^{-1} \sum_{g=1}^G \boldsymbol{X}_g^\top \boldsymbol{u}_g$, we need to prove that

$$(\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-1/2} \boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} \frac{1}{N} \sum_{g=1}^G \boldsymbol{X}_g^\top \boldsymbol{u}_g \xrightarrow{d} \mathrm{N}(0,1). \tag{B.1}$$

We define $z_g = (\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-1/2} \boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} N^{-1} \boldsymbol{X}_g^\top \boldsymbol{u}_g$ with mean and variance given by $\mathrm{E}(z_g) = 0$ and $\mathrm{E}(z_g^2) = (\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-1} N^{-2} \boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} \boldsymbol{X}_g^\top \boldsymbol{\Omega}_g \boldsymbol{X}_g \boldsymbol{Q}_N^{-1} \boldsymbol{a}$, respectively, and note that $\sum_{g=1}^G \mathrm{E}(z_g^2) = 1$. Then (B.1) follows from the Lyapunov Central Limit Theorem for heterogeneous, independent random variables if, for some $\xi > 0$, $\sum_{g=1}^G \mathrm{E}|z_g|^{2+\xi} \to 0$ (Lyapunov's condition). To prove the latter, recall that μ_N denotes the smallest eigenvalue of $\boldsymbol{\Gamma}_N$, such that $\|\boldsymbol{\Gamma}_N^{-1}\| = O(\mu_N^{-1})$ and hence $\|\boldsymbol{V}_N^{-1}\| = O(\mu_N^{-1})$; see (11). Thus,

$$\sum_{g=1}^G \mathrm{E}|z_g|^{2+\xi} \leq (\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-1-\xi/2} \|\boldsymbol{Q}_N^{-1}\|^{2+\xi} N^{-2-\xi} \sum_{g=1}^G \mathrm{E}\|\boldsymbol{X}_g^\top \boldsymbol{u}_g\|^{2+\xi}$$
$$\leq C \mu_N^{-1-\xi/2} N^{-2-\xi} \sum_{g=1}^G N_g^{2+\xi} \leq C \mu_N^{-1-\xi/2} N^{-1-\xi} \sup_{1\leq g\leq G} N_g^{1+\xi} \to 0, \tag{B.2}$$

where the second inequality is due to Assumption 2 and Lemma A.2 with $\xi \leq 2+\lambda$, and the convergence is due to Assumption 3. This completes the proof of (17).

Proof of (18). 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

$$\boldsymbol{A}_{1N} = \frac{1}{N^2}\boldsymbol{Q}_N^{-1}\sum_{g=1}^{G}\boldsymbol{X}_g^\top \boldsymbol{u}_g\boldsymbol{u}_g^\top \boldsymbol{X}_g\boldsymbol{Q}_N^{-1} - \boldsymbol{V}_N,$$

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

$$\boldsymbol{A}_{3N} = \frac{1}{N^2}\boldsymbol{Q}_N^{-1}\sum_{g=1}^{G}\boldsymbol{X}_g^\top \boldsymbol{X}_g(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_G)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_G)^\top \boldsymbol{X}_g^\top \boldsymbol{X}_g\boldsymbol{Q}_N^{-1}.$$

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 $m = 1, 2, 3$. To prove the result for $m = 1$, let $w_g = z_g^2 - G^{-1}$ such that $\sum_{g=1}^{G} w_g = \sum_{g=1}^{G} z_g^2 - 1 = (\boldsymbol{a}^\top \boldsymbol{V}_N\boldsymbol{a})^{-1}\boldsymbol{a}^\top \boldsymbol{A}_{1N}\boldsymbol{a}$. We note that $\mathrm{E}(\sum_{g=1}^{G} w_g) = 0$ and prove convergence in mean-square. Thus, we analyze

$$\mathrm{E}\Big(\sum_{g=1}^{G} w_g\Big)^2 = \mathrm{E}\Big(\sum_{g,h=1}^{G} w_g w_h\Big) = \sum_{g=1}^{G}\mathrm{E}(w_g^2) + \sum_{g,h=1,g\neq h}^{G}\mathrm{E}(w_g w_h), \tag{B.3}$$

where the first term on the right-hand side is

$$\sum_{g=1}^{G}\mathrm{E}(w_g^2) = \sum_{g=1}^{G}\mathrm{E}(z_g^4) - 2G^{-1}\sum_{g=1}^{G}\mathrm{E}(z_g^2) + G^{-1} = \sum_{g=1}^{G}\mathrm{E}(z_g^4) - G^{-1},$$

since $\sum_{g=1}^{G}\mathrm{E}(z_g^2) = 1$. The Lyapunov condition (B.2) with $\xi = 2$ shows that $\sum_{g=1}^{G}\mathrm{E}(z_g^4) \to 0$, and hence $\sum_{g=1}^{G}\mathrm{E}(w_g^2) \to 0$. Because z_g^2 is an independent sequence (Assumption 1), the second term on the right-hand side of (B.3) is

