Djogbenou, Antoine A.; MacKinnon, James G.; Nielsen, Morten Ørregard

Working Paper

Asymptotic Theory and Wild Bootstrap Inference with Clustered Errors

Queen's Economics Department Working Paper, No. 1399

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

Suggested Citation: Djogbenou, Antoine A.; MacKinnon, James G.; Nielsen, Morten Ørregard (2018) : Asymptotic Theory and Wild Bootstrap Inference with Clustered Errors, Queen's Economics Department Working Paper, No. 1399, Queen's University, Department of Economics, Kingston (Ontario)

This Version is available at:
https://hdl.handle.net/10419/188911

Queen's Economics Department Working Paper No. 1399

Asymptotic Theory and Wild Bootstrap Inference with Clustered Errors

Antoine A. Djogbenou
Queen's University

James G. MacKinnon
Queen's University

Morten Orregard Nielsen
Queen's University

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

3-2018

Asymptotic Theory and 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

March 1, 2018

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. Under a somewhat stronger set of conditions, we also derive formal Edgeworth expansions for the asymptotic and bootstrap test statistics. Simulation experiments illustrate the theoretical results, and the Edgeworth expansions explain the overrejection of the asymptotic test and shed light on the choice of auxiliary distribution for the wild bootstrap.

Keywords: Clustered data, cluster-robust variance estimator, CRVE, Edgeworth expansion, 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 are correlated within clusters. In such cases, it is very common to use a cluster-robust variance estimator (CRVE) to calculate t-statistics and Wald statistics, because neglecting the cluster structure can lead to severely biased standard errors and large size distortions (Moulton, 1986).

*An earlier version of this paper was circulated under the title "Validity of Wild Bootstrap Inference with Clustered Errors." We are grateful to the editor, Jianqing Fan, an anonymous associate editor, three anonymous referees, Russell Davidson, Silvia Gonçalves, Bruce Hansen, and seminar participants at NY Camp Econometrics XII, the 2017 CEA Annual Meeting, the 2017 CESG Annual Meeting, and U.C. San Diego for comments. MacKinnon thanks the Social Sciences and Humanities Research Council of Canada (SSHRC) for financial support. Nielsen thanks the SSHRC, the Canada Research Chairs (CRC) program, and the Center for Research in Econometric Analysis of Time Series (CREATES, funded by the Danish National Research Foundation, DNRF78) for financial support. Some of the computations were performed at the Centre for Advanced Computing at Queen's University.

†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.

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.

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 in certain cases; see MacKinnon and Webb (2017b). Interestingly, MacKinnon and Webb (2018) 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 are clustered. 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, allow for heteroskedasticity of unknown form, and do not restrict dependence within clusters.

To assess the accuracy of the bootstrap relative to the asymptotic normal approximation, we derive one- and two-term formal Edgeworth expansions under somewhat stronger assumptions. These expansions explain the overrejection of the asymptotic test found in simulations. We apply the expansions to discuss the choice of auxiliary distribution and give conditions under which the wild cluster bootstrap may provide an asymptotic refinement.

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), Carter, Schnepel, and Steigerwald (2017), and Hansen and Lee (2017), among others. All but the last two of these assume that clusters are equal-sized. Hansen and Lee (2017) derives a law of large numbers and a central limit theorem for clustered samples under conditions that are very similar to ours and apply their results to several different estimation problems, including regression, but do not consider bootstrap inference. Carter et al. (2017) considers linear regression with a cluster structure and studies the effects of heterogeneity across clusters, but it makes much stronger assumptions than we do.

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 cluster bootstrap produced less reliable inferences than the wild cluster bootstrap; for additional simulation evidence, see MacKinnon and Webb (2017a). 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 numbers 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 (first-

order) validity of both the wild cluster bootstrap and the ordinary wild bootstrap. In Section 4, we present the results of some simulation studies. In Section 5, we discuss higher-order asymptotic theory, and Section 6 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 $N_g \times N_g$ 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 (2018). 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.

It is easy to see that $\hat{\boldsymbol{V}}$ is singular whenever $k > G$, because the rank of $\hat{\boldsymbol{V}}$ cannot exceed G. This occurs, for example, whenever there are cluster fixed effects. In that case, the dimension of the square matrix $\boldsymbol{X}^\top \boldsymbol{X}$ increases with, and must always exceed, G. Moreover, the diagonal block of $\hat{\boldsymbol{\Gamma}}$ that corresponds to the fixed effects is a zero matrix, because the vector $\hat{\boldsymbol{u}}_g$ must be orthogonal to the fixed effect for cluster g. This may (but typically does not) cause $\boldsymbol{V}_N$ to have zero diagonal

elements for the coefficients of the fixed effects. However, the presence of cluster fixed effects does not prevent us from using (5) to make inferences about the remaining elements of $\boldsymbol{\beta}$.

A readily implemented solution with cluster fixed effects is to project all other regressors off them so that $\boldsymbol{y}$ and $\boldsymbol{X}$ are expressed as deviations from cluster means; see Pustejovsky and Tipton (2018). Let $\boldsymbol{D}_g$ be an $N_g \times G$ matrix with the g^{th} column equal to a vector of 1s and all other elements equal to 0, and let $\boldsymbol{D}$ be the $N \times G$ matrix formed by stacking the $\boldsymbol{D}_g$. Then $\boldsymbol{M_D} = \mathbf{I}_N - \boldsymbol{D}(\boldsymbol{D}^\top \boldsymbol{D})^{-1}\boldsymbol{D}^\top$ is the projection matrix that takes deviations from cluster means, and we can redefine $\boldsymbol{y}$ as $\boldsymbol{M_D y}$ and $\boldsymbol{X}$ as $\boldsymbol{M_D X}$ so as to partial out the fixed effects. Whenever a model originally involves fixed effects, we will assume that our conditions hold for the model involving the transformed data.

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 needed) against a one-sided or two-sided alternative.

We next derive the asymptotic limit theory for t_a. 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 norm.

Assumption 1. The sequence $\{\boldsymbol{X}_g^\top \boldsymbol{u}_g\}$ is independent across g and satisfies, for all $g \in \mathbb{N}$, that $\mathrm{E}(\boldsymbol{u}_g|\boldsymbol{X}) = \boldsymbol{0}$ and $\mathrm{E}(\boldsymbol{u}_g\boldsymbol{u}_g^\top|\boldsymbol{X}) = \boldsymbol{\Omega}_g$, where $\boldsymbol{\Omega}_g$ is positive definite. In addition, for some $\lambda \geq 0$,

$$\sup_{i,g\in\mathbb{N}} \mathrm{E}\|\boldsymbol{X}_{ig}^\top u_{ig}\|^{4+\lambda} < \infty.$$

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

$$\sup_{i,g\in\mathbb{N}} \mathrm{E}\|\boldsymbol{X}_{ig}\|^{4+\lambda} < \infty,$$

where λ is the same as in Assumption 1. Furthermore, there exists a non-random sequence $\{\mu_N\}$ and a non-random, finite scalar $v_a > 0$ such that $\mu_N \to \infty$ and $\mu_N \boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a} \xrightarrow{P} v_a$.

Assumption 3. For λ defined in Assumption 1 and μ_N defined in Assumption 2,

$$G \to \infty \quad \text{and} \quad \mu_N^{\frac{4+\lambda}{6+2\lambda}} \sup_{g\in\mathbb{N}} \frac{N_g}{N} \to 0.$$

Assumption 1 imposes the conditions that $\{\boldsymbol{X}_g^\top \boldsymbol{u}_g\}$ is independent across clusters, with finite $4+\lambda$ moments, and that $\boldsymbol{u}_g$ has zero conditional mean and constant, but possibly heterogeneous, conditional variance matrix. Conditions like the first part of Assumption 2 are standard in asymptotic theory for linear regressions.

Because of the clustered errors in model (1), the order of magnitude of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ depends in a complicated way on the regressors, the relative cluster sizes, the intra-cluster correlation structure, and interactions among these. This is captured in the second part of Assumption 2, where it is assumed that the conditional variance of $\boldsymbol{a}^\top\hat{\boldsymbol{\beta}}$, multiplied by a non-random sequence $\{\mu_N\}$, converges to a finite, non-zero limit. An important consequence of the studentization in our results is that the rate μ_N does not need to be known, but only needs to exist.

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 variance of $\boldsymbol{a}^\top\hat{\boldsymbol{\beta}}$, i.e. the magnitude of $\boldsymbol{a}^\top\boldsymbol{V}_N\boldsymbol{a}$ as represented by (the inverse of) the sequence μ_N, and to the moment condition in Assumption 1. Thus, μ_N can be interpreted as the rate at which information accumulates.

To analyze the role of μ_N, we investigate two extreme cases, with all other cases lying in between: (i) $\boldsymbol{\Omega}_g$ is diagonal with no intra-cluster correlation at all and (ii) $\boldsymbol{\Omega}_g$ is a dense matrix with constant correlations, and the regressors are correlated. In case (i), it straightforwardly holds that

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

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) we find 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 that

$$\|\boldsymbol{V}_N\| = O_P\Big(N^{-1}\sup_{g\in\mathbb{N}}N_g\Big) \quad \text{and} \quad \mu_N = N/\sup_{g\in\mathbb{N}}N_g. \tag{9}$$

Therefore, in case (ii), $\hat{\boldsymbol{\beta}}$ converges at rate $O_P\big(N^{-1/2}\sup_{g\in\mathbb{N}}N_g^{1/2}\big)$. In general, it follows from (7) and (9) that, under Assumptions 1 and 2,

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

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

Clearly, (7) implies a stronger condition in Assumption 3 than (9). Specifically, in case (ii), where the $\boldsymbol{\Omega}_g$ are dense, Assumption 3 is 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 amount of independent information contained in a cluster) is smaller than the actual cluster size (N_g), more heterogeneity in N_g is allowed by the second condition of Assumption 3.

Because the exponent on μ_N in Assumption 3 is decreasing in λ, the condition is stronger when fewer moments are assumed to exist, i.e. when λ is lower, cf. Assumption 1. Thus, a sufficient condition for Assumption 3 that does not depend on λ is

$$G\to\infty \quad \text{and} \quad \mu_N^{2/3}\sup_{g\in\mathbb{N}}\frac{N_g}{N}\to 0. \tag{11}$$

Alternatively, in view of (7) and (9), we can find a sufficient condition for Assumption 3 that does not depend on μ_N, namely,

$$G\to\infty \quad \text{and} \quad \sup_{g\in\mathbb{N}}N_g = o\Big(N^{\frac{2+\lambda}{6+2\lambda}}\Big). \tag{12}$$

The exponent in (12) is increasing in λ, and a sufficient condition that does not depend on either λ or μ_N is that

$$G\to\infty \quad \text{and} \quad \sup_{g\in\mathbb{N}}N_g = o(N^{1/3}). \tag{13}$$

The second condition of Assumption 3, or either of the sufficient conditions in (11)–(13), allow 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., $N_g = c_gN^\alpha$, where c_g and α are fixed constants. The

former case, with the N_g being fixed constants, could be considered a prototypical case. When this holds, then $\hat{\boldsymbol{\beta}}$ is in fact $O_P(G^{-1/2})$; see also Assumption 5 in Section 5.

Because $\mu_N \to \infty$, 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, allowing any intra-cluster correlation structure, including independence, (13) shows that any $\alpha < 1/3$ satisfies Assumption 3. However, in case (ii) above, where the $\boldsymbol{\Omega}_g$ are dense, more heterogeneity of cluster sizes is allowed, and any $\alpha < 1$ satisfies (11). In that case, 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, characterizing the degree of cross-sectional dependence.

Our first result in Theorem 2.1 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. More recently, Carter, Schnepel, and Steigerwald (2017) obtains a result similar to our Theorem 2.1 that allows clusters to be heterogeneous. However, they impose a moment assumption that restricts intra-cluster dependence and rules out, e.g., the random effects model (which is used as their simulation DGP) and even some models with homoskedastic errors that are uncorrelated within clusters. Moreover, they impose very high-level assumptions to restrict cluster-size heterogeneity, and in general it is not clear how to verify, or derive sufficient primitive conditions for, those assumptions. In contrast, our assumptions are primitive and straightforward to interpret. Also very recently (indeed after the first draft of the present paper was written), Hansen and Lee (2017) derives a law of large numbers and a central limit theorem for clustered samples under conditions that are very similar to ours. They apply their results to several different estimation problems, including regression, but do not consider bootstrap inference.

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 et al. (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,

$$\boldsymbol{a}^\top (\boldsymbol{\beta}_N - \boldsymbol{\beta}_0) = (\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{1/2} \delta, \tag{14}$$

which is often referred to as "Pitman drift." Under (14), 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 2.1. *Suppose that Assumptions 1–3 are satisfied and the true value of $\boldsymbol{\beta}$ is given by (14).*

It then holds that

$$\frac{\boldsymbol{a}^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)}{(\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{1/2}} \xrightarrow{d} \mathrm{N}(0,1), \tag{15}$$

$$\frac{\boldsymbol{a}^\top \hat{\boldsymbol{V}} \boldsymbol{a}}{\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a}} \xrightarrow{P} 1, \tag{16}$$

$$t_a \xrightarrow{d} \mathrm{N}(\delta, 1). \tag{17}$$

When the null hypothesis H_0 is true, the following is an immediate consequence of Theorem 2.1.

Corollary 2.1. *Under the assumptions of Theorem 2.1 and H_0, it holds that $t_a \xrightarrow{d} \mathrm{N}(0,1)$.*

The result in Corollary 2.1 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 2.1 and Corollary 2.1 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 (15), which is $(\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{1/2}$, or equivalently $\mu_N^{-1/2}$, instead of $N^{-1/2}$; see also the discussion around (9).

The proof of Theorem 2.1 may be found in Appendix B. In this proof, we make use of the scalars $z_g = v_a^{-1/2} \mu_N^{1/2} N^{-1} \boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} \boldsymbol{X}_g^\top \boldsymbol{u}_g$, 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 $\sup_g N_g$.

Theorem 2.1, specifically (17), 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 two-sided alternative, the probability of rejecting the null hypothesis when the DGP is (14) is given by the asymptotic local power function

$$1 - \Phi(z_{1-\alpha/2} - \delta) + \Phi(-z_{1-\alpha/2} - \delta), \tag{18}$$

where $\Phi(x)$ denotes the cumulative distribution function of the standard normal distribution, and $z_{1-\alpha/2}$ satisfies $\Phi(z_{1-\alpha/2}) = 1 - \alpha/2$. The asymptotic local power function (18) 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 (14). 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 2.1. 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). However, that paper provides no theoretical justification for 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 zero mean and unit variance. 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 we also analyze below. The WB was designed for regression models with independent, heteroskedastic errors but has recently been suggested for the model (1) by MacKinnon and Webb (2018). 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.

3.1 Wild Cluster Bootstrap

We next describe the algorithm needed to implement the WCU and WCR bootstraps for testing the hypothesis H_0 in some detail.[1] 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 realizations of the random variable v^* with zero mean and unit variance;
   (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

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

[1] 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). This is the principal reason for considering WCU.

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

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

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

Our next result demonstrates the validity of the WCB. Let the cumulative distribution function (CDF) of t_a under $\mathrm{H_0}$ be denoted $P_0(t_a \leq x)$. As usual, let P^* denote the probability measure induced by the bootstrap (WCB or WB, as appropriate) conditional on a given sample, and let E^* and Var^* denote the corresponding expectation and variance conditional on a given sample.

Theorem 3.1. *Suppose Assumptions 1–3 are satisfied with $\lambda > 0$, that the true value of $\boldsymbol{\beta}$ is given by (14), and that $\mathrm{E}^*|v^*|^{4+\lambda} < \infty$ with λ as given in Assumption 1. Then, for any $\epsilon > 0$,*

$$P\Big(\sup_{x\in\mathbb{R}} \Big|P^*(t_a^* \leq x) - P_0(t_a \leq x)\Big| > \epsilon\Big) \to 0.$$

When the null hypothesis $\mathrm{H_0}$ is true, that is, when $\delta = 0$ in (14), Theorem 3.1 implies that P values computed in step 4 of the WCU and WCR algorithms are asymptotically valid, as are studentized bootstrap confidence intervals. More generally, Theorem 3.1 shows that, under the sequence of local alternatives (14), the bootstrap distribution $P^*(t_a^* \leq x)$ coincides with that of the original t-statistic under the null hypothesis $\mathrm{H_0}$, $P_0(t_a \leq x)$, in Corollary 2.1. This implies that the WCB test has the same asymptotic local power function (18) as the asymptotic test based on t_a.

3.2 Ordinary Wild Bootstrap

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 clustered 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 realizations of the random variable v^* with zero mean and unit variance.

Note that, although this algorithm relies on the WB to generate the bootstrap errors, u_{ig}^*, and hence the bootstrap data, the WB test statistic is still computed using the CRVE based on the bootstrap data, i.e. using $\hat{\boldsymbol{V}}^*$.

Theorem 3.2. *Suppose that Assumptions 1–3 with $\lambda > 0$ are satisfied, that the true value of $\boldsymbol{\beta}$ is given by (14), and that $\mathrm{E}^*|v^*|^{4+\lambda} < \infty$ with λ as given in Assumption 1. Then, for any $\epsilon > 0$,*

$$P\Big(\sup_{x\in\mathbb{R}} \Big|P^*(t_a^* \leq x) - P_0(t_a \leq x)\Big| > \epsilon\Big) \to 0.$$

Like Theorem 3.1, 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 Theorem 3.2 is obtained under the sequence of local alternatives (14), 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.

Although the result in Theorem 3.2 is identical to that in Theorem 3.1 on the surface, the underlying theory differs in important ways. In particular, the WB is unable to replicate the intra-cluster correlation structure in $\boldsymbol{\Omega}_g$ because the WB multiplies each residual by independent draws of the auxiliary random variable v^*, so that the WB bootstrap DGP has independent (but possibly heteroskedastic) errors, even within clusters. In consequence, the WB estimator $\boldsymbol{a}^\top \hat{\boldsymbol{\beta}}^*$ has a different asymptotic variance matrix (conditional on the original sample) than that of the original sample t-statistic and that of the WCB estimator (conditional on the original sample); cf. (15) and (B.15) in Appendix B. However, the fact that $\boldsymbol{a}^\top \hat{\boldsymbol{\beta}}^*$ has the "wrong" variance does not invalidate the WB, because t_a^* is studentized appropriately and thus has the correct asymptotic distribution.

Furthermore, because the normalization of $\boldsymbol{a}^\top \hat{\boldsymbol{\beta}}^*$ under the WB is in fact of order $N^{1/2}$ (see (B.15) and (B.19) in Appendix B), the distribution of t_a^* for the WB will in general approach the asymptotic $\mathrm{N}(0,1)$ distribution more rapidly than the distribution of t_a. This rules out the possibility of asymptotic refinements for the WB. On the other hand, asymptotic refinements are possible for the WCB, and we investigate them in Section 5. In practice, these issues might well make it more difficult for the WB than for the WCB to mimic the distribution of t_a when μ_N is small, e.g. when G is small or the cluster sizes are heterogeneous and the $\boldsymbol{\Omega}_g$ are dense. We study 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.

Most of our experiments are based on the DGP

$$\boldsymbol{y}_g = \beta_1 + \beta_2 \boldsymbol{x}_g + \boldsymbol{u}_g, \quad \mathrm{E}(\boldsymbol{u}_g \boldsymbol{u}_g^\top) = \boldsymbol{\Omega}_g, \quad g = 1, \ldots, G, \tag{19}$$

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.[2] 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.[3] We obtained very similar results using both methods. The null hypothesis is that $\beta_2 = 0$; this is equivalent to setting $\boldsymbol{a} = [0 \;\; 1]^\top$. Every experiment has 100,000 replications.

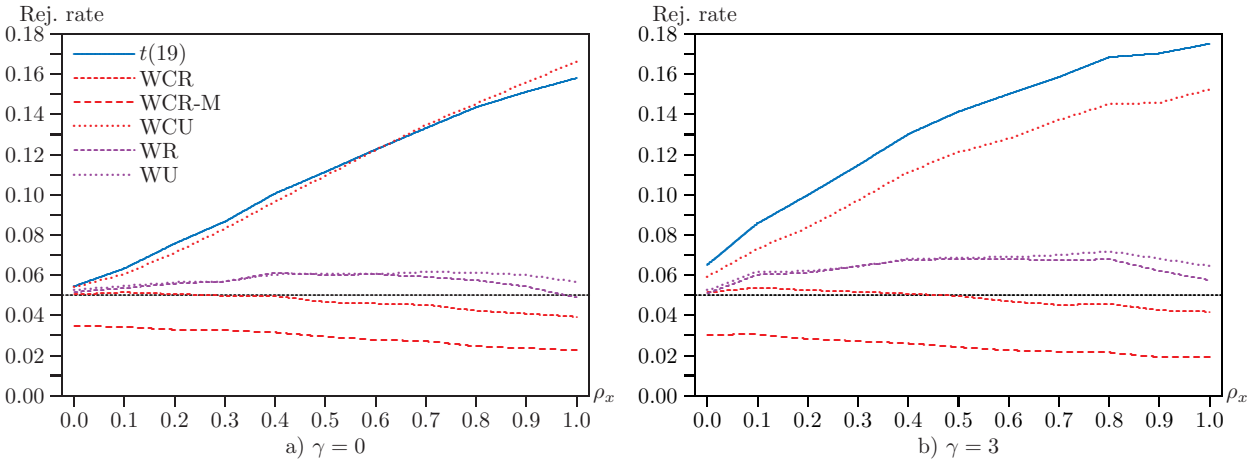
Since we have to impose conditions like Assumption 3 on the cluster sizes, we expect inference to be harder when cluster sizes are not all the same; see MacKinnon and Webb (2017b) for evidence on this point. In order to allow cluster sizes to vary systematically, we initially allocate N observations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

$$N_g = \left[\frac{N \exp(\gamma g / G)}{\sum_{j=1}^{G} \exp(\gamma j / G)} \right], \quad \text{for } g = 1, \ldots, G-1, \tag{20}$$

[2] Specifically, $u_{ig} = (1-\rho)^{1/2} \varepsilon_{ig} + \rho^{1/2} e_g$, where ε_{ig} and e_g are mutually independent i.i.d. N(0,1) random variables.