$$\sum_{g,h=1,g\neq h}^{G}\mathrm{E}(w_g w_h) = \sum_{g,h=1,g\neq h}^{G}\mathrm{E}(w_g)\mathrm{E}(w_h) = \sum_{g,h=1}^{G}\mathrm{E}(w_g)\mathrm{E}(w_h) - \sum_{g=1}^{G}\big(\mathrm{E}(w_g)\big)^2 = -\sum_{g=1}^{G}\big(\mathrm{E}(w_g)\big)^2$$

because $\sum_{g=1}^{G}\mathrm{E}(w_g) = 0$. By Jensen's inequality, $\sum_{g=1}^{G}\big(\mathrm{E}(w_g)\big)^2 \leq \sum_{g=1}^{G}\mathrm{E}(w_g^2) \to 0$ using (B.2), which proves the result for $m = 1$.

Next, we analyze the case of $m = 2$, where, using the fact that $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_G)^\top \boldsymbol{X}_g^\top \boldsymbol{X}_g\boldsymbol{Q}_N^{-1}\boldsymbol{a}$ is a scalar, we find that

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

We first note that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_G\| = O_P(\|\boldsymbol{V}_N\|^{1/2}) = O_P(N^{-1/2}\sup_{1\leq g\leq G} N_g^{1/2})$; see (9). In addition, $(\boldsymbol{a}^\top \boldsymbol{V}_N\boldsymbol{a})^{-1} = O_P(\mu_N^{-1})$; see (11) and Assumption 2. Then, since

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

which is $O(N^{-3}\sup_{1\leq g\leq G} N_g^3)$ using Assumption 2, the Cauchy-Schwarz inequality, and Lemma A.2, we obtain that

$$(\boldsymbol{a}^\top \boldsymbol{V}_N\boldsymbol{a})^{-1}\boldsymbol{a}^\top \boldsymbol{A}_{2N}\boldsymbol{a} = O_P\left(\left(\mu_N^{-1/2}\frac{\sup_{1\leq g\leq G} N_g}{N}\right)^2\right) = o_P(1)$$

under Assumption 3; see also (14). Finally, the proof for $m = 3$ is nearly identical to that for $m = 2$, using the bound

$$\left\|(\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-1} \boldsymbol{a}^\top \boldsymbol{A}_{3N} \boldsymbol{a}\right\| \leq (\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-1} \frac{1}{N^2} \|\boldsymbol{Q}_N^{-1}\|^2 \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_G\|^2 \sum_{g=1}^{G} \|\boldsymbol{X}_g^\top \boldsymbol{X}_g\|^2$$
$$= O_P\left(\left(\mu_N^{-1/2} \frac{\sup_{1\leq g\leq G} N_g}{N}\right)^2\right) = o_P(1).$$

Proof of (19). We use (16) to decompose the t-statistic (6) as

$$t_a = \left(\frac{\boldsymbol{a}^\top \hat{\boldsymbol{V}} \boldsymbol{a}}{\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a}}\right)^{-1/2} \left((\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-1/2} \boldsymbol{a}^\top (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_G) + \delta\right),$$

and the result then follows directly from (17), (18), and Slutsky's Theorem.

B.2 Proof of Theorem 2

We give the proof conditional on $\boldsymbol{X}$, which is sufficient. It follows the same main outline as that of Theorem 1. Under the WCB probability measure, we let $\ddot{\boldsymbol{V}} = \boldsymbol{Q}_N^{-1} \ddot{\boldsymbol{\Gamma}} \boldsymbol{Q}_N^{-1}$ and $\ddot{\boldsymbol{\Gamma}} = N^{-2} \sum_{g=1}^{G} \boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g \ddot{\boldsymbol{u}}_g^\top \boldsymbol{X}_g$ denote the bootstrap true values. First we note that, by identical steps to those in the proof of Theorem 1, it holds more generally that, under (16),

$$\frac{\boldsymbol{a}^\top (\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_G)}{(\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{1/2}} = O_P(1) \quad \text{and} \quad \frac{\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a}}{\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a}} \xrightarrow{P} 1. \tag{B.4}$$

Proof of (21). Noting (B.4) and that $(\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-1/2} \boldsymbol{a}^\top (\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}}) = (\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-1/2} \boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} N^{-1} \boldsymbol{X}^\top \boldsymbol{u}^*$, we show that

$$(\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-1/2} \boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} N^{-1} \boldsymbol{X}^\top \boldsymbol{u}^* = (\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-1/2} \boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} N^{-1} \sum_{g=1}^{G} \boldsymbol{X}_g^\top \boldsymbol{u}_g^* \xrightarrow{d^*} \mathrm{N}(0,1). \tag{B.5}$$

To show (B.5), we follow the proof of (B.1). We define $z_g^* = (\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-1/2} \boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} N^{-1} \boldsymbol{X}_g^\top \boldsymbol{u}_g^*$ and apply the Lyapunov Central Limit Theorem to $\sum_{g=1}^{G} z_g^*$. Since $\mathrm{E}^*(z_g^*) = 0$ and $\sum_{g=1}^{G} \mathrm{E}^*(z_g^{*2}) = 1$ (because $\mathrm{E}^*(v_g^*) = 0$ and $\mathrm{E}^*(v_g^{*2}) = 1$ for all g), this requires verifying the Lyapunov condition for some $\xi > 0$; that is, we need to show that $\sum_{g=1}^{G} \mathrm{E}^*|z_g^*|^{2+\xi} \xrightarrow{P} 0$.

We first use $\ddot{\boldsymbol{u}}_g = \boldsymbol{u}_g - \boldsymbol{X}_g(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_G)$ and the c_r inequality to find that, for $0 \leq \xi \leq \lambda/2$,

$$\sup_{1\leq g\leq G} N_g^{-2-\xi} \|\boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g\|^{2+\xi} \leq 2^{1+\xi} \sup_{1\leq g\leq G} N_g^{-2-\xi} \|\boldsymbol{X}_g^\top \boldsymbol{u}_g\|^{2+\xi}$$
$$+ 2^{1+\xi} \|\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_G\|^{2+\xi} \sup_{1\leq g\leq G} N_g^{-2-\xi} \|\boldsymbol{X}_g^\top \boldsymbol{X}_g\|^{2+\xi} = O_P(1) \tag{B.6}$$

using Lemma A.2 and (B.4). It follows from (B.6) that

$$\sup_{1\leq g\leq G} N_g^{-2-\xi} \mathrm{E}^* \|\boldsymbol{X}_g^\top \boldsymbol{u}_g^*\|^{2+\xi} \leq \sup_{1\leq g\leq G} N_g^{-2-\xi} \|\boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g\|^{2+\xi} \mathrm{E}^* |v_g^*|^{2+\xi} = O_P(1) \tag{B.7}$$

for $0 \leq \xi \leq \min\{\eta - 2, \lambda/2\}$. We then find that

$$\sum_{g=1}^{G} \mathrm{E}^* |z_g^*|^{2+\xi} \leq \left(\frac{\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a}}{\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a}}\right)^{-1-\xi/2} (\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-1-\xi/2} \|\boldsymbol{Q}_N^{-1}\|^{2+\xi} N^{-2-\xi} \sum_{g=1}^{G} \mathrm{E}^* \|\boldsymbol{X}_g^\top \boldsymbol{u}_g^*\|^{2+\xi}$$
$$= O_P\left(\mu_N^{-1-\xi/2} N^{-2-\xi} \sum_{g=1}^{G} N_g^{2+\xi}\right) = O_P\left(\mu_N^{-1-\xi/2} N^{-1-\xi} N_g^{1+\xi}\right) \tag{B.8}$$

using (B.4), (B.7), and $(\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-1} = O_P(\mu_N^{-1})$. Thus, it holds that $\sum_{g=1}^G \mathrm{E}^*|z_g^*|^{2+\xi}$ is $o_P(1)$ under Assumption 3 as in (B.2).

Proof of (22). We note that $\boldsymbol{X}_g^\top \hat{\boldsymbol{u}}_g^* = \boldsymbol{X}_g^\top \boldsymbol{u}_g^* - \boldsymbol{X}_g^\top \boldsymbol{X}_g(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})$, which implies the decomposition

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

where

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

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

$$\boldsymbol{B}_{3N}^* = \boldsymbol{Q}_N^{-1} \frac{1}{N^2} \sum_{g=1}^G \boldsymbol{X}_g^\top \boldsymbol{X}_g (\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})^\top \boldsymbol{X}_g^\top \boldsymbol{X}_g \boldsymbol{Q}_N^{-1}.$$

Using this decomposition it is sufficient to prove that $(\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-1} \boldsymbol{a}^\top \boldsymbol{B}_{mN}^* \boldsymbol{a} = o_{P^*}(1)$, in probability, for $m = 1, 2, 3$. The proofs for each term roughly follow those for the corresponding term in the proof of (18). For $m = 1$, we let $w_g^* = z_g^{*2} - G^{-1}$ such that $(\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-1} \boldsymbol{a}^\top \boldsymbol{B}_{1N} \boldsymbol{a} = \sum_{g=1}^G w_g^* = \sum_{g=1}^G z_g^{*2} - 1$. We prove convergence in mean-square by analyzing

$$\mathrm{E}^* \Big(\sum_{g=1}^G w_g^* \Big)^2 = \mathrm{E}^* \Big(\sum_{g,h=1}^G w_g^* w_h^* \Big) = \sum_{g=1}^G \mathrm{E}^*(w_g^{*2}) + \sum_{g,h=1, g \neq h}^G \mathrm{E}^*(w_g^* w_h^*), \tag{B.9}$$