[3] Let $v_{m,ig} = (1-\rho_1)^{1/2} \varepsilon_{m,ig} + \rho_1^{1/2} e_{m,g}$, $m = 1, 2$, where all component random variables are i.i.d. $\mathrm{N}(0,1)$, so that both $v_{1,ig}$ and $v_{2,ig}$ are $\mathrm{N}(0,1)$ with intra-cluster correlation ρ_1. Then u_{ig} equals $\mu_1 + \sigma_1 v_{1,ig}$ with probability p and $\mu_2 + \sigma_2 v_{2,ig}$ with probability $1-p$. To obtain the desired moments and correlations for u_{ig}, in particular intra-cluster correlation $\rho = 0.1$, we used $p = 0.1967$, $\mu_1 = 0.7693$, $\mu_2 = -0.1884$, $\sigma_1 = 1.5734$, $\sigma_2 = 0.6770$, and $\rho_1 = 0.2556$.

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



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, cluster sizes become 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 (before taking the logarithm) of ρ_x, and the error terms are generated by the normal mixture model described above.[4] Figure 1 shows rejection frequencies for five tests at the .05 level when $G = 20$, $N = 4000$, and $\rho = 0.1$. In panel a), where $\gamma = 0$, all clusters have 200 observations. In panel b), where $\gamma = 3$, which is quite a large value, cluster sizes vary from 33 to 598. 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.

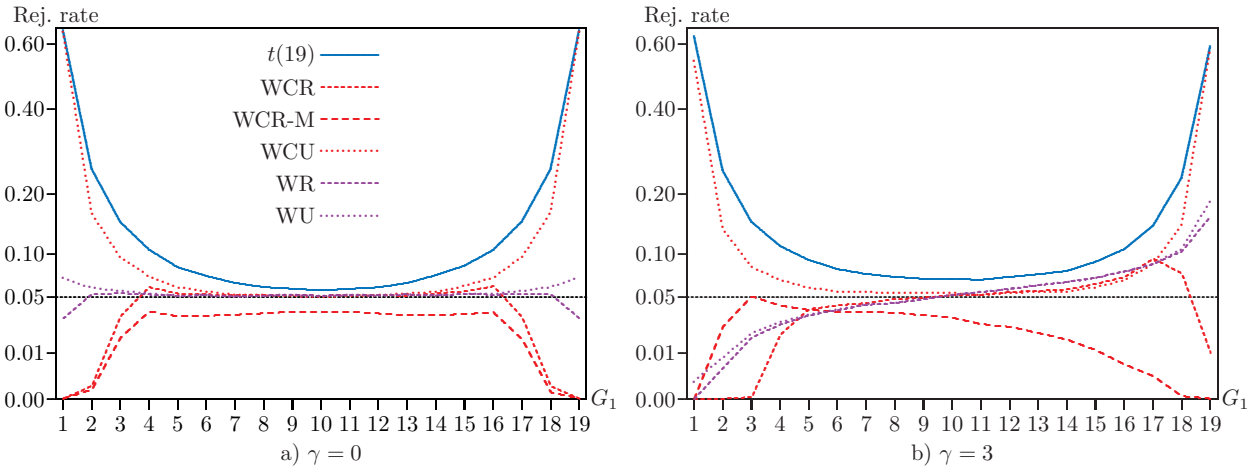
Throughout, we compare bootstrap rejection frequencies with ones for the cluster-robust t-test as implemented in STATA. In particular, we use critical values taken from the $t(G-1)$ distribution instead of the standard normal, as advocated by Bester, Conley, and Hansen (2011), and the CRVE is the one in (5) multiplied by the factor $G(N-1)/((G-1)(N-k))$. Without this factor, or if we had used the standard normal distribution instead of the $t(G-1)$ distribution, the overrejection that is evident in Figure 1 would have been even more severe. For all the bootstrap tests, we report symmetric P values based on $B = 399$ bootstrap samples, where the v^* are drawn from the Rademacher distribution. For the WCR bootstrap test, we also report results using the two-point Mammen (1993) auxiliary distribution, which are labelled WCR-M in the figures.

Both the cluster-robust t-test and the WCU bootstrap test always overreject, and they do so more severely as ρ_x increases. In contrast, the WCR bootstrap works very well in all cases, although it tends to underreject slightly for larger values of ρ_x. However, when the Mammen distribution is used instead of the Rademacher, the WCR bootstrap underrejects quite severely. The reasons for the poor performance of this variant of the WCR bootstrap are analyzed in Section 5.2 using higher-order asymptotic theory. The two ordinary wild bootstraps (WR and WU) perform almost perfectly when $\rho_x = 0$, overreject somewhat for moderate values of ρ_x, but then improve as ρ_x approaches 1. For $\rho_x = 1$, WR actually outperforms WCR in both panels of Figure 1.

Since our focus is on the bootstrap, the only non-bootstrap procedure for which we report

[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.

Figure 2: Rejection frequencies for treatment dummy, $G = 20$, $N = 4000$, $\rho = 0.10$



results is the test implemented in STATA. It is apparent in Figure 1 that rejection frequencies for that test are extremely sensitive to the value of ρ_x. In an actual empirical application, its performance could have been predicted by computing the feasible version of the "effective number of clusters" proposed in Carter, Schnepel, and Steigerwald (2017) and called G^{*A}. For example, in panel a) of Figure 1, the average value of G^{*A} declines from 7.59 to 4.23 as ρ_x increases from 0.0 to 1.0. In panel b), it declines from 5.93 to 3.64. Thus the value of G^{*A} correctly predicts that the usual test will perform better in panel a) than in panel b), especially when ρ_x is small, and that its performance will deteriorate sharply as ρ_x increases.

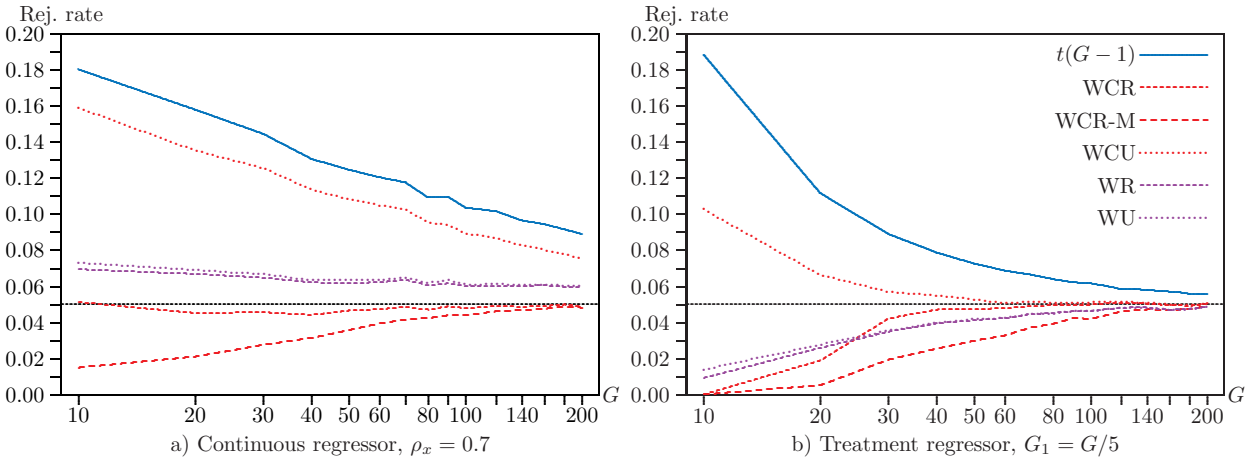
In the next two experiments, a typical element of the test regressor in (19) is a dummy variable that equals 1 for some clusters and 0 for others; it can be thought of as a cluster-level treatment dummy. Many applications of cluster-robust inference involve this type of variable, and it is well-known that inference can be problematical when the number of treated, or untreated, clusters is small; see MacKinnon and Webb (2017a,b). 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 2 shows rejection frequencies for the same five tests when the regressor is a treatment dummy that equals 1 for G_1 out of $G = 20$ clusters with $N = 4000$. Once again, the error terms are drawn from a normal mixture model. In panel a), the clusters are equal in size, with $N_g = 200$. 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. A more complete analysis and explanation of these extreme overrejections and underrejections in the "few treated clusters" case is given in MacKinnon and Webb (2017b, Section 6). However, all the bootstrap methods except WCR-M work very well for $4 \leq G_1 \leq 16$.

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 (2018) 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.

In panel b) of Figure 2, rejection frequencies are shown for a case in which $\gamma = 3$ and clusters

Figure 3: Rejection frequencies as G changes, $\gamma = 3$, $\rho = 0.10$



are treated from smallest to largest.[5] 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 (2018). WCU generally overrejects more 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 $10 \leq G_1 \leq 18$. WCR-M performs surprisingly well for $G_1 = 2$ and $G_1 = 3$, but it underrejects very severely for large values of G_1.

The situation depicted in panel b) of Figure 2 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 panel b) but not as well as they do in panel a).

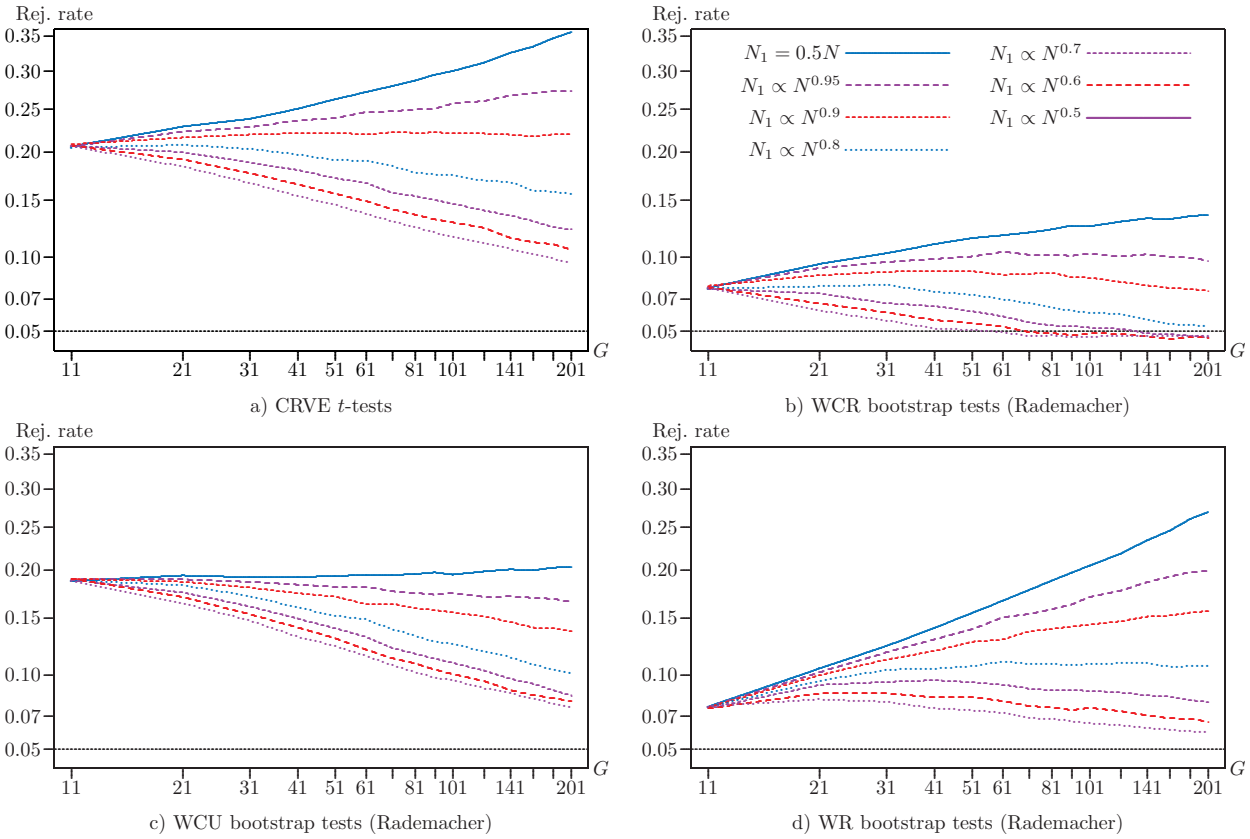
In the next two experiments, we vary the number of clusters G and the sample size together. The results are shown in Figure 3. In panel a), the regressor is continuous, as in Figure 1. We fix ρ_x at 0.7 (which is one of the worst values for the ordinary wild bootstrap tests) and vary G from 10 to 100 by 10 and then from 120 to 200 by 20. The value of γ is 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 sample size is $N = 40,000$.

There are four striking results in panel a) of Figure 3. The first is that all the bootstrap tests reject far less often than the t-test. The second is that WCR performs very much better than WCU. 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 may be 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 Mammen version of WCR underrejects severely when G is small, but the underrejection essentially disappears by the time $G = 200$. The final 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.

In panel b) of Figure 3, we consider what happens as G increases when the regressor is a

[5] 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 4: Rejection frequencies for four tests, continuous regressor with one big cluster

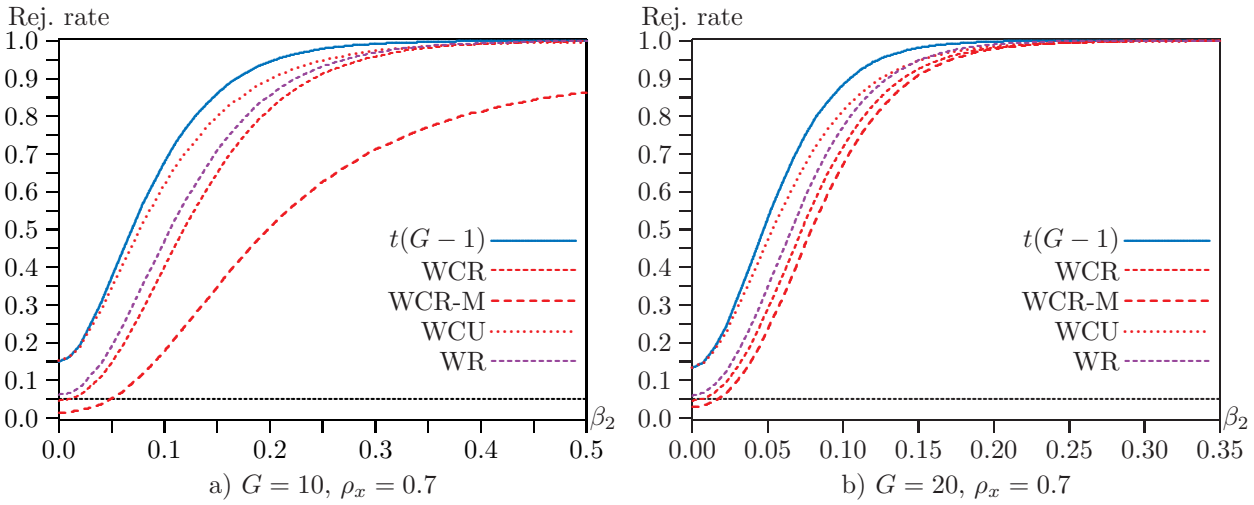


treatment dummy. As in panel b) of Figure 2, $\gamma = 3$. The fraction of treated clusters is held constant, with $G_1/G = 0.2$, and the rejection frequencies for $G = 20$ correspond to the ones for $G_1 = 4$ in panel b) of Figure 2. As the results of Section 3 suggest, all methods improve steadily as G increases. However, the two wild cluster bootstrap methods that use the Rademacher distribution evidently improve faster than WCR-M and the two ordinary wild bootstrap methods. For $G \geq 30$, the best methods are clearly WCR and WCU. These results are consistent with those in panel a), although WCR no longer seems to have a clear advantage over WCU.

In Figure 3, the largest cluster constitutes 27.5% of the sample for $G = 10$ but only 1.8% for $G = 200$. In the next set of 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 $\alpha = 0.5$, $N_1/N = 0.112$.

The four panels of Figure 4 show rejection frequencies for CRVE t-tests and three bootstrap 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 for the CRVE t-test,

Figure 5: Simulated power for continuous regressor, $\gamma = 0$, $\rho = 0.1$



in panel a), 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 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.

Panel b) shows rejection frequencies for the WCR bootstrap for the same set of experiments. These are much smaller than the ones for the CRVE t-test in panel a). 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 slightly.

In contrast, we see from panel c) that rejection frequencies for the WCU bootstrap are quite high when $G = 11$, but they decrease with G for all values of α except $\alpha = 1$. Overall, this procedure always works at least somewhat better than the CRVE t-test, especially for larger values of G. Finally, we see from panel d) 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 as G increases, except for the smallest values of α where WR clearly improves as G increases.

Up to this point, we have only studied test size. Figure 5 investigates the power of alternative tests for the continuous regressor (lognormal) case. The horizontal axis shows the true value of β_2 for tests of $\beta_2 = 0$. All clusters have 200 observations. In the left panel, there are 10 clusters, and in the right panel there are 20. For both values of G, using the $t(G-1)$ distribution leads to substantial overrejection under the null hypothesis and therefore to apparently high (but meaningless) power. Interestingly, however, WCU overrejects just as severely under the null but has noticeably less power for large values of β_2. WCR performs extremely well under the null and therefore has meaningful power. WCR-M is severely lacking in power for $G = 10$, much more so than the extent of its underrejection under the null would suggest, and even for $G = 20$ it has slightly lower power.

Figure 6 investigates power for the treatment dummy case. In the left panel, $G = 10$ and $G_1 = 2$, and in the right panel, $G = 20$ and $G_1 = 4$. In both cases, WCU is seriously lacking in power for large values of β_2, even though it overrejects very substantially under the null. In contrast, even though WCR underrejects severely under the null when $G = 10$, it has more power than any of the other bootstrap tests for large values of β_2. Once again, WCR-M is grossly lacking in power for $G = 10$ but performs quite well for $G = 20$. In the latter case, it actually has more power than WCU for large values of β_2, although it still has less power than WR and WCR.

Figure 6: Simulated power for treatment dummy, $\gamma = 0$, $\rho = 0.1$

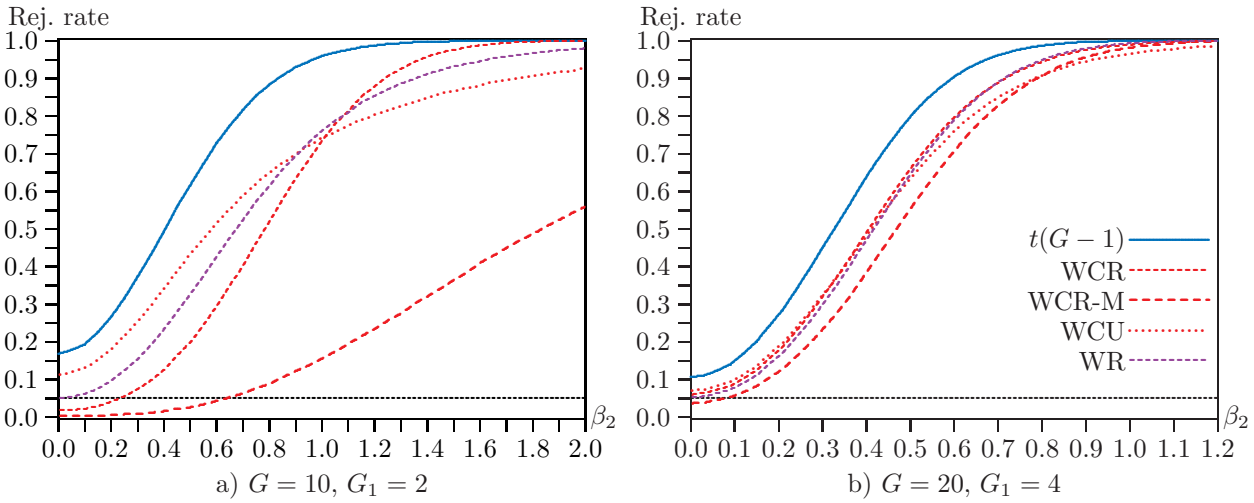
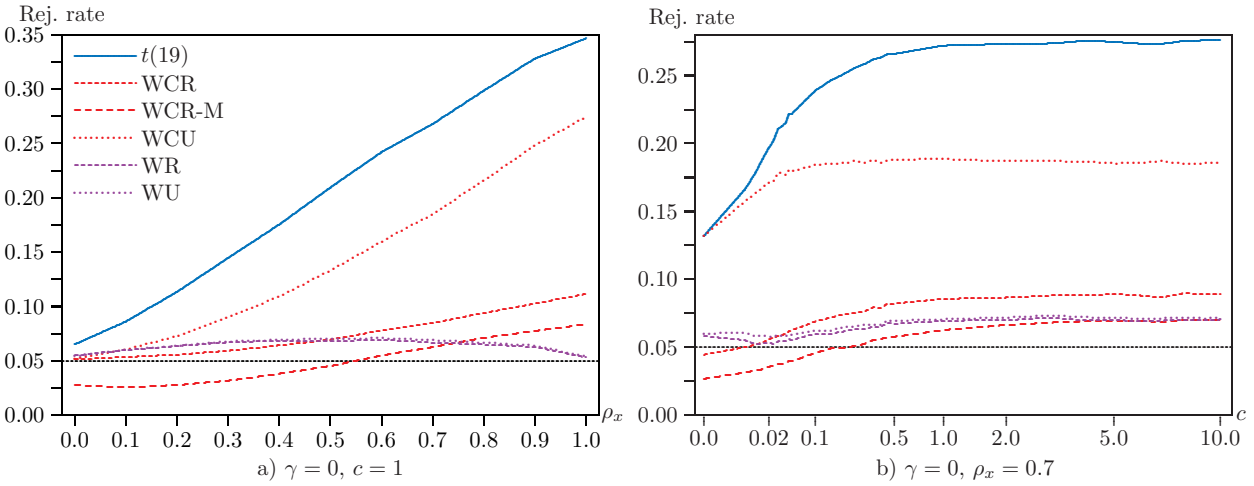


Figure 7: Rejection frequencies with heteroskedastic errors, $G = 20$, $N = 4000$, $\rho = 0.10$

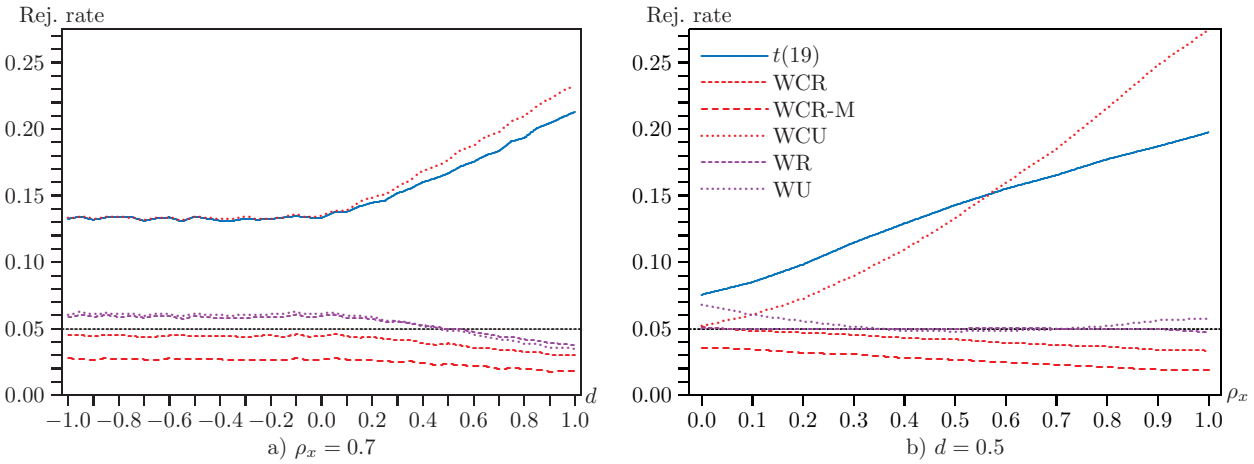


Overall, these results favor the WCR bootstrap using the Rademacher distribution, even in cases where it underrejects under the null, as in the left panel of Figure 6. However, the fact that all the tests seem to be converging to similar power functions as G increases from 10 to 20, which continues (in results that are not reported) as G increases from 20 to 40, suggests that asymptotic theory probably provides a good guide to the power of all tests provided G is not too small.