where the first term on the right-hand side is

$$\sum_{g=1}^G \mathrm{E}^*(w_g^{*2}) = \sum_{g=1}^G \mathrm{E}^*(z_g^{*4}) - 2G^{-1} \sum_{g=1}^G \mathrm{E}^*(z_g^{*2}) + G^{-1} = \sum_{g=1}^G \mathrm{E}^*(z_g^{*4}) - G^{-1} = o_P(1) \tag{B.10}$$

using $\sum_{g=1}^G \mathrm{E}^*(z_g^{*2}) = 1$ and the Lyapunov condition (B.8) with $\xi = 2$. Since z_g^{*2} is an independent sequence under the bootstrap measure, the second term on the right-hand side of (B.9) is

$$\sum_{g,h=1, g \neq h}^G \mathrm{E}^*(w_g^* w_h^*) = \sum_{g,h=1, g \neq h}^G \mathrm{E}^*(w_g^*) \mathrm{E}^*(w_h^*) = \sum_{g,h=1}^G \mathrm{E}^*(w_g^*) \mathrm{E}^*(w_h^*) - \sum_{g=1}^G (\mathrm{E}^*(w_g^*))^2 = -\sum_{g=1}^G (\mathrm{E}^*(w_g^*))^2$$

because $\mathrm{E}^*(v_g^{*2} - 1) = 0$. By Jensen's inequality and (B.10), $\sum_{g=1}^G (\mathrm{E}^*(w_g^*))^2 \leq \sum_{g=1}^G \mathrm{E}^*(w_g^{*2}) = o_P(1)$, which implies the result for $m = 1$. To prove the result for $m = 2$, we write

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

where $(\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-1} (\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a}) \xrightarrow{P} 1$ by (B.4), $(\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-1} = O_P(\mu_N^{-1})$ by (11) and Assumption 2, and $\|\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}}\| = O_{P^*}(\|\ddot{\boldsymbol{V}}\|^{1/2}) = O_{P^*}(N^{-1/2} \sup_{1 \leq g \leq G} N_g^{1/2})$, in probability, as in (9). We then apply the bound

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

which is $O_P(N^{-3}\sup_{1\leq g\leq G} N_g^3)$ using Assumption 2, the Cauchy-Schwarz inequality, Lemma A.2, and (B.7). We then obtain that

$$(\boldsymbol{a}^\top \ddot{\boldsymbol{V}}\boldsymbol{a})^{-1}\boldsymbol{a}^\top \boldsymbol{B}_{2N}^*\boldsymbol{a} = O_{P^*}\left(\left(\mu_N^{-1/2}\frac{\sup_{1\leq g\leq G} N_g}{N}\right)^2\right) = o_{P^*}(1),$$

in probability, under Assumption 3; see also (14). Finally, the proof for $m = 3$ is nearly identical to that for $m = 2$, using the bound

$$\|(\boldsymbol{a}^\top \ddot{\boldsymbol{V}}\boldsymbol{a})^{-1}\boldsymbol{a}^\top \boldsymbol{B}_{3N}^*\boldsymbol{a}\| \leq (\boldsymbol{a}^\top \ddot{\boldsymbol{V}}\boldsymbol{a})^{-1}\|\boldsymbol{Q}_N^{-1}\|^2\|\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}}\|^2 N^{-2}\sum_{g=1}^G \|\boldsymbol{X}_g^\top \boldsymbol{X}_g\|^2,$$

which is of order $O_{P^*}(\mu_N^{-1}N^{-2}\sup_{1\leq g\leq G} N_g^2) = o_{P^*}(1)$ in probability.

Proof of (23). Follows immediately by (21), (22), and Slutsky's Theorem.

B.3 Proof of Theorem 3

To prove (24), (25), and (26), we follow the same main steps as in the proof of Theorem 2, but now let $\ddot{\boldsymbol{V}} = \boldsymbol{Q}_N^{-1}\ddot{\boldsymbol{\Gamma}}\boldsymbol{Q}_N^{-1}$ and $\ddot{\boldsymbol{\Gamma}} = N^{-2}\sum_{g=1}^G\sum_{i=1}^{N_g} \boldsymbol{X}_{ig}^\top \ddot{u}_{ig}^2 \boldsymbol{X}_{ig}$ denote the bootstrap true values under the WB probability measure. First, we note that

$$(\boldsymbol{a}^\top \bar{\boldsymbol{V}}_N\boldsymbol{a})^{-1} = O_P(N) \tag{B.11}$$

in view of (11) and (12).