In all the experiments reported so far, the error terms are homoskedastic. Simulation results in MacKinnon and Webb (2018) suggest that, when error variances differ across clusters, several procedures, including the asymptotic test and the WCB, can be less reliable than in the homoskedastic case. Those results were for difference-in-differences regressions. Here we investigate the effects of heteroskedasticity in the model (19) with a lognormal regressor. The error terms in that equation are now multiplied by $(1 + cx_{ig}^2)^{1/2}$, where c is a constant that we specify. When $c = 0$, the errors are homoskedastic, as before, and as c increases the errors are increasingly heteroskedastic.

The left panel of Figure 7 is comparable to the left panel of Figure 1. In both cases, there are

Figure 8: Rejection frequencies with heterogeneous regressor, $G = 20$, $N = 4000$, $\gamma = 0$, $\rho = 0.10$



20 clusters, each with 200 observations. However, in Figure 7, the value of c is 1, which implies that there is substantial heteroskedasticity. Even when $\rho_x = 0$, so that the heteroskedasticity is solely at the individual level, all procedures perform a bit less well in Figure 7 than in Figure 1. As ρ_x increases, so that more and more of the heteroskedasticity is at the cluster level, the differences between the two figures become much more striking. For larger values of ρ_x, the conventional procedure based on $t(19)$ critical values overrejects much more severely than it did before. So does the WCU bootstrap, although it now performs better relative to the conventional procedure. Instead of underrejecting for large values of ρ_x, the WCR bootstrap now overrejects for both the Mammen and Rademacher distributions. The only procedures that perform about the same as before are the two ordinary wild bootstraps, WR and WU. They both work extremely well for $\rho_x = 0$ and $\rho_x = 1$, but they overreject slightly for intermediate values.

The right panel of Figure 7 shows rejection frequencies as a function of c for $\rho_x = 0.7$. Note that the horizontal axis has been subjected to a cube root transformation, because rejection frequencies are most sensitive to the value of c when it is very small. Even a small amount of heteroskedasticity that varies at the cluster level evidently has a noticeable effect on rejection frequencies. On the other hand, the difference between $c = 1$ (the case in the left panel) and $c = 10$ is quite small.

In all the experiments with a continuous regressor reported so far, the regressor was lognormally distributed, with the same distribution for all clusters. In Figure 8, we relax this assumption by allowing for heterogeneity across clusters. We introduce a parameter $d \geq -1$ which is used to generate the elements x_{ig} of the vector $\boldsymbol{x}_g$ in eq. (19) according to

$$x_{ig} = \exp\Big(\Big(1 + d\frac{g-1}{G-1}\Big)w_{ig}\Big), \tag{21}$$

where the w_{ig} are distributed as $\mathrm{N}(0,1)$, independent across clusters but with correlation ρ_x between w_{ig} and w_{jg} in the same cluster. The DGP in eq. (21) causes both the variance and the higher moments of the x_{ig} to decrease with g for $d < 0$ and to increase with g for $d > 0$. There is no effect on the first cluster, and the effect is largest for the G^{th} one. Even for relatively small values of d, there is substantial heterogeneity across clusters. In practice, we would be surprised to encounter heterogeneity as extreme as that for the larger values of d in the left panel of the figure.

The left panel of Figure 8 shows rejection frequencies as functions of d when $\rho_x = 0.7$, and the right panel shows them as functions of ρ_x when $d = 0.5$. Not surprisingly, the effects of both

parameters depend strongly on the value of the other. The value of d has very little effect for $d < 0$. In contrast, as d increases above 0, the t-test and the WCU bootstrap overreject more and more severely, and the two restricted wild cluster bootstraps underreject slightly more. The ordinary wild bootstrap tests (both WR and WU) overreject slightly for negative and small positive values of d, but they also start to underreject as d becomes relatively large.

In the right panel of Figure 8, we see that the effect of ρ_x is much stronger for WCU when $d = 0.5$ than it is when $d = 0$ (the latter situation is depicted in the left panel of Figure 1). For large values of ρ_x, WCU actually overrejects quite a lot more severely than the t-test. On the other hand, for this particular value of d, WR works remarkably well for all values of ρ_x, while WU is moderately sensitive to the value of ρ_x.

The results in Figure 8 suggest that heterogeneity in the regressors across clusters can significantly affect the performance of some of the methods we consider. The least affected methods are the two ordinary wild bootstrap procedures and the WCR using the Rademacher auxiliary distribution, which seem quite robust to heterogeneity of the type considered here.

5 Higher-Order Asymptotic Theory

In this section, we first derive formal Edgeworth expansions of the CDFs of the sample t-statistic and the WCB t-statistic. We apply these expansions to investigate the impact of the choice of auxiliary distribution in the WCB and to study whether the WCB can yield an asymptotic refinement over the normal approximation under H_0; that is, whether the difference between $P^*(t_a^* \leq x)$ and $P_0(t_a \leq x)$ in Theorem 3.1 can be made smaller than $o_P(1)$, uniformly in x.

5.1 Formal Edgeworth Expansions

For the higher-order theory, the analysis will be exclusively under the null hypothesis, so that P and P_0 are the same, and to simplify notation we use only the former. Furthermore, we strengthen Assumptions 2 and 3 as follows.

Assumption 4. The regressor matrix $\boldsymbol{X}$ is non-random and satisfies $\boldsymbol{Q}_N \to \boldsymbol{Q}$, where $\boldsymbol{Q}$ is finite and positive definite.

Assumption 5. The number of clusters $G \to \infty$, and the cluster sizes satisfy $\sup_{g\in\mathbb{N}} N_g < \infty$.

In Assumption 4, we assume that the regressor $\boldsymbol{X}$ is non-random, which is necessary to keep the theory tractable. Furthermore, Assumption 4 implies that Assumption 1 reduces to:

Assumption 6. The errors $\{\boldsymbol{u}_g\}$ are independent across g and satisfy, for all $g \in \mathbb{N}$, that $\mathrm{E}(\boldsymbol{u}_g) = \boldsymbol{0}$, $\mathrm{E}(\boldsymbol{u}_g \boldsymbol{u}_g^\top) = \boldsymbol{\Omega}_g$, where $\boldsymbol{\Omega}_g$ is positive definite, and $\sup_{i,g\in\mathbb{N}} \mathrm{E}|u_{ig}|^{4+\lambda} < \infty$ for some $\lambda > 0$.

Although Assumption 6 is implied by Assumptions 1 and 4, we include it here for ease of reference. In what follows, we shall also make use of Assumption 6 for a higher value of λ than previously (where only $\lambda > 0$ was assumed), i.e. a stronger moment condition relative to Assumption 1.

We note that, under Assumption 5, the rates μ_N, N, and G are asymptotically proportional. This must be the case because, as $N \to \infty$, no cluster can have more than $N_c^{\max} = \sup_{g\in\mathbb{N}} N_g < \infty$ observations. Therefore, eventually, G must be proportional to N. The rate of convergence of $\ddot{\boldsymbol{\beta}}$ can be described in terms of (the square-root of) any of the three rates. That is, for some positive, finite constants c_1, c_2, and c_3,

$$\frac{\mu_N}{N} \to c_1, \quad \frac{G}{N} \to c_2, \quad \frac{G}{\mu_N} \to c_3, \quad \sqrt{G}(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = O_P(1), \quad \text{and} \quad \mathrm{E}\|\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 = O(G^{-1}); \tag{22}$$

see also Theorem 2.1 and (B.8). Many summations that will be encountered in the higher-order theory contain G terms, and, to avoid an asymptotic factor of proportionality, it will be important to use $\sqrt{G}$ as the rate of convergence of $\ddot{\boldsymbol{\beta}}$. Consequently, all expansions will be in terms of powers of $\sqrt{G}$. This once more emphasizes the important role of G, and not N, as the most relevant notion of sample size in the context of cluster-robust inference.

We consider both one- and two-term Edgeworth expansions. Following, e.g., Hall (1992, Ch. 2), the formal m-term Edgeworth expansion ($m = 1, 2$) of the CDF of t_a is given, uniformly in x, by

$$P(t_a \leq x) = \Phi(x) + \sum_{j=1}^{m} G^{-j/2} q_j(x)\phi(x) + o(G^{-m/2}), \tag{23}$$

where Φ and ϕ are the standard normal CDF and probability density function (PDF), respectively, and q_1 and q_2 are even and odd functions, respectively. For the bootstrap, the formal expansion is

$$P^*(t_a^* \leq x) = \Phi(x) + \sum_{j=1}^{m} G^{-j/2} \ddot{q}_j(x)\phi(x) + o_P(G^{-m/2}), \tag{24}$$

where $\ddot{q}_1$ and $\ddot{q}_2$ are even and odd functions, respectively. The bootstrap is said to provide an asymptotic refinement if the first or both of the higher-order terms of the CDFs of t_a and t_a^* agree, i.e., if $\ddot{q}_1(x) \xrightarrow{P} q_1(x)$ uniformly in x and possibly also $\ddot{q}_2(x) \xrightarrow{P} q_2(x)$ uniformly in x.

Furthermore, for two-sided symmetric tests, we have the formal two-term ($m = 2$) expansion

$$P(|t_a| \leq x) = P(t_a \leq x) - P(t_a \leq -x) = 2\Phi(x) - 1 + 2G^{-1} q_2(x)\phi(x) + o(G^{-1}), \quad x \geq 0, \tag{25}$$

because ϕ and q_1 are even functions, while q_2 is an odd function, and similarly for the bootstrap counterpart. Thus, q_1 plays no role in two-term Edgeworth expansions for two-sided symmetric tests, where the bootstrap provides an asymptotic refinement if $\ddot{q}_2(x) \xrightarrow{P} q_2(x)$ uniformly in x.

To find the functions q_j and $\ddot{q}_j$, for $j = 1, 2$, we first write the sample t-statistic as

$$t_a = \left(\frac{1}{G} \sum_{g=1}^{G} \frac{\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} \frac{G}{N^2} (\boldsymbol{X}_g^\top \hat{\boldsymbol{u}}_g \hat{\boldsymbol{u}}_g^\top \boldsymbol{X}_g) \boldsymbol{Q}_N^{-1} \boldsymbol{a}}{\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a}} \right)^{-1/2} \frac{1}{\sqrt{G}} \sum_{g=1}^{G} \frac{\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} \frac{\sqrt{G}}{N} \boldsymbol{X}_g^\top \boldsymbol{u}_g}{(\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-1/2}}$$

and then we use the decomposition $\hat{\boldsymbol{u}}_g = \boldsymbol{u}_g - \boldsymbol{X}_g(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ to rewrite

$$t_a = \left(\frac{1}{G} \sum_{g=1}^{G} W_g^2 + \frac{1}{G} \sum_{g=1}^{G} Z_g^2 - \frac{2}{G} \sum_{g=1}^{G} W_g Z_g \right)^{-1/2} \frac{1}{\sqrt{G}} \sum_{g=1}^{G} W_g, \tag{26}$$

where we have defined

$$W_g = (\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-1/2} \boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} \frac{\sqrt{G}}{N} \boldsymbol{X}_g^\top \boldsymbol{u}_g, \text{ and} \tag{27}$$

$$Z_g = (\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-1/2} \boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} \frac{\sqrt{G}}{N} \boldsymbol{X}_g^\top \boldsymbol{X}_g \boldsymbol{Q}_N^{-1} \frac{1}{N} \boldsymbol{X}^\top \boldsymbol{u}. \tag{28}$$

We define W_g^* and Z_g^* entirely analogously, simply replacing the error vector $\boldsymbol{u}_g$ with its bootstrap analog $\boldsymbol{u}_g^*$ and replacing $\boldsymbol{V}_N$ with $\ddot{\boldsymbol{V}}$.

We note from (26) that Z_g, and specifically the term $G^{-1} \sum_{g=1}^{G} Z_g^2 - 2G^{-1} \sum_{g=1}^{G} W_g Z_g$, arise from the estimation of the asymptotic variance using residuals $\hat{\boldsymbol{u}}$ rather than errors $\boldsymbol{u}$, and thus reflect the bias in this estimation.

Theorem 5.1. *Suppose Assumptions 4–6 are satisfied with $\lambda = 2m$ and that H_0 is true. Then the formal m-term Edgeworth expansions of the CDF of t_a are given by (23) for $m = 1, 2$, while that of $|t_a|$ is given by (25) for $m = 2$ with*

$$q_1(x) = \frac{1}{6}\gamma_N(2x^2+1) \quad \textit{and}$$
$$q_2(x) = -\frac{1}{2}(2\gamma_N^2 - \tau_{1N} + 2\tau_{2N})x - \frac{1}{24}(16\gamma_N^2 - 2\xi_N - 6\tau_{1N} - 6\tau_{3N})(x^3 - 3x) - \frac{1}{18}\gamma_N^2(x^5 - 10x^3 + 15x),$$

where $\gamma_N = G^{-1}\sum_{g=1}^G \mathrm{E}(W_g^3)$, $\xi_N = G^{-1}\sum_{g=1}^G \mathrm{E}(W_g^4)$, and

$$\tau_{1N} = G^{-1}\sum_{g_1,g_2,g_3=1}^G \mathrm{E}(W_{g_1}W_{g_2}Z_{g_3}^2) - 4G^{-1}\sum_{g_1,g_2=1}^G \mathrm{E}(W_{g_1}^2 W_{g_2}Z_{g_1}), \quad \tau_{2N} = \sum_{g=1}^G \mathrm{E}(W_g Z_g), \quad \tau_{3N} = \sum_{g=1}^G \mathrm{E}(Z_g^2).$$

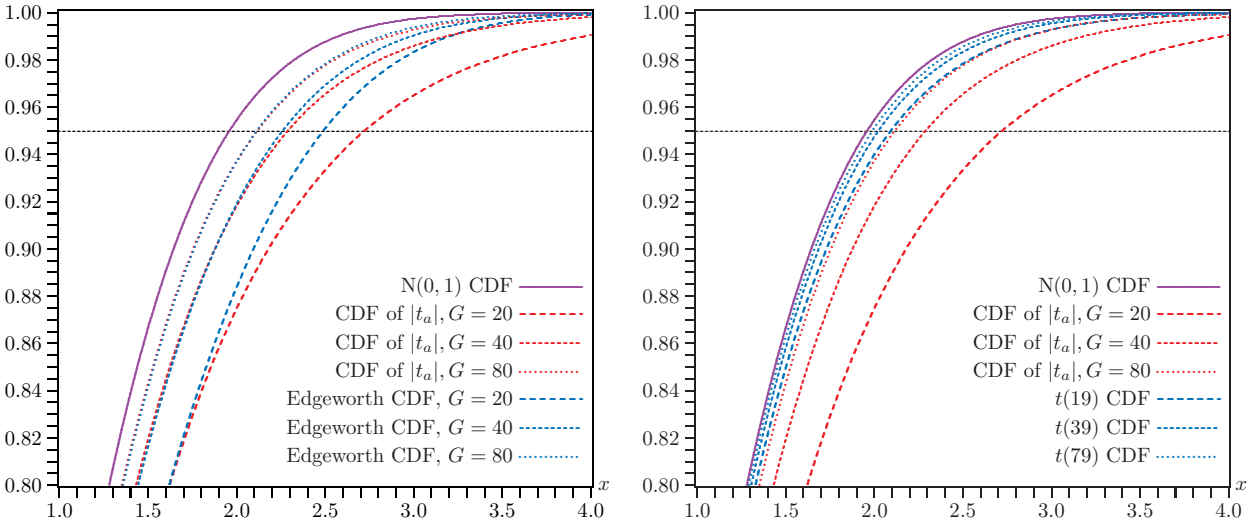
If, in addition, $\mathrm{E}^*|v^*|^{4+2m} < \infty$, then the formal Edgeworth expansions of the CDFs of t_a^* and $|t_a^*|$ are given by the same expressions as those of t_a, but with $\ddot{q}_j$ instead of q_j; see also (24). The functions $\ddot{q}_j$ are obtained from q_j by replacing the population mean $\mathrm{E}(\cdot)$ by the bootstrap analog $\mathrm{E}^*(\cdot)$ and replacing W_g and Z_g by W_g^* and Z_g^*, respectively.

Validity of the formal Edgeworth expansions given in Theorem 5.1 requires further regularity conditions. In particular, for the validity of the Edgeworth expansion of the CDF of t_a in (23), a sufficient condition would be "Cramér's condition" on the characteristic function of $\boldsymbol{u}$; see, e.g., Hall (1992, Thm. 2.2). This condition is satisfied if the distribution of $\boldsymbol{u}$ is sufficiently smooth (has a nondegenerate absolutely continuous component). A similar condition would be required on the characteristic function of the wild bootstrap auxiliary random variables v_g^*. In the bootstrap literature there are two common approaches. In one approach, the Cramér condition is imposed, which is theoretically appealing but rules out all commonly applied discrete distributions for v_g^*. See, for example, Liu (1988) or Kline and Santos (2012). Another approach, see e.g. Mammen (1993), is to continue the analysis without discussing Cramér's condition further, and instead focus on using the formal Edgeworth expansions to theoretically explain the overrejection of the asymptotic test and superiority of the bootstrap in finite samples, and also shed light on the choice of the distribution of the auxiliary random variables, v_g^*. We follow the latter approach.

To assess the accuracy of our Edgeworth expansions, we plot in panel a) of Figure 9 the empirical CDFs of $|t_a|$ for 20, 40, and 80 clusters together with the corresponding two-term Edgeworth expansions, which are given in (25) and Theorem 5.1. The setup is the same as that in panel b) of Figure 3, since the treatment regressor can reasonably be argued to satisfy Assumption 4. The standard normal CDF is also included for reference. As a benchmark, we plot in panel b) of Figure 9 the same empirical CDFs together with the CDFs of the $t(19)$, $t(39)$, and $t(79)$ distributions, which are commonly used for inference, as was also the case in Figure 3.

Comparing panels a) and b) of Figure 9, it is clear that the Edgeworth CDFs provide a very substantial improvement over both the reference normal approximation and the t-distribution CDFs. Following the 0.95 percentile horizontally across panel b), we note that the t-distribution CDFs are very far from the empirical CDFs of $|t_a|$, leading to the severe overrejection of the asymptotic test documented in Figure 3. On the other hand, the Edgeworth CDFs track the empirical CDFs of $|t_a|$ extremely closely, except for $G = 20$ in the very tail of the distribution. Again, following the 0.95 percentile horizontally across panel a), the Edgeworth CDFs can perfectly explain the overrejection of the asymptotic test for $G = 40$ and $G = 80$, and almost for $G = 20$.

Figure 9: Edgeworth expansions of two-sided test, treatment dummy, $\gamma = 3$, $\rho = 0.10$, $G_1 = G/5$



a) Absolute test statistics and Edgeworth approximations

b) Absolute test statistics and Student's t approximations

5.2 Refinements and Choice of Auxiliary Distribution

Given our formal expansions in Theorem 5.1, the second-order bootstrap error in estimating $P(t_a \leq x)$ is given, uniformly in x, by

$$P^*(t_a^* \leq x) - P(t_a \leq x) = G^{-1/2}\frac{1}{6}(\ddot{\gamma}_N - \gamma_N)(2x^2+1)\phi(x) + o_P(G^{-1/2}).$$

The next theorem gives an expansion of $\ddot{\gamma}_N - \gamma_N$, and hence conditions under which the formal Edgeworth expansions for the sample t-statistic and the bootstrap t-statistic agree up to $o_P(G^{-1/2})$.

To distinguish between the restricted and unrestricted versions of the WCB, we let A take the values R or U depending on whether the restricted or unrestricted version is considered. We define

$$Z_g(A) = (\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-1/2}\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1}\frac{\sqrt{G}}{N}\boldsymbol{X}_g^\top \boldsymbol{X}_g \boldsymbol{M}_N(A)\boldsymbol{Q}_N^{-1}\frac{1}{N}\boldsymbol{X}^\top \boldsymbol{u}, \quad A \in \{U, R\}, \tag{29}$$

where $\boldsymbol{M}_N(A) = \boldsymbol{I}_k - \boldsymbol{Q}_N^{-1}\boldsymbol{a}(\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1}\boldsymbol{a})^{-1}\boldsymbol{a}^\top \mathbb{I}(A = R)$ and $\mathbb{I}(A = R)$ equals one if and only if the restricted estimator is considered. Note that $Z_g(U) = Z_g$, which was defined in (28).