Proof of (24). We now have $(\boldsymbol{a}^\top \bar{\boldsymbol{V}}_N\boldsymbol{a})^{-1/2}\boldsymbol{a}^\top(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}}) = (\boldsymbol{a}^\top \bar{\boldsymbol{V}}_N\boldsymbol{a})^{-1/2}\boldsymbol{a}^\top\boldsymbol{Q}_N^{-1}N^{-1}\boldsymbol{X}^\top\boldsymbol{u}^*$. Under the WB probability measure, u_{ig}^* is heteroskedastic, but independent across both i and g. We let $z_{ig}^* = (\boldsymbol{a}^\top \bar{\boldsymbol{V}}_N\boldsymbol{a})^{-1/2}\boldsymbol{a}^\top\boldsymbol{Q}_N^{-1}N^{-1}\boldsymbol{X}_{ig}^\top u_{ig}^*$, where $\mathrm{E}^*(z_{ig}^*) = 0$ and $\sum_{g=1}^G\sum_{i=1}^{N_g}\mathrm{E}^*(z_{ig}^{*2}) = 1$. The desired result now follows by application of the Lyapunov Central Limit Theorem to $\sum_{g=1}^G\sum_{i=1}^{N_g} z_{ig}^*$, which requires verifying the Lyapunov condition that, for some $\xi > 0$, $\sum_{g=1}^G\sum_{i=1}^{N_g}\mathrm{E}^*|z_{ig}^*|^{2+\xi} \xrightarrow{P} 0$. Noting (B.11) and $\mathrm{E}^*\|\boldsymbol{X}_{ig}^\top u_{ig}^*\|^{2+\xi} = O_P(1)$—see also (B.6) and (B.7)—we find that

$$\sum_{g=1}^G\sum_{i=1}^{N_g}\mathrm{E}^*|z_{ig}^*|^{2+\xi} \leq (\boldsymbol{a}^\top \bar{\boldsymbol{V}}_N\boldsymbol{a})^{-1-\xi/2}\|\boldsymbol{Q}_N^{-1}\|^{2+\xi}N^{-2-\xi}\sum_{g=1}^G\sum_{i=1}^{N_g}\mathrm{E}^*\|\boldsymbol{X}_{ig}^\top u_{ig}^*\|^{2+\xi} = O_P(N^{-\xi/2})$$

for $\xi \leq \min\{\lambda/2, \eta - 2\}$, which proves (24).

Proof of (25). To prove (22), we show that

$$\frac{\boldsymbol{a}^\top \ddot{\boldsymbol{V}}\boldsymbol{a}}{\boldsymbol{a}^\top \bar{\boldsymbol{V}}_N\boldsymbol{a}} \xrightarrow{P} 1 \quad \text{and} \quad \frac{\boldsymbol{a}^\top \hat{\boldsymbol{V}}^*\boldsymbol{a}}{\boldsymbol{a}^\top \ddot{\boldsymbol{V}}\boldsymbol{a}} \xrightarrow{P^*} 1, \tag{B.12}$$

in probability. For the first statement in (B.12) we use the decomposition

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

where

$$\boldsymbol{C}_{1N} = \boldsymbol{Q}_N^{-1}\frac{1}{N^2}\sum_{g=1}^{G}\sum_{i=1}^{N_g}\boldsymbol{X}_{ig}^\top\left(u_{ig}^2 - \mathrm{E}(u_{ig}^2|\boldsymbol{X})\right)\boldsymbol{X}_{ig}\boldsymbol{Q}_N^{-1},$$

$$\boldsymbol{C}_{2N} = \boldsymbol{Q}_N^{-1}\frac{1}{N^2}\sum_{g=1}^{G}\sum_{i=1}^{N_g}\boldsymbol{X}_{ig}^\top\boldsymbol{u}_{ig}(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_G)^\top\boldsymbol{X}_{ig}^\top\boldsymbol{X}_{ig}\boldsymbol{Q}_N^{-1}, \text{ and}$$

$$\boldsymbol{C}_{3N} = \boldsymbol{Q}_N^{-1}\frac{1}{N^2}\sum_{g=1}^{G}\sum_{i=1}^{N_g}\boldsymbol{X}_{ig}^\top\boldsymbol{X}_{ig}(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_G)(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_G)^\top\boldsymbol{X}_{ig}^\top\boldsymbol{X}_{ig}\boldsymbol{Q}_N^{-1},$$

and show that $(\boldsymbol{a}^\top\bar{\boldsymbol{V}}_N\boldsymbol{a})^{-1}\boldsymbol{a}^\top\boldsymbol{C}_{mN}\boldsymbol{a} \xrightarrow{P} 0$ for $m = 1,\ldots,3$. Equivalently, since $(\boldsymbol{a}^\top\bar{\boldsymbol{V}}_N\boldsymbol{a})^{-1} = O_P(N)$, we show that $N\boldsymbol{a}^\top\boldsymbol{C}_{mN}\boldsymbol{a} \xrightarrow{P} 0$ for $m = 1,\ldots,3$. To prove the result for $m = 1$, for any conforming vector, $\boldsymbol{b}$, let $w_{ig} = \boldsymbol{b}^\top\boldsymbol{X}_{ig}^\top(u_{ig}^2 - \mathrm{E}(u_{ig}^2|\boldsymbol{X}))\boldsymbol{X}_{ig}\boldsymbol{b}$, which is independent across g conditional on $\boldsymbol{X}$. By the law of iterated expectations,