Theorem 5.2. *Suppose Assumptions 4–6 are satisfied with $\lambda = 2$, that $\mathrm{E}^*|v^*|^6 < \infty$, and that H_0 is true. Then it holds that*

$$\ddot{\gamma}_N - \gamma_N = \gamma_N(\mathrm{E}^*(v^{*3}) - 1) + O_P(G^{-1/2}).$$

If, in addition, we assume $\lambda > 2$ then

$$\ddot{\gamma}_N - \gamma_N = \gamma_N(\mathrm{E}^*(v^{*3}) - 1) + G^{-1/2}\mathrm{E}^*(v^{*3})\omega_N(A)\mathcal{Z}_N + O_P(G^{-1}),$$

where $\mathcal{Z}_N \xrightarrow{d} \mathrm{N}(0,1)$ and $\omega_N(A), A \in \{U, R\}$, is defined in (B.39).

The leading term in the expansion of $\ddot{\gamma}_N - \gamma_N$ in Theorem 5.2 is $\gamma_N(\mathrm{E}^*(v^{*3}) - 1)$. The theorem thus establishes a second-order asymptotic refinement of the WCB when this leading term is zero. This is stated as a corollary.

Corollary 5.1. *Under the conditions of Theorem 5.2, it holds that* $\ddot{q}_1(x) \xrightarrow{P} q_1(x)$ *uniformly in* x *if and only if either (i)* $\mathrm{E}(u_{ig}^3) = 0$ *for all* i, g *or (ii)* $\mathrm{E}^*(v^{*3}) = 1$. *Under either of these two circumstances,*

$$\sup_{x\in\mathbb{R}} \left| P^*(t_a^* \leq x) - P(t_a \leq x) \right| = o_P(G^{-1/2}).$$

Theorem 5.2 and Corollary 5.1 show that the WCB achieves a second-order refinement under either of two circumstances. The first is when the errors have third moment equal to zero, where it follows easily that $\gamma_N = 0$. The second is when the distribution of the auxiliary random variable v^* has third moment equal to one. This resembles the results found for the wild bootstrap by Wu (1986), Liu (1988), and Mammen (1993). Indeed, our results specialize to their results in the special case with $N_g = 1$ for all g. We next give some further results on the second term in the Edgeworth expansions, and subsequently we return to a discussion of the choice of auxiliary distribution.

The following theorem gives expansions of $\ddot{\xi}_N - \xi_N$ and $\ddot{\tau}_{jN} - \tau_{jN}$, and hence, together with Theorem 5.2, conditions under which the formal Edgeworth expansions of the sample t-statistic and the bootstrap t-statistic agree up to $o_P(G^{-1})$.

Theorem 5.3. *Suppose Assumptions 4–6 are satisfied with* $\lambda = 4$, *that* $\mathrm{E}^*|v^*|^8 < \infty$, *and that* H_0 *is true. Then it holds that*

$$\ddot{\xi}_N - \xi_N = \xi_N(\mathrm{E}^*(v^{*4}) - 1) + o_P(1) \quad \textit{and} \quad \ddot{\tau}_{jN} - \tau_{jN} = o_P(1) \textit{ for } j = 1, 2, 3.$$

Theorem 5.3 establishes that $\ddot{\xi}_N - \xi_N \xrightarrow{P} 0$ only if the auxiliary random variable has fourth moment equal to one, i.e. only if $\mathrm{E}^*(v^{*4}) = 1$, which is satisfied only by the Rademacher distribution.

With the formal expansions in Theorems 5.1–5.3, and noting that Theorem 5.2 implies that $\ddot{\gamma}_N^2 - \gamma_N^2 = \gamma_N^2(\mathrm{E}^*(v^{*3}) - 1)(\mathrm{E}^*(v^{*3}) + 1) + O_P(G^{-1/2})$, the third-order bootstrap error in estimating $P(t_a \leq x)$ is

$$\begin{aligned} P^*(t_a^* \leq x) - P(t_a \leq x) = {} & \frac{1}{6} G^{-1/2} \Big(\gamma_N(\mathrm{E}^*(v^{*3}) - 1) + G^{-1/2} \mathrm{E}^*(v^{*3}) \omega_N(A) \mathcal{Z}_N \Big) (2x^2 + 1)\phi(x) \\ & - \frac{1}{18} G^{-1} \gamma_N^2 (\mathrm{E}^*(v^{*3}) - 1)(\mathrm{E}^*(v^{*3}) + 1)(x^5 + 2x^3 - 3x)\phi(x) \\ & + \frac{1}{12} G^{-1} \xi_N (\mathrm{E}^*(v^{*4}) - 1)(x^3 - 3x) + o_P(G^{-1}), \end{aligned} \tag{30}$$

uniformly in x. Similarly, the third-order bootstrap error for two-sided symmetric tests, i.e. the error in estimating $P(|t_a| \leq x)$, is given, uniformly in x, by

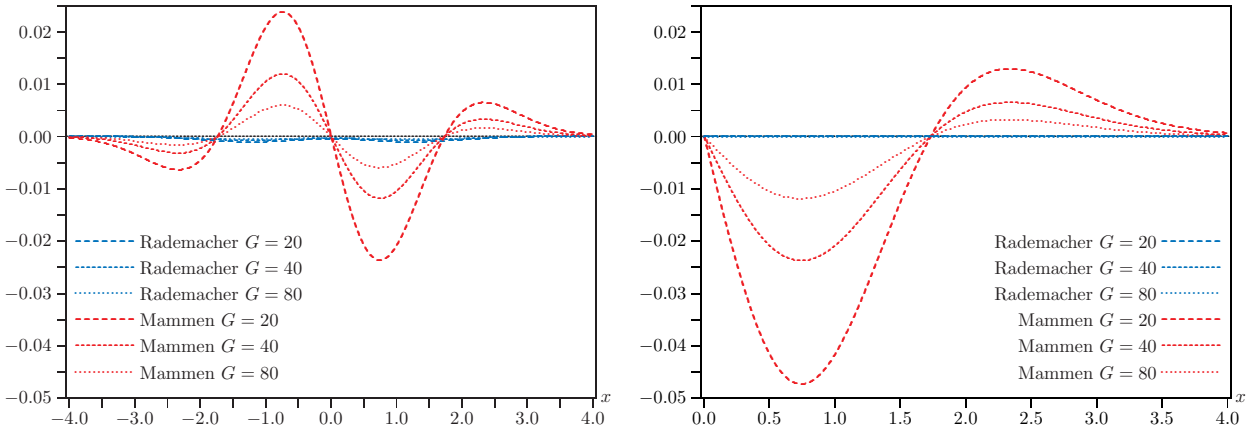
$$\begin{aligned} P^*(|t_a^*| \leq x) - P(|t_a| \leq x) = {} & -\frac{1}{9} G^{-1} \gamma_N^2 (\mathrm{E}^*(v^{*3}) - 1)(\mathrm{E}^*(v^{*3}) + 1)(x^5 + 2x^3 - 3x)\phi(x) \\ & + \frac{1}{6} G^{-1} \xi_N (\mathrm{E}^*(v^{*4}) - 1)(x^3 - 3x) + o_P(G^{-1}). \end{aligned} \tag{31}$$

Thus, in combination with the earlier results, Theorem 5.3 establishes conditions for a third-order asymptotic refinement of the WCB. We state these in two corollaries.

Corollary 5.2. *Under the conditions of Theorem 5.3, it holds that* $\ddot{q}_2(x) \xrightarrow{P} q_2(x)$ *uniformly in* x *if and only if* $\mathrm{E}(u_{ig}^3) = 0$ *for all* i, g *and* $\mathrm{E}^*(v^{*4}) = 1$. *In that case,*

$$\sup_{x\in\mathbb{R}} \left| P^*(|t_a^*| \leq x) - P(|t_a| \leq x) \right| = o_P(G^{-1}).$$

Figure 10: Bootstrap errors and aux. distributions, treatment dummy, $\gamma = 3$, $\rho = 0.10$, $G_1 = G/5$



a) One-sided Edgeworth bootstrap errors

b) Two-sided Edgeworth bootstrap errors

Corollary 5.2 establishes an asymptotic refinement of the two-sided symmetric test if only if the errors have third moment equal to zero (to ensure that $\ddot{\gamma}_N - \gamma_N \xrightarrow{P} 0$) and the auxiliary random variable, v^*, has fourth moment equal to one (to ensure that $\ddot{\xi}_N - \xi_N \xrightarrow{P} 0$). Since the latter condition implies that the third moment of v^* is zero, Corollary 5.2 thus shows that only the Rademacher auxiliary distribution has the potential to achieve an asymptotic refinement for the two-sided symmetric test.

Corollary 5.3. *Under the conditions of Theorem 5.3, it holds that* $G^{1/2}(\ddot{q}_1(x) - q_1(x)) \xrightarrow{P} 0$ *uniformly in* x *and* $\ddot{q}_2(x) \xrightarrow{P} q_2(x)$ *uniformly in* x *if and only if* $\mathrm{E}(u_{ig}^3) = 0$ *for all* i, g, $\mathrm{E}^*(v^{*3}) = 0$, *and* $\mathrm{E}^*(v^{*4}) = 1$. *In that case,*

$$\sup_{x \in \mathbb{R}} \left| P^*(t_a^* \leq x) - P(t_a \leq x) \right| = o_P(G^{-1}).$$

For the one-sided case in Corollary 5.3, we note that a third-order asymptotic refinement of the WCB in the one-sided case is achieved under the additional condition that v^* has third moment equal to zero. This condition is required to eliminate the term of order $G^{-1/2}$ in the expansion of $\ddot{\gamma}_N - \gamma_N$, and hence to make $\ddot{q}_1(x) - q_1(x)$ of order $o_P(G^{-1/2})$ uniformly in x. Thus, as in Corollary 5.2, the result in Corollary 5.3 shows that only the Rademacher auxiliary distribution has the potential to achieve a third-order asymptotic refinement in the one-sided case.

The above analysis, specifically the result in Corollary 5.1, shows theoretical conditions under which a v^* with $\mathrm{E}^*(v^{*3}) = 1$ should be preferred; see also Wu (1986), Liu (1988), and Mammen (1993). However, there is a good deal of simulation evidence that, for the ordinary wild bootstrap without clustering, using such a v^* often does not, in fact, work particularly well; see, e.g., Davidson, Monticini, and Peel (2007) and Davidson and Flachaire (2008). This evidence is also supported by our simulation results in Section 4, where we compare the Rademacher and Mammen distributions. Furthermore, the expansion in Theorem 5.2 and the results in Corollaries 5.2 and 5.3 suggest that a v^* with $\mathrm{E}^*(v^{*3}) = 0$ and $\mathrm{E}^*(v^{*4}) = 1$, i.e. the Rademacher distribution, may be preferred.

In Figure 10, we plot the bootstrap errors, i.e. the right-hand sides of (30) and (31), for two common choices of auxiliary distribution, namely, the Rademacher and Mammen (1993) distributions. We ignore the o_P-terms on the right-hand sides, and the random variable $\mathcal{Z}_N$ is set equal to its expectation, which is zero. As in Figure 9, the setup is the same as that in panel b) of Figure 3. In particular, therefore, the errors are skewed, suggesting that the Mammen distribution

may have an advantage in this case. As discussed in the previous paragraph, the Rademacher and Mammen distributions trade off the relative importance of the third and fourth moments, in the sense that the Mammen distribution satisfies $\mathrm{E}^*(v^{*3}) = 1$ and the Rademacher distribution satisfies $\mathrm{E}^*(v^{*4}) = 1$; c.f. Theorems 5.2 and 5.3.

For both the one-sided test in panel a) of Figure 10 and the two-sided test in panel b), the Rademacher auxiliary distribution has very much smaller bootstrap error than the Mammen distribution. This may be surprising, because the errors are skewed, which should favor the Mammen distribution. It appears that, even though the bootstrap error with the Mammen distribution vanishes at rate $O(G^{-1})$, while that with the Rademacher distribution vanishes at rate $O(G^{-1/2})$, the skewness correction is much less important than the kurtosis correction, which results in the superiority of the Rademacher auxiliary distribution in this case.

Comparing Figures 9 and 10, we note that, if the bootstrap errors are very small (or zero as with a refinement), then the bootstrap achieves the same rejection frequency as the Edgeworth CDFs in Figure 9. Thus, the very small bootstrap errors for the Rademacher distribution in panel b) of Figure 10 explain the superior finite-sample size of the (restricted) bootstrap tests based on the Rademacher distribution in panel b) of Figure 3, especially for $G \geq 30$. In contrast, the bootstrap errors with the Mammen distribution in panel b) of Figure 10 are negative and quite large for $x \leq 2$, meaning there is not enough mass in that part of the distribution, and hence too much mass in the right tail, leading to negative size distortion (underrejection) as found in panel b) of Figure 3.

6 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 key results 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 sets. First, Theorem 2.1 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. Second, Theorems 3.1 and 3.2 provide a similar foundation for the wild cluster bootstrap (WCB) and ordinary wild bootstrap (WB), respectively, in both their restricted and unrestricted versions. Third, Theorems 5.1–5.3 provide higher-order asymptotic theory that we use to shed light on the choice of auxiliary distribution in the WCB and to give conditions under which the WCB may attain a higher-order asymptotic refinement. Simulation evidence and higher-order theory suggest that the restricted WCB using the Rademacher auxiliary distribution is generally the best choice.

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 $\{w_g\}$ be an independent sequence of random variables with mean zero satisfying $\sup_{g\in\mathbb{N}} \mathrm{E}|w_g|^\theta < \infty$ for some $\theta \geq 1$. Then $\sum_{g=1}^G w_g = O_P(G^{\max\{1/\theta, 1/2\}})$.*

Proof. First suppose $1 \leq \theta \leq 2$. Let $\epsilon > 0$ be arbitrary and choose K such that $K^\theta = 2\epsilon^{-1} \sup_g \mathrm{E}|w_g|^\theta$. By Markov's inequality and the von Bahr-Esseen inequality,

$$P\left(\sum_{g=1}^{G} w_g > KG^{1/\theta}\right) \leq \frac{\mathrm{E}|\sum_{g=1}^{G} w_g|^\theta}{K^\theta G} \leq \frac{2\sum_{g=1}^{G} \mathrm{E}|w_g|^\theta}{K^\theta G} \leq \frac{2\sup_{g\in\mathbb{N}} \mathrm{E}|w_g|^\theta}{K^\theta} = \epsilon.$$

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

Lemma A.2. *Let Assumptions 1 and 2 be satisfied. Then,*

$$\sup_{g\in\mathbb{N}} N_g^{-\theta} \mathrm{E}\|\boldsymbol{X}_g^\top \boldsymbol{u}_g\|^\theta = O(1) \text{ for } 1 \leq \theta \leq 4 + \lambda,$$
$$\sup_{g\in\mathbb{N}} N_g^{-\theta} \mathrm{E}\|\boldsymbol{X}_g^\top \boldsymbol{X}_g\|^\theta = O(1) \text{ for } 1 \leq \theta \leq 2 + \lambda/2.$$

Proof. By the triangle and c_r inequalities, for $\theta \geq 1$,

$$\mathrm{E}\|\boldsymbol{X}_g^\top \boldsymbol{u}_g\|^\theta = \mathrm{E}\left\|\sum_{i=1}^{N_g} \boldsymbol{X}_{ig}^\top u_{ig}\right\|^\theta \leq \mathrm{E}\left(\sum_{i=1}^{N_g} \|\boldsymbol{X}_{ig}^\top u_{ig}\|\right)^\theta \leq N_g^{\theta-1} \sum_{i=1}^{N_g} \mathrm{E}\|\boldsymbol{X}_{ig}^\top u_{ig}\|^\theta. \tag{A.1}$$

By Assumption 1, $\sup_{i,g\in\mathbb{N}} \mathrm{E}\|\boldsymbol{X}_{ig}^\top u_{ig}\|^\theta \leq C$ when $\theta \leq 4 + \lambda$, in which case (A.1) implies that $\mathrm{E}\|\boldsymbol{X}_g^\top \boldsymbol{u}_g\|^\theta \leq CN_g^\theta$. It follows that $\sup_{g\in\mathbb{N}} N_g^{-\theta}\mathrm{E}\|\boldsymbol{X}_g^\top \boldsymbol{u}_g\|^\theta \leq C$ for $\theta \leq 4 + \lambda$, which proves the first result. The second result follows in the same way after replacing $\boldsymbol{u}_g$ by $\boldsymbol{X}_g$ in (A.1), noting that $\|\boldsymbol{X}_{ig}^\top \boldsymbol{X}_{ig}\|^\theta \leq \|\boldsymbol{X}_{ig}\|^{2\theta}$, and applying the uniform moment condition in Assumption 2. □

Lemma A.3. *Let W_g and Z_g be given by (27) and (28), and also let*

$$Z_g(A) = \frac{1}{G}\sum_{h=1}^{G} V_{gh}(A) \text{ with } V_{gh}(A) = (\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-1/2} \boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} \frac{\sqrt{G}}{N} \boldsymbol{X}_g^\top \boldsymbol{X}_g \boldsymbol{M}_N(A) \boldsymbol{Q}_N^{-1} \frac{G}{N} \boldsymbol{X}_h^\top \boldsymbol{u}_h. \tag{A.2}$$

If Assumptions 4–6 are satisfied then,

$$\sup_{g\in\mathbb{N}} \mathrm{E}|W_g|^\theta = O(1), \quad \sup_{g\in\mathbb{N}} \mathrm{E}|Z_g(A)|^\theta = O(G^{-\theta/2}), \quad \sup_{g,h\in\mathbb{N}} \mathrm{E}|V_{gh}(A)|^\theta = O(1),$$

for $1 \leq \theta \leq 4 + \lambda$ and $A \in \{U, R\}$.

Proof. We first note that, under Assumptions 4–6 and using (22) and Lemma A.2,

$$\sup_{g\in\mathbb{N}} \mathrm{E}|W_g|^\theta \leq (\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-\theta/2} \|\boldsymbol{Q}_N^{-1}\|^\theta \frac{G^{\theta/2}}{N^\theta} \sup_{g\in\mathbb{N}} \mathrm{E}\|\boldsymbol{X}_g^\top \boldsymbol{u}_g\|^\theta = O(1).$$

Second, $\boldsymbol{M}_N(A)$ has the useful properties that $\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = \boldsymbol{M}_N(A)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ and $\ddot{\boldsymbol{u}}_g = \boldsymbol{u}_g - \boldsymbol{X}_g \boldsymbol{M}_N(A)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, so that $Z_g(A) = (\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-1/2} \boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} \frac{\sqrt{G}}{N} \boldsymbol{X}_g^\top \boldsymbol{X}_g (\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$; see (29). Then,

$$\sup_{g\in\mathbb{N}} \mathrm{E}|Z_g(A)|^\theta \leq (\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-\theta/2} \|\boldsymbol{Q}_N^{-1}\|^{2\theta} \frac{G^{\theta/2}}{N^\theta} \sup_{g\in\mathbb{N}} \|\boldsymbol{X}_g^\top \boldsymbol{X}_g\|^\theta \mathrm{E}\|\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^\theta = O(G^{-\theta/2}),$$

using again Assumptions 4–6, (22), and Lemma A.2. Finally, by the same arguments,

$$\sup_{g,h\in\mathbb{N}} \mathrm{E}|V_{gh}(A)|^\theta \leq (\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-\theta/2} \|\boldsymbol{Q}_N^{-1}\|^{2\theta} \frac{G^{3\theta/2}}{N^{2\theta}} \sup_{g\in\mathbb{N}} \|\boldsymbol{X}_g^\top \boldsymbol{X}_g\| \|\boldsymbol{M}(A)\|^\theta \sup_{h\in\mathbb{N}} \mathrm{E}\|\boldsymbol{X}_h^\top \boldsymbol{u}_h\|^\theta = O(1).$$

□

Lemma A.4. *Let W_g and Z_g be given by (27) and (28), and further define $S_N = G^{-1/2}\sum_{g=1}^G W_g$, $U_N = G^{-1/2}\sum_{g=1}^G (W_g^2 - 1)$, and $T_N = \sum_{g=1}^G Z_g(Z_g - 2W_g)$.*

(i) *If Assumptions 4–6 are satisfied then*

$$\begin{aligned}
&\mathrm{E}(S_N) = 0, \quad \mathrm{E}(S_N^2) = 1, \quad \mathrm{E}(S_N^3) = G^{-1/2}\gamma_N, \quad \mathrm{E}(S_N^4) = 3 + G^{-1}(\xi_N - 3\xi_{2N}),\\
&\mathrm{E}(S_N U_N) = \gamma_N, \quad \mathrm{E}(S_N^2 U_N) = G^{-1/2}(\xi_N - \xi_{2N}),\\
&\mathrm{E}(S_N T_N) = O(G^{-1/2}), \quad \mathrm{E}(S_N^2 T_N) = \tau_{1N} - 2\tau_{2N} + O(G^{-1}),
\end{aligned}$$

where γ_N, ξ_N, ξ_{2N}, τ_{1N}, and τ_{2N} are defined in Theorem 5.1.

(ii) *If, in addition, Assumption 6 is satisfied with $\lambda = 2$, then it also holds that*

$$\begin{aligned}
&\mathrm{E}(S_N^3 U_N) = 3\gamma_N + O(G^{-1}), \quad \mathrm{E}(S_N^4 U_N) = G^{-1/2}(4\gamma_N^2 + 6(\xi_N - \xi_{2N})) + O(G^{-3/2}),\\
&\mathrm{E}(S_N U_N^2) = O(G^{-1/2}), \quad \mathrm{E}(S_N^2 U_N^2) = 2\gamma_N^2 + \xi_N - \xi_{2N} + O(G^{-1}),\\
&\mathrm{E}(S_N^3 T_N) = O(G^{-1/2}), \quad \mathrm{E}(S_N^4 T_N) = 6\tau_{1N} - 6\tau_{2N} + 3\tau_{3N} + O(G^{-1}),
\end{aligned}$$

where τ_{3N} is defined in Theorem 5.1.