$$\mathrm{E}\left(\left(\sum_{g=1}^{G}\sum_{i=1}^{N_g}w_{ig}\right)^2\right) = \sum_{g=1}^{G}\mathrm{E}\left(\left(\sum_{i=1}^{N_g}w_{ig}\right)^2\right) \leq \sum_{g=1}^{G}N_g\sum_{i=1}^{N_g}\mathrm{E}(w_{ig}^2) \leq CN\sup_{1\leq g\leq G}N_g,$$

using also the c_r inequality and Assumptions 1 and 2. It follows by Assumption 2 and (10) that $N\boldsymbol{a}^\top\boldsymbol{C}_{1N}\boldsymbol{a} = O_P(N^{-1/2}\sup_{1\leq g\leq G}N_g^{1/2} = o_P(1)$. For $m = 2$, we apply the bound

$$\|N\boldsymbol{a}^\top\boldsymbol{C}_{2N}\boldsymbol{a}\| \leq N\|\boldsymbol{Q}_N^{-1}\|^2\|\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_G\|\frac{1}{N^2}\sum_{g=1}^{G}\sum_{i=1}^{N_g}\|\boldsymbol{X}_{ig}\|^3\|u_{ig}\| = O_P(N^{-1/2}\sup_{1\leq g\leq G}N_g) = o_P(1),$$

using $\|\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_G\| = O_P(N^{-1/2}\sup_{1\leq g\leq G}N_g^{1/2})$, $\boldsymbol{Q}_N^{-1} = O_P(1)$, (10), and Assumptions 1 and 2. Finally, we turn to $m = 3$, where, by an identical argument, we obtain

$$\|N\boldsymbol{a}^\top\boldsymbol{C}_{3N}\boldsymbol{a}\| \leq N\|\boldsymbol{Q}_N^{-1}\|^2\|\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_G\|^2\frac{1}{N^2}\sum_{g=1}^{G}\sum_{i=1}^{N_g}\|\boldsymbol{X}_{ig}\|^4 = O_P(N^{-1}\sup_{1\leq g\leq G}N_g) = o_P(1).$$

To prove the second statement in (B.12), we apply the decomposition

$$\boldsymbol{a}^\top(\hat{\boldsymbol{V}}^* - \ddot{\boldsymbol{V}})\boldsymbol{a} = \boldsymbol{a}^\top\left(\boldsymbol{D}_{1N}^* + \boldsymbol{D}_{2N}^* - \boldsymbol{D}_{3N}^* - \boldsymbol{D}_{3N}^{*\top} + \boldsymbol{D}_{4N}^*\right)\boldsymbol{a},$$

where

$$\boldsymbol{D}_{1N}^* = \boldsymbol{Q}_N^{-1}\frac{1}{N^2}\sum_{g=1}^{G}\sum_{i=1}^{N_g}\boldsymbol{X}_{ig}^\top\ddot{u}_{ig}^2\boldsymbol{X}_{ig}(v_{ig}^{*2} - 1)\boldsymbol{Q}_N^{-1},$$

$$\boldsymbol{D}_{2N}^* = \boldsymbol{Q}_N^{-1}\frac{1}{N^2}\sum_{g=1}^{G}\sum_{i\neq j=1}^{N_g}\boldsymbol{X}_{ig}^\top\ddot{u}_{ig}\ddot{u}_{jg}\boldsymbol{X}_{jg}v_{ig}^*v_{jg}^*\boldsymbol{Q}_N^{-1},$$

$$\boldsymbol{D}_{3N}^* = \boldsymbol{Q}_N^{-1}\frac{1}{N^2}\sum_{g=1}^{G}\boldsymbol{X}_g^\top\boldsymbol{u}_g^*(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})^\top\boldsymbol{X}_g^\top\boldsymbol{X}_g\boldsymbol{Q}_N^{-1}, \text{ and}$$

$$\boldsymbol{D}_{4N}^* = \boldsymbol{Q}_N^{-1}\frac{1}{N^2}\sum_{g=1}^{G}\boldsymbol{X}_g^\top\boldsymbol{X}_g(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})^\top\boldsymbol{X}_g^\top\boldsymbol{X}_g\boldsymbol{Q}_N^{-1}.$$

In view of (B.11) and the first part of (B.12), it suffices to prove that $(\boldsymbol{a}^\top \bar{\boldsymbol{V}}_N \boldsymbol{a})^{-1} \boldsymbol{a}^\top \boldsymbol{D}^*_{mN} \boldsymbol{a} \xrightarrow{P^*} 0$, in probability, for $m = 1, \ldots, 4$, or equivalently that $N \boldsymbol{a}^\top \boldsymbol{D}^*_{mN} \boldsymbol{a} \xrightarrow{P^*} 0$, in probability, also for $m = 1, \ldots, 4$.