(iii) *If, in addition, Assumption 6 is satisfied with $\lambda = 4$, then it also holds that*

$$\mathrm{E}(S_N^3 U_N^2) = O(G^{-1/2}), \quad \mathrm{E}(S_N^4 U_N^2) = 12\gamma_N^2 + 3(\xi_N - \xi_{2N}) + O(G^{-1}).$$

Proof. Part (i): Clearly, because W_g is mean zero and independent of W_h for $h \neq g$, it easily follows that $\mathrm{E}(S_N) = 0$, $\mathrm{E}(S_N^2) = G^{-1}\sum_{g=1}^G \mathrm{E}(W_g^2) = 1$, and $\mathrm{E}(S_N^3) = G^{-3/2}\sum_{g=1}^G \mathrm{E}(W_g^3) = G^{-1/2}\gamma_N$. For the fourth moment we find

$$\mathrm{E}(S_N^4) = G^{-2}\mathrm{E}\Big(\sum_{g_1,g_2,g_3,g_4=1}^G W_{g_1}W_{g_2}W_{g_3}W_{g_4}\Big),$$

where, because $\mathrm{E}(W_g) = 0$, none of the summation indexes $g_1, \ldots, g_4$ can be different from all the remaining indexes, i.e. the indexes must either all be equal or be equal in pairs. It follows that

$$\begin{aligned}
\mathrm{E}(S_N^4) &= G^{-2}\sum_{g=1}^G \mathrm{E}(W_g^4) + 3G^{-2}\sum_{g_1,g_2=1,g_1\neq g_2}^G (\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2}^2)\\
&= G^{-1}\xi_N + 3G^{-2}\sum_{g_1,g_2=1}^G (\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2}^2) - 3G^{-2}\sum_{g=1}^G (\mathrm{E}W_g^2)^2 = G^{-1}\xi_N + 3 - 3G^{-1}\xi_{2N}.
\end{aligned}$$

Next, for the cross-moments, we similarly find that $\mathrm{E}(S_N U_N) = G^{-1}\sum_{g=1}^G \mathrm{E}(W_g^3) = \gamma_N$ and

$$\mathrm{E}(S_N^2 U_N) = G^{-3/2}\mathrm{E}\Big(\sum_{g_1,g_2,g_3=1}^G W_{g_1}W_{g_2}(W_{g_3}^2 - 1)\Big),$$

where we note that the summation indexes must satisfy $g_1 = g_2$. Consequently,

$$\begin{aligned}
\mathrm{E}(S_N^2 U_N) &= G^{-3/2}\sum_{g=1}^G \mathrm{E}(W_g^2(W_g^2 - 1)) + G^{-3/2}\sum_{g_1,g_2=1,g_1\neq g_2}^G (\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2}^2 - 1)\\
&= G^{-3/2}\sum_{g=1}^G \mathrm{E}(W_g^2(W_g^2 - 1)) + G^{-3/2}\sum_{g_1,g_2=1}^G (\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2}^2 - 1) - G^{-3/2}\sum_{g=1}^G (\mathrm{E}W_g^2)(\mathrm{E}W_g^2 - 1))\\
&= G^{-3/2}\sum_{g=1}^G (\mathrm{E}(W_g^4) - \mathrm{E}(W_g^2)) - G^{-3/2}\sum_{g=1}^G ((\mathrm{E}W_g^2)^2 - \mathrm{E}(W_g^2)) = G^{-1/2}(\xi_N - \xi_{2N}).
\end{aligned}$$

The result for $\mathrm{E}(S_N T_N)$ follows directly from (A.2) and Lemma A.3. For $\mathrm{E}(S_N^2 T_N)$ we find that

$$\mathrm{E}(S_N^2 T_N) = G^{-1} \sum_{g_1,g_2,g_3=1}^{G} \mathrm{E}(W_{g_1} W_{g_2} Z_{g_3}^2) - 2G^{-1} \sum_{g_1,g_2,g_3=1}^{G} \mathrm{E}(W_{g_1} W_{g_2} W_{g_3} Z_{g_3}),$$

where the first term is part of τ_{1N}. For the second term, we note from (A.2) that the summation is non-zero if either $g_1 = g_3$ (or identically $g_2 = g_3$) or if $g_1 = g_2$. In the first case, the contribution is $-4G^{-1} \sum_{g_1,g_2}^{G} \mathrm{E}(W_{g_1}^2 W_{g_2} Z_{g_1})$, which is the remaining part of τ_{1N}. In the second case, the contribution is

$$-2G^{-1} \sum_{g_1,g_2=1,g_1\neq g_2}^{G} (\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2} Z_{g_2}) - 2G^{-1} \sum_{g=1}^{G} \mathrm{E}(W_g^3 Z_g) = -2G^{-1} \sum_{g_1,g_2=1}^{G} (\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2} Z_{g_2}) + O(G^{-1}),$$

which equals $-2\tau_{2N} + O(G^{-1})$ and where the $O(G^{-1})$ term is due to Lemma A.3.

Part (ii): Because we now assume $\lambda = 2$, six moments of u_{ig}, and hence of W_g and Z_g, exist, which implies that the required cross-moments of S_N, U_N, and T_N exist. Thus, similarly to the previous moments, we find

$$\mathrm{E}(S_N^3 U_N) = G^{-2} \mathrm{E}\Big(\sum_{g_1,g_2,g_3,g_4=1}^{G} W_{g_1} W_{g_2} W_{g_3} (W_{g_4}^2 - 1) \Big),$$

where none of the summation indexes $g_1, \ldots, g_3$ can be different from all the remaining indexes. It follows that

$$\begin{aligned}
\mathrm{E}(S_N^3 U_N) &= G^{-2} \sum_{g=1}^{G} (\mathrm{E}(W_g^5) - \mathrm{E}(W_g^3)) + G^{-2} \sum_{g_1,g_2=1,g_1\neq g_2}^{G} (\mathrm{E}W_{g_1}^3)(\mathrm{E}W_{g_2}^2 - 1) + 3G^{-2} \sum_{g_1,g_2=1,g_1\neq g_2}^{G} (\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2}^3) \\
&= O(G^{-1}) + 3G^{-2} \sum_{g_1,g_2=1}^{G} \mathrm{E}(W_{g_1}^2)\mathrm{E}(W_{g_2}^3) - 3G^{-2} \sum_{g=1}^{G} (\mathrm{E}W_g^2)(\mathrm{E}W_g^3) = 3\gamma_N + O(G^{-1}),
\end{aligned}$$

where the $O(G^{-1})$ terms are due to Lemma A.3. Next, we find in the same way that $\mathrm{E}(S_N^4 U_N)$ contains five summation indexes, out of which the four associated with an S_N cannot be different from all the remaining indexes, i.e. those indexes must either be all equal, equal in pairs, or equal in one triplet. In the first case, Lemma A.3 easily shows that the contribution is $O(G^{-3/2})$. Thus,

$$\begin{aligned}
\mathrm{E}(S_N^4 U_N) &= G^{-5/2} \mathrm{E}\Big(\sum_{g_1,g_2,g_3,g_4,g_5=1}^{G} W_{g_1} W_{g_2} W_{g_3} W_{g_4} (W_{g_5}^2 - 1) \Big) \\
&= 4G^{-5/2} \sum_{g_1,g_2=1,g_1\neq g_2}^{G} \mathrm{E}(W_{g_1}^3 (W_{g_2}^3 - W_{g_2})) + 6G^{-5/2} \sum_{g_1,g_2=1,g_1\neq g_2}^{G} \mathrm{E}(W_{g_1}^2 W_{g_2}^2 (W_{g_2}^2 - 1)) \\
&\quad + 3G^{-5/2} \sum_{\substack{g_1,g_2,g_3=1 \\ g_1\neq g_2\neq g_3}}^{G} \mathrm{E}(W_{g_1}^2 W_{g_2}^2 (W_{g_3}^2 - 1)) + O(G^{-3/2}).
\end{aligned}$$

When a term has only one summation, then it is $O(G^{-3/2})$ by Lemma A.3 because of the normalization by $G^{-5/2}$. It follows that the first and second terms of $\mathrm{E}(S_N^4 U_N)$ are

$$4G^{-5/2} \sum_{g_1,g_2=1}^{G} (\mathrm{E}W_{g_1}^3)(\mathrm{E}(W_{g_2}^3) - \mathrm{E}(W_{g_2})) + O(G^{-3/2}) = 4G^{-1/2}\gamma_N^2 + O(G^{-3/2}),$$

$$6G^{-5/2} \sum_{g_1,g_2=1}^{G} (\mathrm{E}W_{g_1}^2)(\mathrm{E}(W_{g_2}^4) - \mathrm{E}(W_{g_2}^2)) + O(G^{-3/2}) = 6G^{-1/2}(\xi_N - 1) + O(G^{-3/2}).$$

The third term of $\mathrm{E}(S_N^4 U_N)$ is

$$3G^{-5/2}\sum_{g_1,g_2,g_3=1}^{G}(\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2}^2)(\mathrm{E}W_{g_3}^2-1)-3G^{-5/2}\sum_{g=1}^{G}(\mathrm{E}W_g^2)^2(\mathrm{E}W_g^2-1)$$
$$-3G^{-5/2}\sum_{g_1,g_2=1,g_1\neq g_2}^{G}(\mathrm{E}W_{g_1}^2)^2(\mathrm{E}W_{g_2}^2-1)-6G^{-5/2}\sum_{g_1,g_2=1,g_1\neq g_2}^{G}(\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2}^2)(\mathrm{E}W_{g_2}^2-1),$$

of which the first term is zero, the next two are $O(G^{-3/2})$ by Lemma A.3, and the final term is

$$-6G^{-5/2}\sum_{g_1,g_2=1}^{G}(\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2}^2)(\mathrm{E}W_{g_2}^2-1)+O(G^{-3/2})=-6G^{-1/2}(\xi_{2N}-1)+O(G^{-3/2})$$

by the same arguments as above.

Next, $\mathrm{E}(S_N U_N^2)$ contains three summation indexes, of which the index associated with S_N cannot be different from the other two, and the result follows immediately from Lemma A.3. Finally, $\mathrm{E}(S_N^2 U_N^2)$ contains four summation indexes, where the two indexes associated with an S_N cannot be different from all the other indexes. Hence,

$$\begin{aligned}\mathrm{E}(S_N^2U_N^2)&=G^{-2}\mathrm{E}\Big(\sum_{g_1,g_2,g_3,g_4=1}^{G}W_{g_1}W_{g_2}(W_{g_3}^2-1)(W_{g_4}^2-1)\Big)\\&=G^{-2}\sum_{g_1,g_2,g_3=1}^{G}\mathrm{E}(W_{g_1}^2(W_{g_2}^2-1)(W_{g_3}^2-1))+2G^{-2}\sum_{g_1,g_2=1,g_1\neq g_2}^{G}\mathrm{E}(W_{g_1}W_{g_2}(W_{g_1}^2-1)(W_{g_2}^2-1)).\end{aligned}$$

Again using Lemma A.3 and noting the normalization by G^{-2}, the second term on the right-hand side is $2G^{-2}\sum_{g_1,g_2=1}^{G}(\mathrm{E}W_{g_1}^3)(\mathrm{E}W_{g_2}^3)+O(G^{-1})=2\gamma_N^2+O(G^{-1})$. The first term of $\mathrm{E}(S_N^2U_N^2)$ has either $g_2=g_3$, in which case the contribution is

$$G^{-2}\sum_{g_1,g_2=1,g_1\neq g_2}^{G}(\mathrm{E}W_{g_1}^2)(\mathrm{E}(W_{g_2}^4)-2\mathrm{E}(W_{g_2}^2)+1)+G^{-2}\sum_{g=1}^{G}\mathrm{E}(W_g^2(W_g^2-1)^2)=\xi_N-1+O(G^{-1}),$$

or it has $g_2\neq g_3$, in which case the contribution is

$$G^{-2}\sum_{\substack{g_1,g_2,g_3=1\\g_1\neq g_2\neq g_3}}^{G}(\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2}^2-1)(\mathrm{E}W_{g_3}^2-1)+2G^{-2}\sum_{g_1,g_2=1,g_1\neq g_2}^{G}(\mathrm{E}W_{g_1}^4-\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2}^2-1),$$

where the second term is $O(G^{-1})$ by Lemma A.3 and the first term is

$$\begin{aligned}&G^{-2}\sum_{g_1,g_2,g_3=1}^{G}(\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2}^2-1)(\mathrm{E}W_{g_3}^2-1)-G^{-2}\sum_{g=1}^{G}(\mathrm{E}W_g^2)(\mathrm{E}W_g^2-1)(\mathrm{E}W_g^2-1)\\&-G^{-2}\sum_{g_1,g_2=1,g_1\neq g_2}^{G}(\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2}^2-1)^2-2G^{-2}\sum_{g_1,g_2=1,g_1\neq g_2}^{G}(\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_1}^2-1)(\mathrm{E}W_{g_2}^2-1)\\=\;&-G^{-2}\sum_{g_1,g_2=1}^{G}(\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2}^2-1)^2+O(G^{-1})=-G^{-1}\sum_{g=1}^{G}(\mathrm{E}W_{g_2}^2-1)^2+O(G^{-1}),\end{aligned}$$

which equals $1-\xi_{2N}+O(G^{-1})$.

For the cross-moments with T_N, we use (A.2), let $V_{gh} = V_{gh}(U)$, and find that

$$\mathrm{E}(S_N^3 T_N) = G^{-7/2} \sum_{g_1,\ldots,g_6}^{G} \mathrm{E}(W_{g_1} W_{g_2} W_{g_3} V_{g_6 g_4} V_{g_6 g_5}) - 2G^{-5/2} \sum_{g_1,\ldots,g_5}^{G} \mathrm{E}(W_{g_1} W_{g_2} W_{g_3} W_{g_4} V_{g_4 g_5}),$$

where the subscripts $g_1, \ldots, g_5$ on both sides must be equal at least in pairs. This eliminates three summations in both terms, and the result then follows easily from Lemma A.3. Next, $\mathrm{E}(S_N^4 T_N)$ also contains two terms, which we investigate in turn. By (A.2) the first term is given by $G^{-4} \sum_{g_1,\ldots,g_7=1}^{G} \mathrm{E}(W_{g_1} \cdots W_{g_4} V_{g_7 g_5} V_{g_7 g_6})$, where the subscripts $g_1, \ldots, g_6$ must be equal in three pairs, two triplets, or one pair and one quadruplet. In the latter two cases there are at most three summations, so the contribution is $O(G^{-1})$ by Lemma A.3. This leaves the contribution

$$\begin{aligned}
&3G^{-4} \sum_{\substack{g_1,\ldots,g_4=1 \\ g_1 \neq g_2 \neq g_3}}^{G} (\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2}^2)(\mathrm{E}V_{g_4 g_3}^2) + 6G^{-4} \sum_{\substack{g_1,\ldots,g_4=1 \\ g_1 \neq g_2 \neq g_3}}^{G} (\mathrm{E}W_{g_1}^2)\mathrm{E}(W_{g_2} W_{g_3} V_{g_4 g_2} V_{g_4 g_3}) \\
&= 3G^{-4} \sum_{g_1,\ldots,g_4=1}^{G} (\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2}^2)(\mathrm{E}V_{g_4 g_3}^2) + 6G^{-4} \sum_{g_1,\ldots,g_4=1}^{G} (\mathrm{E}W_{g_1}^2)\mathrm{E}(W_{g_2} W_{g_3} V_{g_4 g_2} V_{g_4 g_3}) + O(G^{-1}) \\
&= 3G^{-2} \sum_{g_1,g_2=1}^{G} (\mathrm{E}V_{g_1 g_2}^2) + 6G^{-3} \sum_{g_1,g_2,g_3=1}^{G} \mathrm{E}(W_{g_1} W_{g_2} V_{g_3 g_1} V_{g_3 g_2}) + O(G^{-1}) \\
&= 3 \sum_{g=1}^{G} \mathrm{E}(Z_g^2) + 6G^{-1} \sum_{g_1,g_2,g_3=1}^{G} \mathrm{E}(W_{g_1} W_{g_2} Z_{g_3}^2) + O(G^{-1}),
\end{aligned}$$

using Lemma A.3. The first term on the right-hand side is $3\tau_{3N}$, and the second is six times the first part of τ_{1N}. Similarly, the second term of $\mathrm{E}(S_N^4 T_N)$ is $-2G^{-3} \sum_{g_1,\ldots,g_6=1}^{G} \mathrm{E}(W_{g_1} \cdots W_{g_5} V_{g_5 g_6})$, where the subscripts $g_1, \ldots, g_6$ must be equal in three pairs since otherwise the contribution is $O(G^{-1})$ by Lemma A.3. This leaves

$$\begin{aligned}
&- 6G^{-3} \sum_{\substack{g_1,g_2,g_3=1 \\ g_1 \neq g_2 \neq g_3}}^{G} (\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2}^2)(\mathrm{E}W_{g_3} V_{g_3 g_3}) - 24G^{-3} \sum_{\substack{g_1,g_2,g_3=1 \\ g_1 \neq g_2 \neq g_3}}^{G} (\mathrm{E}W_{g_1}^2)\mathrm{E}(W_{g_2}^2 W_{g_3} V_{g_2 g_3}) \\
&= - 6G^{-3} \sum_{g_1,g_2,g_3=1}^{G} (\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2}^2)(\mathrm{E}W_{g_3} V_{g_3 g_3}) - 24G^{-3} \sum_{g_1,g_2,g_3=1}^{G} (\mathrm{E}W_{g_1}^2)\mathrm{E}(W_{g_2}^2 W_{g_3} V_{g_2 g_3}) + O(G^{-1}) \\
&= - 6 \sum_{g=1}^{G} \mathrm{E}(W_g Z_g) - 24G^{-1} \sum_{g_1,g_2=1}^{G} \mathrm{E}(W_{g_1}^2 W_{g_2} Z_{g_1}) + O(G^{-1}),
\end{aligned}$$

using Lemma A.3. The first term on the right-hand side is $-6\tau_{2N}$ and the second is six times the second part of τ_{1N}.

Part (iii): Since $\lambda = 4$, eight moments of u_{ig}, and hence of W_g, must exist, which implies that the required cross-moments of S_N and U_N exist. First, $\mathrm{E}(S_N^3 U_N^2)$ contains five summation indexes, where again those associated with an S_N cannot be different from all other indexes, and if an index from a U_N is different from all the other indexes, then it follows by Lemma A.3 that the contribution is $O(G^{-1/2})$. This leaves at most two summations, so the result is $O(G^{-1/2})$ by Lemma A.3.

Next, we find that $\mathrm{E}(S_N^4 U_N^2)$ contains six summation indexes, of which the four indexes associated with an S_N cannot be different from all the other indexes. Furthermore, terms with only one

or two summations are $O(G^{-1})$ by Lemma A.3 because of the normalization by G^{-3}. Thus,

$$\begin{aligned}
\mathrm{E}(S_N^4 U_N^2) &= G^{-3}\mathrm{E}\Bigg(\sum_{g_1,g_2,g_3,g_4,g_5,g_6=1}^{G} W_{g_1}W_{g_2}W_{g_3}W_{g_4}(W_{g_5}^2-1)(W_{g_6}^2-1)\Bigg) \\
&= 12G^{-3}\sum_{\substack{g_1,g_2,g_3=1 \\ g_1\neq g_2\neq g_3}}^{G}(\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2}^3)(\mathrm{E}W_{g_3}^3) + 3G^{-3}\sum_{\substack{g_1,g_2,g_3=1 \\ g_1\neq g_2\neq g_3}}^{G}(\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2}^2)(\mathrm{E}(W_{g_3}^2-1)^2) \\
&\quad + 8G^{-3}\sum_{\substack{g_1,g_2,g_3=1 \\ g_1\neq g_2\neq g_3}}^{G}(\mathrm{E}W_{g_1}^3)(\mathrm{E}W_{g_2}^3)(\mathrm{E}W_{g_3}^2-1) + 3G^{-3}\sum_{\substack{g_1,g_2,g_3,g_4=1 \\ g_1\neq g_2\neq g_3\neq g_4}}^{G}(\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2}^2)(\mathrm{E}W_{g_3}^2-1)(\mathrm{E}W_{g_4}^2-1).
\end{aligned}$$

The first three terms of $\mathrm{E}(S_N^4 U_N^2)$ are

$$\begin{aligned}
12G^{-3}\sum_{g_1,g_2,g_3=1}^{G}(\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2}^3)(\mathrm{E}W_{g_3}^3) + O(G^{-1}) &= 12\gamma_N^2 + O(G^{-1}), \\
3G^{-3}\sum_{g_1,g_2,g_3=1}^{G}(\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2}^2)(\mathrm{E}(W_{g_3}^2-1)^2) + O(G^{-1}) &= 3(\xi_N-1) + O(G^{-1}), \\
8G^{-3}\sum_{g_1,g_2,g_3=1}^{G}(\mathrm{E}W_{g_1}^3)(\mathrm{E}W_{g_2}^3)(\mathrm{E}W_{g_3}^2-1) + O(G^{-1}) &= O(G^{-1}).
\end{aligned}$$

For the final term of $\mathrm{E}(S_N^4 U_N^2)$, we first note that if the summation index g_4 is unrestricted then the contribution is zero. Thus, the final term of $\mathrm{E}(S_N^4 U_N^2)$ is

$$-3G^{-3}\sum_{\substack{g_1,g_2,g_3=1 \\ g_1\neq g_2\neq g_3}}^{G}(\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2}^2)(\mathrm{E}W_{g_3}^2-1)^2 - 6G^{-3}\sum_{\substack{g_1,g_2,g_3=1 \\ g_1\neq g_2\neq g_3}}^{G}(\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2}^2)(\mathrm{E}W_{g_3}^2-1)(\mathrm{E}W_{g_1}^2-1),$$

where the second term is $O(G^{-1})$ because when g_3 is unrestricted the contribution is zero and when g_3 is restricted there are only two summations remaining. This leaves the contribution

$$-3G^{-3}\sum_{\substack{g_1,g_2,g_3=1 \\ g_1\neq g_2\neq g_3}}^{G}(\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2}^2)(\mathrm{E}W_{g_3}^2-1)^2 = -3G^{-3}\sum_{g_1,g_2,g_3=1}^{G}(\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2}^2)(\mathrm{E}W_{g_3}^2-1)^2 + O(G^{-1}),$$

which equals $-3(\xi_{2N}-1)+O(G^{-1})$. □

Appendix B: Proofs of Main Results

B.1 Proof of Theorem 2.1

Proof of (15). The left-hand side of (15) is

$$(\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-1/2}\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} N^{-1}\sum_{g=1}^{G}\boldsymbol{X}_g^\top \boldsymbol{u}_g = v_a^{-1/2}\mu_N^{1/2}\boldsymbol{a}^\top \boldsymbol{Q}^{-1} N^{-1}\sum_{g=1}^{G}\boldsymbol{X}_g^\top \boldsymbol{u}_g(1+o_P(1))$$

by Assumption 2 and Slutsky's Theorem. Thus, we need to prove that

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

We define $z_g = v_a^{-1/2}\mu_N^{1/2}N^{-1}\boldsymbol{a}^\top\boldsymbol{Q}^{-1}\boldsymbol{X}_g^\top\boldsymbol{u}_g$, which, by Assumption 1 is an independent sequence with mean zero and conditional variance given by $\mathrm{E}(z_g^2|\boldsymbol{X}) = v_a^{-1}\mu_N N^{-2}\boldsymbol{a}^\top\boldsymbol{Q}^{-1}\boldsymbol{X}_g^\top\boldsymbol{\Omega}_g\boldsymbol{X}_g\boldsymbol{Q}^{-1}\boldsymbol{a}$. By Assumption 2, $\sum_{g=1}^G \mathrm{E}(z_g^2|\boldsymbol{X}) \xrightarrow{P} 1$, and because $\{\mathrm{E}(z_g^2|\boldsymbol{X})\}$ is uniformly integrable by the uniform moment bound in Assumption 2, it follows from Vitali's Convergence Theorem (or Lebesgue's Dominated Convergence Theorem) that also $\sum_{g=1}^G \mathrm{E}(z_g^2) \to 1$. Then (B.1) follows from the Lyapunov Central Limit Theorem for heterogeneous, independent random variables if, for some $\xi > 0$, it holds that $\sum_{g=1}^G \mathrm{E}|z_g|^{2+\xi} \to 0$ (Lyapunov's condition). We find that

$$\begin{aligned}\sum_{g=1}^G \mathrm{E}|z_g|^{2+\xi} &\leq v_a^{-1-\xi/2}\mu_N^{1+\xi/2}\|\boldsymbol{a}^\top\boldsymbol{Q}^{-1}\|^{2+\xi}N^{-2-\xi}\sum_{g=1}^G \mathrm{E}\|\boldsymbol{X}_g^\top\boldsymbol{u}_g\|^{2+\xi} \\ &\leq C\mu_N^{1+\xi/2}N^{-2-\xi}\sum_{g=1}^G N_g^{2+\xi} \leq C\mu_N^{1+\xi/2}N^{-1-\xi}\sup_{g\in\mathbb{N}} N_g^{1+\xi} \to 0, \qquad \text{(B.2)}\end{aligned}$$

where the second inequality is due to positive definiteness of $\boldsymbol{Q}$ (Assumption 2) and Lemma A.2 (with $\theta = \xi + 2$), and the convergence is due to Assumption 3 setting $\xi = 2 + \lambda$.