To prove the result for $m = 1$, we let $w^*_{ig} = z^{*2}_{ig} - 1/N$, which is independent across i and g (under the WB probability measure) and satisfies $\sum_{g=1}^G \sum_{i=1}^{N_g} w^*_{ig} = (\boldsymbol{a}^\top \bar{\boldsymbol{V}}_N \boldsymbol{a})^{-1} \boldsymbol{a}^\top \boldsymbol{D}^*_{1N} \boldsymbol{a}$. Moreover,

$$\mathrm{E}^*\Big(\sum_{g=1}^G \sum_{i=1}^{N_g} w^*_{ig} \Big)^2 = \mathrm{E}^*\Big(\sum_{g,h=1}^G \sum_{i,j=1}^{N_g} w^*_{ig} w^*_{jh} \Big) = \sum_{g=1}^G \sum_{i=1}^{N_g} \mathrm{E}^*(w^{*2}_{ig}) + \sum_{\substack{g,h=1 \\ (i,g)\neq(j,h)}}^G \sum_{i,j=1}^{N_g} \mathrm{E}^*(w^*_{ig} w^*_{jh}), \tag{B.13}$$

where, using $\sum_{g=1}^G \sum_{i=1}^{N_g} \mathrm{E}^*(z^{*2}_{ig}) = 1$,

$$\sum_{g=1}^G \sum_{i=1}^{N_g} \mathrm{E}^*(w^{*2}_{ig}) = \sum_{g=1}^G \sum_{i=1}^{N_g} \mathrm{E}^*(z^{*4}_{ig}) - 1/N = o_P(1) \tag{B.14}$$

by an application of the Lyapunov condition for $\xi = 2$. Thus, we only need to verify that the second term on the right-hand side of (B.13) is $o_P(1)$. Using $\mathrm{E}^*(v^{*2}_{ig} - 1) = 0$ and the fact that $z^{*2}_{ig} - 1/N$ is an independent sequence, under the WB probability measure, we have that

$$\sum_{\substack{g,h=1 \\ (i,g)\neq(j,h)}}^G \sum_{i,j=1}^{N_g} \mathrm{E}^*(w^*_{ig} w^*_{jh}) = \sum_{g,h=1}^G \sum_{i,j=1}^{N_g} \mathrm{E}^*(w^*_{ig}) \mathrm{E}^*(w^*_{jh}) - \sum_{g=1}^G \sum_{i=1}^{N_g} \big(\mathrm{E}^*(w^*_{ig})\big)^2 = -\sum_{g=1}^G \sum_{i=1}^{N_g} \big(\mathrm{E}^*(w^*_{ig})\big)^2 = o_P(1)$$

because $\sum_{g=1}^G \sum_{i=1}^{N_g} \big(\mathrm{E}^*(w^*_{ig})\big)^2 \leq \sum_{g=1}^G \sum_{i=1}^{N_g} \mathrm{E}^*(w^{*2}_{ig}) = o_P(1)$ by Jensen's inequality and (B.14).

For $m = 2$, we note that $\mathrm{E}^*(v^*_{ig} v^*_{jg}) = 0$ for $i \neq j$, so that $(\boldsymbol{a}^\top \bar{\boldsymbol{V}}_N \boldsymbol{a})^{-1} \boldsymbol{a}^\top \boldsymbol{D}^*_{2N} \boldsymbol{a} = \sum_{g=1}^G \sum_{i\neq j=1}^{N_g} z^*_{ig} z^*_{jg}$ is a sum of zero mean random variables (conditional on the original sample). Hence,

$$\begin{aligned} \mathrm{E}^*\Big(\big((\boldsymbol{a}^\top \bar{\boldsymbol{V}}_N \boldsymbol{a})^{-1} \boldsymbol{a}^\top \boldsymbol{D}^*_{2N} \boldsymbol{a} \big)^2 \Big) &= \sum_{g=1}^G \sum_{i\neq j=1}^{N_g} \mathrm{E}^*(z^{*2}_{ig}) \mathrm{E}^*(z^{*2}_{jg}) \\ &\leq (\boldsymbol{a}^\top \bar{\boldsymbol{V}}_N \boldsymbol{a})^{-2} \|\boldsymbol{Q}_N^{-1}\|^4 N^{-4} \sum_{g=1}^G \sum_{i\neq j=1}^{N_g} \mathrm{E}^* \|\boldsymbol{X}_{ig}^\top u^*_{ig}\|^2 \mathrm{E}^* \|\boldsymbol{X}_{jg}^\top u^*_{jg}\|^2, \end{aligned}$$

which is $O_P(N^2 N^{-3} \sup_{1\leq g\leq G} N_g) = O_P(N^{-1} \sup_{1\leq g\leq G} N_g) = o_P(1)$ by (B.11) and $\mathrm{E}^* \|\boldsymbol{X}_{ig}^\top u^*_{ig}\|^2 = O_P(1)$; see also (B.6) and (B.7).

For $m = 3$, we first observe that

$$(\boldsymbol{a}^\top \bar{\boldsymbol{V}}_N \boldsymbol{a})^{-1} \boldsymbol{a}^\top \boldsymbol{D}^*_{3N} \boldsymbol{a} = (\boldsymbol{a}^\top \bar{\boldsymbol{V}}_N \boldsymbol{a})^{-1} (\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})^\top \frac{1}{N^2} \sum_{g=1}^G \boldsymbol{X}_g^\top \boldsymbol{X}_g \boldsymbol{Q}_N^{-1} \boldsymbol{a} \boldsymbol{a}^\top \boldsymbol{X}_g^\top \boldsymbol{u}^*_g, \tag{B.15}$$

where

$$\mathrm{E}^* \Big\| \sum_{g=1}^G \boldsymbol{X}_g^\top \boldsymbol{X}_g \boldsymbol{Q}_N^{-1} \boldsymbol{a} \boldsymbol{a}^\top \boldsymbol{X}_g^\top \boldsymbol{u}^*_g \Big\|^2 \leq \|\boldsymbol{Q}_N^{-1}\|^2 \sum_{g=1}^G \|\boldsymbol{X}_g^\top \boldsymbol{X}_g\|^2 \mathrm{E}^* \|\boldsymbol{X}_g^\top \boldsymbol{u}^*_g\|^2 = O_P\Big(N \sup_{1\leq g\leq G} N_g^2 \Big) \tag{B.16}$$

because $\boldsymbol{X}_g^\top \boldsymbol{u}^*_g = \sum_{i=1}^{N_g} \boldsymbol{X}_{ig}^\top u^*_{ig}$, where $\boldsymbol{X}_{ig}^\top u^*_{ig}$ is an independent, zero mean sequence with finite variance under the WB probability measure, and by application of Lemma A.1 we thus obtain $\mathrm{E}^* \|\boldsymbol{X}_g^\top \boldsymbol{u}^*_g\|^2 = O_P(N_g)$. Combining (B.11), (B.16), and $\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}} = O_P(N^{-1/2})$, (B.15) is

$$(\boldsymbol{a}^\top \bar{\boldsymbol{V}}_N \boldsymbol{a})^{-1} \boldsymbol{a}^\top \boldsymbol{D}^*_{3N} \boldsymbol{a} = O_{P^*}\Big(N^{-1} \sup_{1\leq g\leq G} N_g \Big) \xrightarrow{P^*} 0,$$

in probability, using (10). Finally, by very similar arguments, we find for $m = 4$ that

$$\|(\boldsymbol{a}^\top \bar{\boldsymbol{V}}_N \boldsymbol{a})^{-1} \boldsymbol{a}^\top \boldsymbol{D}^*_{4N} \boldsymbol{a}\| \leq (\boldsymbol{a}^\top \bar{\boldsymbol{V}}_N \boldsymbol{a})^{-1} \|\boldsymbol{Q}_N^{-1}\|^2 \frac{1}{N^2} \sum_{g=1}^{G} \|\boldsymbol{X}_g^\top \boldsymbol{X}_g\|^2 \|\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}}\|^2 = O_{P^*}\Big(N^{-1} \sup_{1 \leq g \leq G} N_g\Big) \xrightarrow{P^*} 0,$$

in probability, using also Lemma A.2.

Proof of (26). Follows immediately by (24), (25), and Slutsky's Theorem.

References

Arellano, M. (1987). Computing robust standard errors for within groups estimators. *Oxford Bulletin of Economics and Statistics 49*, 431–434.

Bell, R. M. and D. F. McCaffrey (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology 28*, 169–181.

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. 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.

Eicker, F. (1963). Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *Annals of Mathematical Statistics 34*, 447–456.

Gonçalves, S. (2011). The moving blocks bootstrap for panel linear regression models with individual fixed effects. *Econometric Theory 27*, 1048–1082.

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.

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

Kauermann, G. and R. J. Carroll (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association 96*, 1387–1396.

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

MacKinnon, J. G. (2002). Bootstrap inference in econometrics. *Canadian Journal of Economics 35*, 615–645.

MacKinnon, J. G. (2016). Inference with large clustered datasets. *L'Actualité Économique 92*, to appear.

MacKinnon, J. G. and M. D. Webb (2017a). The subcluster wild bootstrap for few (treated) clusters. QED Working Paper 1364, Queen's University.

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

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

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

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 and Economic Statistics 35*, to appear.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica 48*, 817–838.

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