Proof of (16). We start with the decomposition

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

where we used Assumption 2 and

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

Thus, we need to show that $\mu_N\boldsymbol{a}^\top\boldsymbol{A}_{mN}\boldsymbol{a} \xrightarrow{P} 0$ for $m = 1, 2, 3$. To prove the result for $m = 1$, let $w_g = z_g^2 - \mathrm{E}(z_g^2)$ such that, by the law of iterated expectations, $\sum_{g=1}^G w_g = v_a^{-1}\mu_N\boldsymbol{a}^\top\boldsymbol{A}_{1N}\boldsymbol{a}$. Clearly $\mathrm{E}(\sum_{g=1}^G w_g) = 0$, and we prove convergence in mean-square,

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

where the first equality follows from independence across clusters. The Lyapunov condition (B.2) with $\xi = 2$ shows that $\sum_{g=1}^G \mathrm{E}(z_g^4) \to 0$, and hence also $\sum_{g=1}^G (\mathrm{E}z_g^2)^2 \to 0$ by Jensen's inequality, which proves the result for $m = 1$.

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

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

We first note that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N\| = O_P(\|\boldsymbol{V}_N\|^{1/2}) = O_P(N^{-1/2} \sup_{g\in\mathbb{N}} N_g^{1/2})$; see (9). Then,

$$\begin{aligned} \mathrm{E}\Big\| \sum_{g=1}^{G} \boldsymbol{X}_g^\top \boldsymbol{X}_g \boldsymbol{Q}^{-1} \boldsymbol{a}\boldsymbol{a}^\top \boldsymbol{Q}^{-1} \boldsymbol{X}_g^\top \boldsymbol{u}_g \Big\| &\leq \|\boldsymbol{Q}^{-1}\|^2 \sum_{g=1}^{G} \mathrm{E}\Big\| \boldsymbol{X}_g^\top \boldsymbol{X}_g \boldsymbol{X}_g^\top \boldsymbol{u}_g \Big\| \\ &\leq \|\boldsymbol{Q}^{-1}\|^2 \sum_{g=1}^{G} \sum_{i,j=1}^{N_g} \mathrm{E}\Big\| \boldsymbol{X}_{ig}^\top \boldsymbol{X}_{ig} \boldsymbol{X}_{jg}^\top u_{jg} \Big\|, \end{aligned} \tag{B.3}$$

where, by the Cauchy-Schwarz inequality and Assumptions 1 and 2,

$$\mathrm{E}\|\boldsymbol{X}_{ig}^\top \boldsymbol{X}_{ig} \boldsymbol{X}_{jg}^\top u_{jg}\| \leq (\mathrm{E}\|\boldsymbol{X}_{ig}^\top \boldsymbol{X}_{ig}\|^2)^{1/2} (\mathrm{E}\|\boldsymbol{X}_{jg}^\top u_{jg}\|^2)^{1/2} \leq C,$$

so that the left-hand side of (B.3) is $O_P(N \sup_{g\in\mathbb{N}} N_g)$. It follows that

$$\|\mu_N \boldsymbol{a}^\top \boldsymbol{A}_{2N} \boldsymbol{a}\| = O_P\Big(\mu_N N^{-3/2} \sup_{g\in\mathbb{N}} N_g^{3/2}\Big) = o_P(1)$$

under Assumption 3; see also (11).

Finally, the proof for $m = 3$ is similar to that for $m = 2$, but simpler. We find the bound

$$\|\mu_N \boldsymbol{a}^\top \boldsymbol{A}_{3N} \boldsymbol{a}\| \leq \mu_N \frac{1}{N^2} \|\boldsymbol{Q}^{-1}\|^2 \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N\|^2 \sum_{g=1}^{G} \|\boldsymbol{X}_g^\top \boldsymbol{X}_g\|^2,$$

where $\sum_{g=1}^{G} \|\boldsymbol{X}_g^\top \boldsymbol{X}_g\|^2 = O_P(\sum_{g=1}^{G} N_g^2) = O_P(N \sup_{g\in\mathbb{N}} N_g)$ by Lemma A.2. It follows that

$$\|\mu_N \boldsymbol{a}^\top \boldsymbol{A}_{3N} \boldsymbol{a}\| = O_P\Big(\mu_N N^{-2} \sup_{g\in\mathbb{N}} N_g^2\Big) = o_P(1).$$

Proof of (17). We use (14) to decompose the t-statistic (6) as

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

and the result then follows directly from (15), (16), and Slutsky's Theorem.

B.2 Proof of Theorem 3.1

We first give the bootstrap analogs of Theorem 2.1, which establish the asymptotic normality of the WCB estimator and t-statistic. That is, for all $x \in \mathbb{R}$ and for all $\epsilon > 0$,

$$P^*\left(\frac{\boldsymbol{a}^\top (\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})}{(\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{1/2}} \leq x \right) \xrightarrow{P} \Phi(x), \tag{B.4}$$

$$P^*\left(\left| \frac{\boldsymbol{a}^\top \hat{\boldsymbol{V}}^* \boldsymbol{a}}{\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a}} - 1 \right| > \epsilon \right) \xrightarrow{P} 0, \tag{B.5}$$

$$P^*(t_a^* \leq x) \xrightarrow{P} \Phi(x). \tag{B.6}$$

From Corollary 2.1 and (B.6) it follows that

$$P_0(t_a \leq x) \to \Phi(x) \text{ and } P^*(t_a^* \leq x) \xrightarrow{P} \Phi(x),$$

respectively. The desired result then follows by application of the triangle inequality and Polya's Theorem, given that $\Phi(x)$ is everywhere continuous.

We thus need to prove (B.4)–(B.6), and we do so following the same outline as in the proof of Theorem 2.1. Under the WCB probability measure, we let $\ddot{\boldsymbol{\Gamma}} = N^{-2}\sum_{g=1}^{G} \boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g \ddot{\boldsymbol{u}}_g^\top \boldsymbol{X}_g$ and $\ddot{\boldsymbol{V}} = \boldsymbol{Q}_N^{-1}\ddot{\boldsymbol{\Gamma}}\boldsymbol{Q}_N^{-1}$ denote the bootstrap true values (i.e., the values generating the bootstrap data). First note that, by identical steps to those in the proof of Theorem 2.1, it holds that, under (14),

$$\frac{\boldsymbol{a}^\top(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)}{(\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{1/2}} = O_P(1) \quad \text{and} \quad \frac{\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a}}{\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a}} \xrightarrow{P} 1. \tag{B.7}$$

It follows from (B.7) that $\boldsymbol{a}^\top(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_N) = O_P(\mu_N^{-1})$. However, a more readily applicable consequence of (9), (B.7), and Assumption 2 is that

$$\|\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_N\| = O_P\Big(N^{-1/2} \sup_{g\in\mathbb{N}} N_g^{1/2}\Big) \quad \text{and} \quad (\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-1} = O_P(\mu_N). \tag{B.8}$$

Proof of (B.4). We define $z_g^* = (\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-1/2} \boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} N^{-1} \boldsymbol{X}_g^\top \boldsymbol{u}_g^*$ so that $(\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-1/2} \boldsymbol{a}^\top (\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}}) = \sum_{g=1}^{G} z_g^*$, and show that, for all $x \in \mathbb{R}$,

$$P^*\Big(\sum_{g=1}^{G} z_g^* \le x\Big) \xrightarrow{P} \Phi(x). \tag{B.9}$$

In view of (B.7), this suffices to prove (B.4). To show (B.9), we apply the Lyapunov Central Limit Theorem. 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 only requires verifying that the Lyapunov condition holds under the WCB probability measure for some $\xi > 0$ with P-probability converging to one; that is, we need to show that $\sum_{g=1}^{G} \mathrm{E}^*|z_g^*|^{2+\xi} \xrightarrow{P} 0$.

We first find that, because $H_N = \sum_{g=1}^{G} \|\boldsymbol{X}_g^\top \boldsymbol{X}_g\|^\theta$ is a non-negative random variable, $H_N = O_P(\mathrm{E}(H_N))$, and similarly for $\sum_{g=1}^{G} \|\boldsymbol{X}_g^\top \boldsymbol{u}_g\|^\theta$, and it then follows from Lemma A.2 that

$$\sum_{g=1}^{G} \|\boldsymbol{X}_g^\top \boldsymbol{X}_g\|^\theta = O_P\Big(N \sup_{g\in\mathbb{N}} N_g^{\theta-1}\Big) \quad \text{and} \quad \sum_{g=1}^{G} \|\boldsymbol{X}_g^\top \boldsymbol{u}_g\|^\theta = O_P\Big(N \sup_{g\in\mathbb{N}} N_g^{\theta-1}\Big) \tag{B.10}$$

for $1 \le \theta \le 4 + \lambda$ and $1 \le \theta \le 2 + \lambda/2$, respectively. We then find, because $\mathrm{E}^*|v_g|^\theta$ is a finite constant that does not depend on g and using the decomposition $\ddot{\boldsymbol{u}}_g = \boldsymbol{u}_g - \boldsymbol{X}_g(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)$ together with the c_r inequality,

$$\begin{aligned}
\mathrm{E}^* \sum_{g=1}^{G} \|\boldsymbol{X}_g^\top \boldsymbol{u}_g^*\|^\theta = \mathrm{E}^* \sum_{g=1}^{G} \|\boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g v_g^*\|^\theta &\le C \sum_{g=1}^{G} \|\boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g\|^\theta \\
&\le C \sum_{g=1}^{G} \|\boldsymbol{X}_g^\top \boldsymbol{u}_g\|^\theta + C \sum_{g=1}^{G} \|\boldsymbol{X}_g^\top \boldsymbol{X}_g\|^\theta \|\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_N\|^\theta = O_P\Big(N \sup_{g\in\mathbb{N}} N_g^{\theta-1}\Big),
\end{aligned} \tag{B.11}$$

where the last equality in (B.11) is due to (B.8) and (B.10). It then holds that

$$\sum_{g=1}^{G} \mathrm{E}^*|z_g^*|^{2+\xi} \le (\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-1-\xi/2} \|\boldsymbol{Q}_N^{-1}\|^{2+\xi} N^{-2-\xi} \mathrm{E}^* \sum_{g=1}^{G} \|\boldsymbol{X}_g^\top \boldsymbol{u}_g^*\|^{2+\xi} = O_P\Big(\mu_N^{1+\xi/2} \sup_{g\in\mathbb{N}} \frac{N_g^{1+\xi}}{N^{1+\xi}}\Big) \tag{B.12}$$

by (B.8) and (B.11). The right-hand side of (B.12) is $o_P(1)$ by Assumption 3 setting $\xi = \lambda/2 > 0$.

Proof of (B.5). 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

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

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

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

Using this decomposition, it suffices to prove that, for any $\epsilon > 0$, $P^*(|(\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-1} \boldsymbol{a}^\top \boldsymbol{B}_{mN}^* \boldsymbol{a}| > \epsilon) \xrightarrow{P} 0$ for $m = 1, 2, 3$. The proofs for each term roughly follow those for the corresponding term in the proof of (16).

For $m = 1$, use $\ddot{\boldsymbol{u}}_g = \boldsymbol{u}_g - \boldsymbol{X}_g(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)$ to write $\boldsymbol{B}_{1N}^* = \boldsymbol{B}_{11N}^* - \boldsymbol{B}_{12N}^* - \boldsymbol{B}_{12N}^{*\top} + \boldsymbol{B}_{13N}^*$ with

$$\boldsymbol{B}_{11N}^* = \boldsymbol{Q}_N^{-1} \frac{1}{N^2} \sum_{g=1}^G \boldsymbol{X}_g^\top \boldsymbol{u}_g \boldsymbol{u}_g^\top \boldsymbol{X}_g \boldsymbol{Q}_N^{-1} (v_g^{*2} - 1),$$

$$\boldsymbol{B}_{12N}^* = \boldsymbol{Q}_N^{-1} \frac{1}{N^2} \sum_{g=1}^G \boldsymbol{X}_g^\top \boldsymbol{u}_g (\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)^\top \boldsymbol{X}_g^\top \boldsymbol{X}_g \boldsymbol{Q}_N^{-1} (v_g^{*2} - 1), \text{ and}$$

$$\boldsymbol{B}_{13N}^* = \boldsymbol{Q}_N^{-1} \frac{1}{N^2} \sum_{g=1}^G \boldsymbol{X}_g^\top \boldsymbol{X}_g (\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)^\top \boldsymbol{X}_g^\top \boldsymbol{X}_g \boldsymbol{Q}_N^{-1} (v_g^{*2} - 1).$$

We first note that $|\boldsymbol{a}^\top \boldsymbol{B}_{12N}^* \boldsymbol{a}| \leq (\boldsymbol{a}^\top \boldsymbol{B}_{11N}^* \boldsymbol{a})^{1/2} (\boldsymbol{a}^\top \boldsymbol{B}_{13N}^* \boldsymbol{a})^{1/2}$ by the Cauchy-Schwarz inequality, so it suffices to prove the result for $j = 1$ and $j = 3$. Because $\mathrm{E}^*(v_g^{*2}) = 1$ we find that $\mathrm{E}^*((\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-1} \boldsymbol{a}^\top \boldsymbol{B}_{1jN}^* \boldsymbol{a}) = 0$ for $j = 1, 2, 3$. For $j = 1$ we find that $(\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-1} \boldsymbol{a}^\top \boldsymbol{B}_{11N}^* \boldsymbol{a} = \sum_{g=1}^G z_{1g}^*$, where $z_{1g}^* = (\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-1} \boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} N^{-2} \boldsymbol{X}_g^\top \boldsymbol{u}_g \boldsymbol{u}_g^\top \boldsymbol{X}_g \boldsymbol{Q}_N^{-1} (v_g^{*2} - 1)$, and we prove convergence in mean-square. Thus, by independence of z_{1g}^* (under the WCB probability measure),

$$\mathrm{Var}^*\Big(\sum_{g=1}^G z_{1g}^* \Big) = \sum_{g=1}^G \mathrm{Var}^*(z_{1g}^*) = \sum_{g=1}^G \mathrm{E}^*(z_{1g}^{*2}) \leq \mathrm{E}^*((v_g^{*2} - 1)^2) (\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-2} \|\boldsymbol{Q}_N^{-1}\|^4 N^{-4} \sum_{g=1}^G \|\boldsymbol{X}_g^\top \boldsymbol{u}\|^4,$$

which is $O_P(\mu_N^2 N^{-3} \sup_{g \in \mathbb{N}} N_g^3)$ by (B.8), (B.10), Assumption 2, and because $\mathrm{E}^*((v_g^{*2} - 1)^2)$ is a constant that does not depend on g. The result for $j = 1$ then follows from Assumption 3; see also (11). For $j = 3$ we prove convergence in L_1-norm, which implies convergence in probability. Thus,

$$\mathrm{E}^*|(\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-1} \boldsymbol{a}^\top \boldsymbol{B}_{13N}^* \boldsymbol{a}| \leq \|\boldsymbol{Q}_N^{-1}\|^2 (\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-1} \frac{1}{N^2} \Big\| \sum_{g=1}^G \boldsymbol{X}_g^\top \boldsymbol{X}_g (\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_N) \boldsymbol{X}_g^\top \boldsymbol{X}_g \Big\| \mathrm{E}^*|v_g^{*2} - 1|,$$

where $\mathrm{E}^*|v_g^{*2} - 1|$ is a finite constant that does not depend on g, $\|\boldsymbol{Q}_N^{-1}\|^2 = O_P(1)$ by Assumption 2, and $(\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-1} = O_P(\mu_N)$ by (B.8). We also find, by Minkowski's inequality,

$$\Big\| \sum_{g=1}^G \boldsymbol{X}_g^\top \boldsymbol{X}_g (\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_N) \boldsymbol{X}_g^\top \boldsymbol{X}_g \Big\| \leq \|\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_N\|^2 \sum_{g=1}^G \|\boldsymbol{X}_g^\top \boldsymbol{X}_g\|^2 = O_P\Big(\sup_{g \in \mathbb{N}} N_g^2 \Big),$$

where we used (B.8) and (B.10). It follows that

$$\mathrm{E}^*|(\boldsymbol{a}^\top \ddot{\boldsymbol{V}}\boldsymbol{a})^{-1}\boldsymbol{a}^\top \boldsymbol{B}^*_{13N}\boldsymbol{a}| = O_P\Big(\mu_N N^{-2}\sup_{g\in\mathbb{N}} N_g^2\Big) = o_P(1)$$

by Assumption 3; see also (11). This proves the result for $j = 3$ and hence for $m = 1$.

To prove the result for $m = 2$, we first apply the Cauchy-Schwarz inequality to obtain the bound

$$\begin{aligned}|\boldsymbol{a}^\top \boldsymbol{B}^*_{2N}\boldsymbol{a}| &\le \frac{1}{N^2}\Big(\sum_{g=1}^G (\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1}\boldsymbol{X}_g^\top \boldsymbol{u}_g^*)^2\Big)^{1/2}\Big(\sum_{g=1}^G ((\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})^\top \boldsymbol{X}_g^\top \boldsymbol{X}_g \boldsymbol{Q}_N^{-1}\boldsymbol{a})^2\Big)^{1/2} \\ &\le O_P(N^{-2})\|\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}}\|\Big(\sum_{g=1}^G \|\boldsymbol{X}_g^\top \boldsymbol{u}_g^*\|^2\Big)^{1/2}\Big(\sum_{g=1}^G \|\boldsymbol{X}_g^\top \boldsymbol{X}_g\|^2\Big)^{1/2}. \end{aligned} \tag{B.13}$$

Here, $\mathrm{E}^*\|\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}}\|^2 = O_P(\|\ddot{\boldsymbol{V}}\|) = O_P(N^{-1}\sup_{g\in\mathbb{N}} N_g)$, so for any $\zeta > 0$, by Chebyshev's inequality,

$$P^*(\|\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}}\| > \zeta^{-1/2}N^{-1/2}\sup_{g\in\mathbb{N}} N_g^{1/2}) \le \zeta N^{-1}\sup_{g\in\mathbb{N}} N_g \mathrm{E}^*\|\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}}\|^2 = \zeta O_P(1) = o_P(1) \tag{B.14}$$

by choosing ζ sufficiently small; cf. (B.8). It now follows from (B.13) and (B.14), together with (B.8), (B.10), and (B.11), that

$$P^*\Big(|(\boldsymbol{a}^\top \ddot{\boldsymbol{V}}\boldsymbol{a})^{-1}\boldsymbol{a}^\top \boldsymbol{B}^*_{2N}\boldsymbol{a}| > \zeta^{-1}\mu_N N^{-3/2}\sup_{g\in\mathbb{N}} N_g^{3/2}\Big) = \zeta O_P(1) = o_P(1)$$

by choosing ζ sufficiently small. Because $\mu_N N^{-3/2}\sup_{g\in\mathbb{N}} N_g^{3/2} \to 0$ under Assumption 3 (see also (11)), it follows that, for any $\epsilon > 0$, we can choose N large enough that $\zeta^{-1}\mu_N N^{-3/2}\sup_{g\in\mathbb{N}} N_g^{3/2} \le \epsilon$, which proves the result for $m = 2$.

Finally, the proof for $m = 3$ is similar to, but simpler than, that for $m = 2$. We use the bound

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

so that, as for $m = 2$,

$$P^*\Big(|(\boldsymbol{a}^\top \ddot{\boldsymbol{V}}\boldsymbol{a})^{-1}\boldsymbol{a}^\top \boldsymbol{B}^*_{3N}\boldsymbol{a}| > \zeta^{-1}\mu_N N^{-2}\sup_{g\in\mathbb{N}} N_g^2\Big) = \zeta O_P(1) = o_P(1)$$

and $\mu_N N^{-2}\sup_{g\in\mathbb{N}} N_g^2 \to 0$ under Assumption 3.

Proof of (B.6). Follows immediately by (B.4), (B.5), and Slutsky's Theorem.

B.3 Proof of Theorem 3.2

We first define some notation. 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, $\bar{\boldsymbol{V}}_N \ne \boldsymbol{V}_N$. We also 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 (note that these are not calculated under the WB algorithm, but serve only as useful constructions for the proof of Theorem 3.2).

The WB analogs of (B.4)–(B.6), which establish the asymptotic normality of the WB estimator and t-statistic, are as follows: for all $x \in \mathbb{R}$ and for all $\epsilon > 0$,

$$P^*\left(\frac{\boldsymbol{a}^\top(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})}{(\boldsymbol{a}^\top \bar{\boldsymbol{V}}_N \boldsymbol{a})^{1/2}} \leq x\right) \xrightarrow{P} \Phi(x), \tag{B.15}$$

$$P^*\left(\left|\frac{\boldsymbol{a}^\top \hat{\boldsymbol{V}}^* \boldsymbol{a}}{\boldsymbol{a}^\top \bar{\boldsymbol{V}}_N \boldsymbol{a}} - 1\right| > \epsilon\right) \xrightarrow{P} 0, \tag{B.16}$$

$$P^*(t_a^* \leq x) \xrightarrow{P} \Phi(x). \tag{B.17}$$

From Corollary 2.1 and (B.17) it follows that

$$P_0(t_a \leq x) \to \Phi(x) \text{ and } P^*(t_a^* \leq x) \xrightarrow{P} \Phi(x), \tag{B.18}$$

respectively. The desired result then follows by application of the triangle inequality and Polya's Theorem, given that $\Phi(x)$ is everywhere continuous.

We note that (B.15)–(B.17) in fact hold without Assumption 3, but instead imposing only the weaker condition in (10). This will be evident from the proofs given subsequently. However, this is only a theoretical curiosity because the use of Corollary 2.1 in (B.18) requires Assumption 3.

Before proving (B.15)–(B.17), we note that

$$(\boldsymbol{a}^\top \bar{\boldsymbol{V}}_N \boldsymbol{a})^{-1} = O_P(N), \quad \text{and} \quad \frac{\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a}}{\boldsymbol{a}^\top \bar{\boldsymbol{V}}_N \boldsymbol{a}} \xrightarrow{P} 1, \tag{B.19}$$

where the first statement follows directly from Assumption 2 and (7). To prove the second statement in (B.19) 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

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

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

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

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,

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

using the c_r inequality and Assumptions 1 and 2. It follows by Assumption 2 and (10) that $|N \boldsymbol{a}^\top \boldsymbol{C}_{1N} \boldsymbol{a}| = O_P(N^{-1/2} \sup_{g \in \mathbb{N}} N_g^{1/2}) = o_P(1)$.

For $m = 2$, we apply the bound

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

using the Cauchy-Schwarz inequality, (B.8), $\boldsymbol{Q}_N^{-1} = O_P(1)$, (10), and Assumptions 1 and 2. Finally, we turn to $m = 3$, where, by an identical argument, we obtain

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

Proof of (B.15). We now have $(\boldsymbol{a}^\top\bar{\boldsymbol{V}}_N\boldsymbol{a})^{-1/2}\boldsymbol{a}^\top(\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}}) = (\boldsymbol{a}^\top\ddot{\boldsymbol{V}}\boldsymbol{a})^{-1/2}(1+o_P(1))\boldsymbol{a}^\top\boldsymbol{Q}_N^{-1}N^{-1}\boldsymbol{X}^\top\boldsymbol{u}^*$ by (B.19). Under the WB probability measure, u_{ig}^* is heteroskedastic, but independent across both i and g. Let $z_{ig}^* = (\boldsymbol{a}^\top\ddot{\boldsymbol{V}}\boldsymbol{a})^{-1/2}\boldsymbol{a}^\top\boldsymbol{Q}_N^{-1}N^{-1}\boldsymbol{X}_{ig}^\top u_{ig}^*$, with $\mathrm{E}^*(z_{ig}^*) = 0$ and $\sum_{g=1}^{G}\sum_{i=1}^{N_g}\mathrm{E}^*(z_{ig}^{*2}) = 1$. The result 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 $\xi > 0$, $\sum_{g=1}^{G}\sum_{i=1}^{N_g}\mathrm{E}^*|z_{ig}^*|^{2+\xi} \xrightarrow{P} 0$.

By the c_r inequality,

$$\sum_{g=1}^{G}\sum_{i=1}^{N_g}\mathrm{E}^*|z_{ig}^*|^{2+\xi} \leq 2^{1+\xi}\sum_{g=1}^{G}\sum_{i=1}^{N_g}\mathrm{E}^*|z_{1ig}^*|^{2+\xi} + 2^{1+\xi}\sum_{g=1}^{G}\sum_{i=1}^{N_g}\mathrm{E}^*|z_{2ig}^*|^{2+\xi},$$

where $z_{1ig}^* = (\boldsymbol{a}^\top\ddot{\boldsymbol{V}}\boldsymbol{a})^{-1/2}\boldsymbol{a}^\top\boldsymbol{Q}_N^{-1}N^{-1}\boldsymbol{X}_{ig}^\top u_{ig}v_{ig}^*$ and $z_{2ig}^* = (\boldsymbol{a}^\top\ddot{\boldsymbol{V}}\boldsymbol{a})^{-1/2}\boldsymbol{a}^\top\boldsymbol{Q}_N^{-1}N^{-1}\boldsymbol{X}_{ig}^\top\boldsymbol{X}_{ig}(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)v_{ig}^*$. We first obtain the bound

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

Since $H_N = \sum_{g=1}^{G}\sum_{i=1}^{N_g}\mathrm{E}^*\|\boldsymbol{X}_{ig}^\top u_{ig}v_{ig}^*\|^{2+\xi}$ is a non-negative random variable, $H_N = O_P(\mathrm{E}(H_N))$, and we find that

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

which is $O(N)$ by Assumption 1 for $\xi \leq 2 + \lambda$. It follows, using also (B.19), that

$$\sum_{g=1}^{G}\sum_{i=1}^{N_g}\mathrm{E}^*|z_{1ig}^*|^{2+\xi} = O_P(N^{-\xi/2}) = o_P(1) \tag{B.20}$$

by choosing $0 < \xi \leq 2 + \lambda$. Next, by (B.8) and (B.19),

$$\begin{aligned}\mathrm{E}^*|z_{2ig}^*|^{2+\xi} &\leq \mathrm{E}^*|v_{ig}^*|^{2+\xi}(\boldsymbol{a}^\top\ddot{\boldsymbol{V}}\boldsymbol{a})^{-1-\xi/2}\Big|\boldsymbol{a}^\top\boldsymbol{Q}_N^{-1}N^{-1}\boldsymbol{X}_{ig}^\top\boldsymbol{X}_{ig}(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)\Big|^{2+\xi} \\ &= O_P\Big(N^{1+\xi/2}N^{-3-3\xi/2}\sup_{g\in\mathbb{N}} N_g^{1+\xi/2}\Big)\|\boldsymbol{X}_{ig}^\top\boldsymbol{X}_{ig}\|^{2+\xi}.\end{aligned}$$

As in (B.10), $\sum_{g=1}^{G}\sum_{i=1}^{N_g}\|\boldsymbol{X}_{ig}^\top\boldsymbol{X}_{ig}\|^{2+\xi}=O_P(N)$ by Assumption 2 with $0<\xi\leq\lambda/2$, so that

$$\sum_{g=1}^{G}\sum_{i=1}^{N_g}\mathrm{E}^*|z_{2ig}^*|^{2+\xi}=O_P\Big(N^{-2-\xi}\sup_{g\in\mathbb{N}}N_g^{1+\xi/2}\Big)\sum_{g=1}^{G}\sum_{i=1}^{N_g}\|\boldsymbol{X}_{ig}^\top\boldsymbol{X}_{ig}\|^{2+\xi}=O_P\Big(N^{-1-\xi}\sup_{g\in\mathbb{N}}N_g^{1+\xi/2}\Big),$$

which is $=o_P(1)$ by (10), and this proves (B.15).

Proof of (B.16). In light of the two results in (B.19), the result (B.5) follows if, for any $\epsilon>0$, $P^*\big(|(\boldsymbol{a}^\top\ddot{\boldsymbol{V}}\boldsymbol{a})^{-1}\boldsymbol{a}^\top\hat{\boldsymbol{V}}^*\boldsymbol{a}-1|>\epsilon\big)\xrightarrow{P}0$. To prove this, we apply the decomposition

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

where

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

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

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

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

It suffices to prove that, for any $\epsilon>0$, $P^*\big(|(\boldsymbol{a}^\top\bar{\boldsymbol{V}}_N\boldsymbol{a})^{-1}\boldsymbol{a}^\top\boldsymbol{D}_{mN}^*\boldsymbol{a}|>\epsilon\big)\xrightarrow{P}0$, in probability, for $m=1,\ldots,4$. Equivalently, by (B.19), we can replace $(\boldsymbol{a}^\top\bar{\boldsymbol{V}}_N\boldsymbol{a})^{-1}$ by either $(\boldsymbol{a}^\top\ddot{\boldsymbol{V}}\boldsymbol{a})^{-1}$ or by N.

To prove the result for $m=1$, we use $\ddot{u}_{ig}=u_{ig}-\boldsymbol{X}_{ig}(\ddot{\boldsymbol{\beta}}-\boldsymbol{\beta}_N)$ to decompose $\boldsymbol{D}_{1N}^*=\boldsymbol{D}_{11N}^*-\boldsymbol{D}_{12N}^*-\boldsymbol{D}_{12N}^{*\top}+\boldsymbol{D}_{13N}^*$, where

$$\boldsymbol{D}_{11N}^*=\boldsymbol{Q}_N^{-1}\frac{1}{N^2}\sum_{g=1}^{G}\sum_{i=1}^{N_g}\boldsymbol{X}_g^\top u_{ig}u_{ig}^\top\boldsymbol{X}_{ig}\boldsymbol{Q}_N^{-1}(v_{ig}^{*2}-1),$$

$$\boldsymbol{D}_{12N}^*=\boldsymbol{Q}_N^{-1}\frac{1}{N^2}\sum_{g=1}^{G}\sum_{i=1}^{N_g}\boldsymbol{X}_{ig}^\top u_{ig}(\ddot{\boldsymbol{\beta}}-\boldsymbol{\beta}_N)^\top\boldsymbol{X}_{ig}^\top\boldsymbol{X}_{ig}\boldsymbol{Q}_N^{-1}(v_{ig}^{*2}-1),\text{ and}$$

$$\boldsymbol{D}_{13N}^*=\boldsymbol{Q}_N^{-1}\frac{1}{N^2}\sum_{g=1}^{G}\sum_{i=1}^{N_g}\boldsymbol{X}_{ig}^\top\boldsymbol{X}_{ig}(\ddot{\boldsymbol{\beta}}-\boldsymbol{\beta}_N)(\ddot{\boldsymbol{\beta}}-\boldsymbol{\beta}_N)^\top\boldsymbol{X}_{ig}^\top\boldsymbol{X}_{ig}\boldsymbol{Q}_N^{-1}(v_{ig}^{*2}-1).$$

First note that $|\boldsymbol{a}^\top\boldsymbol{D}_{12N}^*\boldsymbol{a}|\leq(\boldsymbol{a}^\top\boldsymbol{D}_{11N}^*\boldsymbol{a})^{1/2}(\boldsymbol{a}^\top\boldsymbol{D}_{13N}^*\boldsymbol{a})^{1/2}$ by the Cauchy-Schwarz inequality, so it suffices to prove the result for $j=1$ and $j=3$. Because $\mathrm{E}^*(v_{ig}^{*2})=1$ we find that $\mathrm{E}^*(N\boldsymbol{a}^\top\boldsymbol{D}_{1jN}^*\boldsymbol{a})=0$ for $j=1,2,3$. For $j=1$, $(\boldsymbol{a}^\top\ddot{\boldsymbol{V}}\boldsymbol{a})^{-1}\boldsymbol{a}^\top\boldsymbol{D}_{11N}^*\boldsymbol{a}=\sum_{g=1}^{G}\sum_{i=1}^{N_g}z_{1ig}^{*2}-1$, and we prove convergence in mean-square. By independence of z_{1ig}^{*2} across i and g (under the WB probability measure),

$$\mathrm{Var}^*\Big(\sum_{g=1}^{G}\sum_{i=1}^{N_g}z_{1ig}^{*2}-1\Big)=\sum_{g=1}^{G}\sum_{i=1}^{N_g}\mathrm{Var}^*(z_{1ig}^{*2})=\sum_{g=1}^{G}\sum_{i=1}^{N_g}\mathrm{E}^*(z_{1ig}^{*4})-\sum_{g=1}^{G}\sum_{i=1}^{N_g}\big(\mathrm{E}^*(z_{1ig}^{*2})\big)^2=o_P(1)\quad\text{(B.21)}$$

using the Lyapunov condition (B.20) for $\xi = 2$ and Jensen's inequality, which proves the result for $j = 1$. For $j = 3$ we prove convergence in L_1-norm. Thus,

$$\mathrm{E}^*\Big|N\boldsymbol{a}^\top \boldsymbol{D}^*_{13N}\boldsymbol{a}\Big| \le N\|\boldsymbol{Q}_N^{-1}\|^2\frac{1}{N^2}\Big\|\sum_{g=1}^{G}\sum_{i=1}^{N_g}\boldsymbol{X}_{ig}^\top\boldsymbol{X}_{ig}(\ddot{\boldsymbol{\beta}}-\boldsymbol{\beta}_N)(\ddot{\boldsymbol{\beta}}-\boldsymbol{\beta}_N)\boldsymbol{X}_{ig}^\top\boldsymbol{X}_{ig}\Big\|\mathrm{E}^*|v_{ig}^{*2}-1|$$
$$\le O_P\Big(N^{-2}\sup_{g\in\mathbb{N}}N_g\Big)\sum_{g=1}^{G}\sum_{i=1}^{N_g}\|\boldsymbol{X}_{ig}^\top\boldsymbol{X}_{ig}\|^2 = O_P\Big(N^{-1}\sup_{g\in\mathbb{N}}N_g\Big) = o_P(1)$$

as above, using that $\|\boldsymbol{X}_{ig}^\top\boldsymbol{X}_{ig}\|^\theta$ is a non-negative random variable, so that, by Assumption 2,

$$\sum_{g=1}^{G}\sum_{i=1}^{N_g}\|\boldsymbol{X}_{ig}^\top\boldsymbol{X}_{ig}\|^\theta = O_P\Big(\sum_{g=1}^{G}\sum_{i=1}^{N_g}\mathrm{E}\|\boldsymbol{X}_{ig}^\top\boldsymbol{X}_{ig}\|^\theta\Big) = O_P(N) \tag{B.22}$$

for $\theta \le 2 + \lambda/2$; see also (B.10). This proves the result for $j = 3$ and hence for $m = 1$.

For $m = 2$, we again decompose $\boldsymbol{D}^*_{2N} = \boldsymbol{D}^*_{21N} + \boldsymbol{D}^*_{22N} + \boldsymbol{D}^{*\top}_{22N} + \boldsymbol{D}^*_{23N}$, where

$$\boldsymbol{D}^*_{21N} = \boldsymbol{Q}_N^{-1}\frac{1}{N^2}\sum_{g=1}^{G}\sum_{i\neq j=1}^{N_g}\boldsymbol{X}_{ig}^\top u_{ig}u_{jg}\boldsymbol{X}_{jg}\boldsymbol{Q}_N^{-1}v^*_{ig}v^*_{jg},$$
$$\boldsymbol{D}^*_{22N} = \boldsymbol{Q}_N^{-1}\frac{1}{N^2}\sum_{g=1}^{G}\sum_{i\neq j=1}^{N_g}\boldsymbol{X}_{ig}^\top u_{ig}(\ddot{\boldsymbol{\beta}}-\boldsymbol{\beta}_N)^\top\boldsymbol{X}_{jg}^\top\boldsymbol{X}_{jg}\boldsymbol{Q}_N^{-1}v^*_{ig}v^*_{jg}, \text{ and}$$
$$\boldsymbol{D}^*_{23N} = \boldsymbol{Q}_N^{-1}\frac{1}{N^2}\sum_{g=1}^{G}\sum_{i\neq j=1}^{N_g}\boldsymbol{X}_{ig}^\top\boldsymbol{X}_{ig}(\ddot{\boldsymbol{\beta}}-\boldsymbol{\beta}_N)(\ddot{\boldsymbol{\beta}}-\boldsymbol{\beta}_N)^\top\boldsymbol{X}_{jg}^\top\boldsymbol{X}_{jg}\boldsymbol{Q}_N^{-1}v^*_{ig}v^*_{jg},$$

and by the Cauchy-Schwarz inequality we only need to prove the result for $\boldsymbol{D}^*_{2jN}$ with $j = 1$ and $j = 3$. For $j = 1$, we use independence of v^*_{ig} across both i and g and prove convergence in mean-square. Hence,

$$\mathrm{E}^*(N\boldsymbol{a}^\top\boldsymbol{D}^*_{21N}\boldsymbol{a})^2 \le \|\boldsymbol{Q}_N^{-1}\|^4\frac{1}{N^2}\sum_{g=1}^{G}\sum_{i\neq j=1}^{N_g}\|\boldsymbol{X}_{ig}^\top u_{ig}\|^2\|\boldsymbol{X}_{jg}^\top u_{jg}\|^2, \tag{B.23}$$

where the summation on the right-hand side is a non-negative random variable with mean

$$\sum_{g=1}^{G}\sum_{i\neq j=1}^{N_g}\mathrm{E}\Big(\|\boldsymbol{X}_{ig}^\top u_{ig}\|^2\|\boldsymbol{X}_{jg}^\top u_{jg}\|^2\Big) \le \sum_{g=1}^{G}\sum_{i\neq j=1}^{N_g}\Big(\mathrm{E}\|\boldsymbol{X}_{ig}^\top u_{ig}\|^4\Big)^{1/2}\Big(\mathrm{E}\|\boldsymbol{X}_{jg}^\top u_{jg}\|^4\Big)^{1/2},$$

which is $O_P(N\sup_{g\in\mathbb{N}}N_g)$ by Assumption 1. It then follows from (B.23), using also Assumption 2 and Markov's inequality, that, for any $\zeta > 0$, $P^*(|N\boldsymbol{a}^\top\boldsymbol{D}^*_{21N}\boldsymbol{a}| > \zeta^{-1}N^{-1/2}\sup_{g\in\mathbb{N}}N_g^{1/2}) = \zeta O_P(1) = o_P(1)$. This proves the result for $j = 1$ because $N^{-1/2}\sup_{g\in\mathbb{N}}N_g^{1/2} \to 0$ by (10). For $j = 3$ we also prove convergence in mean-square and find

$$\mathrm{E}^*(N\boldsymbol{a}^\top\boldsymbol{D}^*_{23N}\boldsymbol{a})^2 \le \|\boldsymbol{Q}_N^{-1}\|^4\frac{1}{N^2}\|\ddot{\boldsymbol{\beta}}-\boldsymbol{\beta}_N\|^4\sum_{g=1}^{G}\sum_{i\neq j=1}^{N_g}\|\boldsymbol{X}_{ig}^\top\boldsymbol{X}_{ig}\|^2\|\boldsymbol{X}_{jg}^\top\boldsymbol{X}_{jg}\|^2$$
$$\le \|\boldsymbol{Q}_N^{-1}\|^4\frac{1}{N^2}\|\ddot{\boldsymbol{\beta}}-\boldsymbol{\beta}_N\|^4\Big(\sum_{g=1}^{G}\sum_{i=1}^{N_g}\|\boldsymbol{X}_{ig}^\top\boldsymbol{X}_{ig}\|^2\Big)^2 = O_P\Big(N^{-2}\sup_{g\in\mathbb{N}}N_g^2\Big) = o_P(1),$$

where we used (B.8) together with Assumption 2 and (B.22). The last equality follows from (10).

For $m = 3$, we apply the Cauchy-Schwarz inequality as in (B.13) and find

$$|N\boldsymbol{a}^\top \boldsymbol{D}^*_{3N}\boldsymbol{a}| \leq \frac{1}{N}\Big(\sum_{g=1}^{G}(\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1}\boldsymbol{X}_g^\top \boldsymbol{u}_g^*)^2\Big)^{1/2}\Big(\sum_{g=1}^{G}((\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})^\top \boldsymbol{X}_g^\top \boldsymbol{X}_g \boldsymbol{Q}_N^{-1}\boldsymbol{a})^2\Big)^{1/2}. \tag{B.24}$$

The term inside the first large parentheses in (B.24) is a non-negative random variable with mean (under the WB probability measure)

$$\mathrm{E}^*\sum_{g=1}^{G}(\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1}\boldsymbol{X}_g^\top \boldsymbol{u}_g^*)^2 = \sum_{g=1}^{G}\sum_{i=1}^{N_g}\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1}\boldsymbol{X}_{ig}^\top \ddot{u}_{ig}^2 \boldsymbol{X}_{ig}\boldsymbol{Q}_N^{-1}\boldsymbol{a} = N^2\boldsymbol{a}^\top \ddot{\boldsymbol{V}}\boldsymbol{a} = O_P(N) \tag{B.25}$$

by (B.19) and (7). The term inside the second large parentheses in (B.24) is

$$\sum_{g=1}^{G}((\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})^\top \boldsymbol{X}_g^\top \boldsymbol{X}_g \boldsymbol{Q}_N^{-1}\boldsymbol{a})^2 = O_P(1)\|\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}}\|^2\sum_{g=1}^{G}\|\boldsymbol{X}_g^\top \boldsymbol{X}_g\|^2 = \|\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}}\|^2 O_P(N\sup_{g\in\mathbb{N}}N_g) \tag{B.26}$$

using (B.10). By an identical argument to that in (B.14), under the WB probability measure, $P^*(\|\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}}\| > \zeta^{-1}N^{-1/2}) = o_P(1)$. Combining (B.24), (B.25), and (B.26),

$$P^*(|N\boldsymbol{a}^\top \boldsymbol{D}^*_{3N}\boldsymbol{a}| > \zeta^{-1}N^{-1/2}\sup_{g\in\mathbb{N}}N_g^{1/2}) = \zeta O_P(1) = o_P(1),$$

where $N^{-1/2}\sup_{g\in\mathbb{N}}N_g^{1/2} \to 0$ by (10), which proves the result for $m = 3$. Finally, by very similar arguments, we find for $m = 4$ that

$$|N\boldsymbol{a}^\top \boldsymbol{D}^*_{4N}\boldsymbol{a}| \leq N\|\boldsymbol{Q}_N^{-1}\|^2\frac{1}{N^2}\sum_{g=1}^{G}\|\boldsymbol{X}_g^\top \boldsymbol{X}_g\|^2\|\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}}\|^2,$$

which satisfies

$$P^*\Big(|N\boldsymbol{a}^\top \boldsymbol{D}^*_{4N}\boldsymbol{a}| > \zeta^{-1}N^{-1}\sup_{1\leq g\leq G}N_g\Big) = \zeta O_P(1) = o_P(1),$$

and the result for $m = 4$ follows because $N^{-1}\sup_{1\leq g\leq G}N_g \to 0$ by (10).

Proof of (B.17). Follows immediately by (B.15), (B.16), and Slutsky's Theorem.

B.4 Proof of Theorem 5.1

Following Ch. 2 of Hall (1992), in particular Theorems 2.1 and 2.2, we consider Taylor-series approximants, $\tilde{t}_a^k$, to t_a^k and define the approximate cumulant functions

$$\Pi_{1N}(t_a) = \mathrm{E}(\tilde{t}_a), \tag{B.27}$$
$$\Pi_{2N}(t_a) = \mathrm{E}(\tilde{t}_a^2) - (\mathrm{E}(\tilde{t}_a))^2, \tag{B.28}$$
$$\Pi_{3N}(t_a) = \mathrm{E}(\tilde{t}_a^3) - 3\mathrm{E}(\tilde{t}_a^2)\mathrm{E}(\tilde{t}_a) + 2(\mathrm{E}(\tilde{t}_a))^3, \tag{B.29}$$
$$\Pi_{4N}(t_a) = \mathrm{E}(\tilde{t}_a^4) - 4\mathrm{E}(\tilde{t}_a^3)\mathrm{E}(\tilde{t}_a) - 3(\mathrm{E}(\tilde{t}_a^2))^2 + 12\mathrm{E}(\tilde{t}_a^2)(\mathrm{E}(\tilde{t}_a))^2 - 6(\mathrm{E}(\tilde{t}_a))^4. \tag{B.30}$$

Then

$$q_1(x) = -(\kappa_1 + \frac{1}{6}\kappa_3(x^2 - 1)) \text{ and} \tag{B.31}$$
$$q_2(x) = -\frac{1}{2}(\kappa_2 + \kappa_1^2)x - \frac{1}{24}(\kappa_4 + 4\kappa_1\kappa_3)(x^3 - 3x) - \frac{1}{72}\kappa_3^2(x^5 - 10x^3 + 15x), \tag{B.32}$$

where κ_1 and κ_3 are the coefficients of the terms of order $O(G^{-1/2})$ in an asymptotic expansion of $\Pi_{1N}(t_a)$ and $\Pi_{3N}(t_a)$, respectively, while κ_2 and κ_4 are the coefficients of the terms of order $O(G^{-1})$ in an asymptotic expansion of $\Pi_{2N}(t_a)$ and $\Pi_{4N}(t_a)$, respectively. We analogously define the corresponding bootstrap cumulants $\ddot{\Pi}_{jN}(t_a^*)$ for $j = 1, \ldots, 4$, replacing the population mean $\mathrm{E}(\cdot)$ by the bootstrap analog $\mathrm{E}^*(\cdot)$, and deduce $\ddot{\kappa}_j$, and hence $\ddot{q}_1$ and $\ddot{q}_2$, in the same way as κ_j.

The remainder of the proof is divided into three parts. First, we derive the Taylor-series approximants, $\tilde{t}_a^k$, to powers of the sample t-statistic. Then we use these approximants to find expansions of the cumulants $\Pi_{jN}(t_a)$ as needed to determine the coefficients κ_j, for $j = 1, \ldots, 4$. In the final part, we derive the corresponding results for (both versions of) the bootstrap t-statistic.

Taylor-series approximants to t_a^k. From (26) we find that t_a is

$$t_a = S_N\big(1 + G^{-1/2}(U_N + G^{-1/2}T_N)\big)^{-1/2}, \tag{B.33}$$

where $T_N = T_{1N} + T_{2N}$ and

$$S_N = \frac{1}{\sqrt{G}}\sum_{g=1}^{G} W_g = O_P(1), \qquad U_N = \frac{1}{\sqrt{G}}\sum_{g=1}^{G}(W_g^2 - 1) = O_P(1), \tag{B.34}$$

$$T_{1N} = -2\sum_{g=1}^{G} W_g Z_g = O_P(1), \qquad T_{2N} = \sum_{g=1}^{G} Z_g^2 = O_P(1). \tag{B.35}$$

The orders of magnitude in (B.34) and (B.35) are derived as follows. First, S_N and U_N are both sums of independent summands with mean zero, so that (B.34) follows by Lemmas A.1 and A.3 with $\theta = 2$. Next, expanding Z_g as in (A.2) and applying Lemma A.3, (B.35) follows straightforwardly.

By second-order Taylor-series expansion of $(1+x)^{-1/2}$ around $x = 0$, we find

$$\Big(1 + G^{-1/2}(U_N + G^{-1/2}T_N)\Big)^{-1/2} = 1 - G^{-1/2}\frac{1}{2}U_N + G^{-1}\Big(-\frac{1}{2}T_N + \frac{3}{8}U_N^2\Big) + O_P(G^{-3/2}).$$

From (B.33) and the orders in (B.34)–(B.35), we then obtain the approximation

$$t_a = S_N + G^{-1/2}\Big(-\frac{1}{2}S_N U_N\Big) + G^{-1}\Big(-\frac{1}{2}S_N T_N + \frac{3}{8}S_N U_N^2\Big) + O_P(G^{-3/2}). \tag{B.36}$$

Finally, each of the Taylor-series approximants, $\tilde{t}_a^k$, is found by taking the relevant power of (B.36) and eliminating terms that are at most $O_P(G^{-3/2})$.

Expansions of cumulants $\Pi_{jN}(t_a)$. Taking expectations of $\tilde{t}_a^k$ as defined above, and using (B.36) and Lemma A.4, we find

$$\mathrm{E}(\tilde{t}_a) = -\frac{1}{2}G^{-1/2}\gamma_N + O(G^{-3/2}), \quad \mathrm{E}(\tilde{t}_a^2) = 1 + G^{-1}(2\gamma_N^2 - \tau_{1N} + 2\tau_{2N}) + O(G^{-2}),$$

$$\mathrm{E}(\tilde{t}_a^3) = -\frac{7}{2}G^{-1/2}\gamma_N + O(G^{-3/2}), \quad \mathrm{E}(\tilde{t}_a^4) = 3 + G^{-1}(28\gamma_N^2 - 2\xi_N - 12\tau_{1N} + 12\tau_{2N} - 6\tau_{3N}) + O(G^{-2}).$$

Inserting these expressions into (B.27)–(B.30), we obtain the cumulants

$$\Pi_{1N}(t_a) = -\frac{1}{2}G^{-1/2}\gamma_N + O(G^{-3/2}), \quad \Pi_{2N}(t_a) = 1 + G^{-1}\Big(\frac{7}{4}\gamma_N^2 - \tau_{1N} + 2\tau_{2N}\Big) + O(G^{-2}),$$

$$\Pi_{3N}(t_a) = -2G^{-1/2}\gamma_N + O(G^{-3/2}), \quad \Pi_{4N}(t_a) = G^{-1}(12\gamma_N^2 - 2\xi_N - 6\tau_{1N} - 6\tau_{3N}) + O(G^{-2}).$$

We finally conclude that

$$\kappa_1 = -\frac{1}{2}\gamma_N, \qquad \kappa_2 = \frac{7}{4}\gamma_N^2 - \tau_{1N} + 2\tau_{2N}, \qquad \kappa_3 = -2\gamma_N, \qquad \kappa_4 = 12\gamma_N^2 - 2\xi_N - 6\tau_{1N} - 6\tau_{3N}.$$

In view of the moment conditions in Lemma A.4, we note that $\kappa_1, \kappa_2, \kappa_3$ exist under the conditions of the one-term expansion ($m = 1$) of Theorem 5.1, while κ_4 exists under the conditions of the two-term expansion ($m = 2$). Thus, we find the results of Theorem 5.1 from (B.31) and (B.32).

Expansions for bootstrap t-statistic. This proof is identical to that for the sample t-statistic, replacing the population mean $\mathrm{E}(\cdot)$ by the bootstrap analog $\mathrm{E}^*(\cdot)$ and replacing W_g and Z_g by W_g^* and Z_g^*, respectively.

B.5 Proof of Theorem 5.2

First we find that

$$\ddot{\gamma}_N = \frac{1}{G}\sum_{g=1}^{G} \mathrm{E}^*(W_g^{*3}) = \frac{1}{G}\sum_{g=1}^{G} (\boldsymbol{a}^\top \ddot{\boldsymbol{V}}_N \boldsymbol{a})^{-3/2}\, \mathrm{E}^*\Big(\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} \frac{\sqrt{G}}{N} \boldsymbol{X}_g^\top \boldsymbol{u}_g^*\Big)^3.$$

However, $\boldsymbol{u}_g^* = \ddot{\boldsymbol{u}}_g v_g^*$, where v_g^* is a scalar and $\mathrm{E}^*(v_g^{*3}) = \mathrm{E}^*(v^{*3})$ is constant, so that

$$\ddot{\gamma}_N = \mathrm{E}^*(v^{*3}) \frac{1}{G}\sum_{g=1}^{G} (\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-3/2} \Big(\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} \frac{\sqrt{G}}{N} \boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g\Big)^3 = \mathrm{E}^*(v_g^{*3})(\gamma_N + B_{1N} + B_{2N} + B_{3N} + B_{4N}),$$

where

$$B_{1N} = (\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-3/2} \frac{1}{G}\sum_{g=1}^{G} \left(\Big(\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} \frac{\sqrt{G}}{N} \boldsymbol{X}_g^\top \boldsymbol{u}_g\Big)^3 - \mathrm{E}\Big(\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} \frac{\sqrt{G}}{N} \boldsymbol{X}_g^\top \boldsymbol{u}_g\Big)^3 \right),$$

$$B_{2N} = (\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-3/2} \frac{1}{G}\sum_{g=1}^{G} \left(\Big(\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} \frac{\sqrt{G}}{N} \boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g\Big)^3 - \Big(\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} \frac{\sqrt{G}}{N} \boldsymbol{X}_g^\top \boldsymbol{u}_g\Big)^3 \right),$$

$$B_{3N} = ((\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-3/2} - (\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-3/2}) \frac{1}{G}\sum_{g=1}^{G} \Big(\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} \frac{\sqrt{G}}{N} \boldsymbol{X}_g^\top \boldsymbol{u}_g\Big)^3,$$

$$B_{4N} = ((\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-3/2} - (\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-3/2}) \frac{1}{G}\sum_{g=1}^{G} \left(\Big(\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} \frac{\sqrt{G}}{N} \boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g\Big)^3 - \Big(\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} \frac{\sqrt{G}}{N} \boldsymbol{X}_g^\top \boldsymbol{u}_g\Big)^3 \right),$$

and we analyze each term B_{iN}, for $i = 1, \ldots, 4$, in turn.

First note that $B_{1N} = G^{-1}\sum_{g=1}^{G}(W_g^3 - \mathrm{E}(W_g^3))$, where $W_g^3 - \mathrm{E}(W_g^3)$ is an independent, mean-zero sequence with finite second moments by Lemma A.3 since we have assumed $\lambda = 2$ in Assumption 6. It follows from Lemma A.1 that $B_{1N} = O_P(G^{-1/2})$. When $\lambda > 2$ is assumed, we apply the Lyapunov Central Limit Theorem to $z_{1g} = G^{-1/2}(W_g^3 - \mathrm{E}(W_g^3))$ jointly with other terms below.

To analyze B_{2N}, we use the decomposition $\ddot{\boldsymbol{u}}_g = \boldsymbol{u}_g - \boldsymbol{X}_g(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ and find

$$B_{2N} = 3B_{21N} - 3B_{22N} - B_{23N},$$

where

$$B_{21N} = (\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-3/2} \frac{1}{G}\sum_{g=1}^{G} \boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} \frac{\sqrt{G}}{N} \boldsymbol{X}_g^\top \boldsymbol{u}_g \Big(\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} \frac{\sqrt{G}}{N} \boldsymbol{X}_g^\top \boldsymbol{X}_g(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\Big)^2 = \frac{1}{G}\sum_{g=1}^{G} W_g Z_g^2(A),$$

$$B_{22N} = (\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-3/2} \frac{1}{G}\sum_{g=1}^{G} \Big(\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} \frac{\sqrt{G}}{N} \boldsymbol{X}_g^\top \boldsymbol{u}_g\Big)^2 \boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} \frac{\sqrt{G}}{N} \boldsymbol{X}_g^\top \boldsymbol{X}_g(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \frac{1}{G}\sum_{g=1}^{G} W_g^2 Z_g(A),$$

$$B_{23N} = (\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-3/2} \frac{1}{G}\sum_{g=1}^{G} \Big(\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1} \frac{\sqrt{G}}{N} \boldsymbol{X}_g^\top \boldsymbol{X}_g(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\Big)^3 = \frac{1}{G}\sum_{g=1}^{G} Z_g^3(A).$$

It follows directly from Lemma A.3 that $\mathrm{E}|B_{21N}| = O(G^{-1})$ and $\mathrm{E}|B_{23N}| = O(G^{-3/2})$ so that $B_{21N} = O_P(G^{-1})$ and $B_{23N} = O_P(G^{-3/2})$. Next, we write $B_{22N} = G^{-1}\sum_{g=1}^{G}(W_g^2 - \mathrm{E}W_g^2)Z_g(A) +$

$G^{-1}\sum_{g=1}^{G}(\mathrm{E}W_g^2)Z_g(A)$, where the second moment of the first term is $O(G^{-2})$ using (A.2) and Lemma A.3 because $(W_g^2 - \mathrm{E}W_g^2)$ has mean zero. Letting $z_{2g} = -3G^{-1/2}\sum_{h=1}^{G}(\mathrm{E}W_h^2)V_{hg}(A)$, it follows from (A.2) that $G^{1/2}B_{2N} = \sum_{g=1}^{G} z_{2g} + O_P(G^{-1/2})$, where we again apply the Lyapunov Central Limit Theorem to z_{2g} jointly with other terms below.

For the analysis of B_{3N}, we first find, by Taylor-series expansion,

$$(\boldsymbol{a}^\top \ddot{\boldsymbol{V}}\boldsymbol{a})^{-3/2} - (\boldsymbol{a}^\top \boldsymbol{V}_N\boldsymbol{a})^{-3/2} = -\frac{3}{2}(\boldsymbol{a}^\top \boldsymbol{V}_N\boldsymbol{a})^{-5/2}\boldsymbol{a}^\top(\ddot{\boldsymbol{V}} - \boldsymbol{V}_N)\boldsymbol{a}(1 + O_P(G^{-1})),$$

which implies

$$B_{3N} = -\frac{3}{2}(\boldsymbol{a}^\top \boldsymbol{V}_N\boldsymbol{a})^{-3/2}\frac{1}{G}\sum_{g=1}^{G}\Big(\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1}\frac{\sqrt{G}}{N}\boldsymbol{X}_g^\top \boldsymbol{u}_g\Big)^3(\boldsymbol{a}^\top \boldsymbol{V}_N\boldsymbol{a})^{-1}\boldsymbol{a}^\top(\ddot{\boldsymbol{V}} - \boldsymbol{V}_N)\boldsymbol{a}(1 + O_P(G^{-1})). \tag{B.37}$$

Next, we note from the analysis of B_{1N} above that

$$(\boldsymbol{a}^\top \boldsymbol{V}_N\boldsymbol{a})^{-3/2}G^{-1}\sum_{g=1}^{G}\Big(\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1}\frac{\sqrt{G}}{N}\boldsymbol{X}_g^\top \boldsymbol{u}_g\Big)^3 - \gamma_N = B_{1N} = O_P(G^{-1/2}). \tag{B.38}$$

Then, using $\ddot{\boldsymbol{u}}_g = \boldsymbol{u}_g - \boldsymbol{X}_g(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, we find that

$$\begin{aligned}(\boldsymbol{a}^\top \boldsymbol{V}_N\boldsymbol{a})^{-1}\boldsymbol{a}^\top(\ddot{\boldsymbol{V}} - \boldsymbol{V}_N)\boldsymbol{a} &= (\boldsymbol{a}^\top \boldsymbol{V}_N\boldsymbol{a})^{-1}\frac{1}{N^2}\sum_{g=1}^{G}\Big((\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1}\boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g)^2 - \mathrm{E}((\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1}\boldsymbol{X}_g^\top \boldsymbol{u}_g)^2)\Big)\\ &= B_{31N} - 2B_{32N} + B_{33N}\end{aligned}$$

with

$$B_{31N} = (\boldsymbol{a}^\top \boldsymbol{V}_N\boldsymbol{a})^{-1}\frac{1}{G}\sum_{g=1}^{G}\Big(\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1}\frac{\sqrt{G}}{N}\boldsymbol{X}_g^\top \boldsymbol{X}_g(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\Big)^2 = \frac{1}{G}\sum_{g=1}^{G}Z_g^2(A),$$

$$B_{32N} = (\boldsymbol{a}^\top \boldsymbol{V}_N\boldsymbol{a})^{-1}\frac{1}{G}\sum_{g=1}^{G}\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1}\frac{\sqrt{G}}{N}\boldsymbol{X}_g^\top \boldsymbol{u}_g\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1}\frac{\sqrt{G}}{N}\boldsymbol{X}_g^\top \boldsymbol{X}_g(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \frac{1}{G}\sum_{g=1}^{G}W_gZ_g(A),$$

$$B_{33N} = (\boldsymbol{a}^\top \boldsymbol{V}_N\boldsymbol{a})^{-1}\frac{1}{G}\sum_{g=1}^{G}\Big(\Big(\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1}\frac{\sqrt{G}}{N}\boldsymbol{X}_g^\top \boldsymbol{u}_g\Big)^2 - \mathrm{E}\Big(\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1}\frac{\sqrt{G}}{N}\boldsymbol{X}_g^\top \boldsymbol{u}_g\Big)^2\Big) = \frac{1}{G}\sum_{g=1}^{G}(W_g^2 - \mathrm{E}(W_g^2)).$$

It follows directly from Lemma A.3 that $\mathrm{E}|B_{31N}| = O(G^{-1})$, so that $B_{31N} = O_P(G^{-1})$. For B_{32N} we find, using (A.2) and Lemma A.3, that

$$\mathrm{E}(B_{32N}^2) = G^{-4}\sum_{g_1,\ldots,g_4}^{G}\mathrm{E}(W_{g_1}W_{g_2}V_{g_1g_3}(A)V_{g_2g_4}(A)) = O(G^{-2}),$$

because the subscripts $g_1, \ldots, g_4$ must be equal at least in pairs. This implies that $B_{32N} = O_P(G^{-1})$. Combining (B.37), (B.38), and the bounds on B_{31N}, B_{32N}, we have shown that $B_{3N} = -(3/2)\gamma_N B_{33N} + O_P(G^{-1})$. Finally, $W_g^2 - \mathrm{E}(W_g^2)$ is an independent, mean-zero sequence with finite second moment by Lemma A.3, such that Lemma A.1 implies that $B_{33N} = O_P(G^{-1/2})$. Thus, we will apply the Lyapunov Central Limit Theorem to $z_{3g} = -(3/2)\gamma_N G^{-1/2}(W_g^2 - \mathrm{E}(W_g^2))$ jointly with other terms below.

For B_{4N} we find, by the same analysis as for B_{3N} and using the above results, that

$$\begin{aligned} B_{4N} = & -\frac{3}{2}(\boldsymbol{a}^\top \boldsymbol{V}_N \boldsymbol{a})^{-5/2} \boldsymbol{a}^\top (\ddot{\boldsymbol{V}} - \boldsymbol{V}_N)\boldsymbol{a}(1 + O_P(G^{-1})) \\ & \times \frac{1}{G}\sum_{g=1}^{G}\left(\left(\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1}\frac{G}{N}\boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g\right)^3 - \left(\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1}\frac{G}{N}\boldsymbol{X}_g^\top \boldsymbol{u}_g\right)^3\right) \\ = & -\frac{3}{2}(B_{31N} - 2B_{32N} + B_{33N})(1 + O_P(G^{-1}))B_{2N} = O_P(G^{-1}). \end{aligned}$$

Finally, collecting the above results we have shown that

$$G^{1/2}(B_{1N} + B_{2N} + B_{3N} + B_{4N}) = \sum_{g=1}^{G} z_g + O_P(G^{-1/2}),$$

where $z_g = z_{1g} + z_{2g} + z_{3g}$ with

$$z_{1g} = G^{-1/2}(W_g^3 - \mathrm{E}(W_g^3)), \quad z_{2g} = -3G^{-1/2}\sum_{h=1}^{G}(\mathrm{E}W_h^2)V_{hg}(A), \quad z_{3g} = -\frac{3}{2}\gamma_N G^{-1/2}(W_g^2 - \mathrm{E}(W_g^2)),$$

see (A.2). We apply the Lyapunov Central Limit Theorem to $\sum_{g=1}^{G} z_g$. Clearly, z_g depends only on $\boldsymbol{u}_g$ and is independent across g with zero mean and variance, apart from smaller-order terms,

$$\begin{aligned} \omega_N^2(A) = \sum_{g=1}^{G}\mathrm{E}(z_g^2) = & G^{-1}\sum_{g=1}^{G}\left(\mathrm{E}(W_g^6) - (\mathrm{E}(W_g^3))^2\right) + 9G^{-1}\sum_{g_1,g_2=1}^{G}(\mathrm{E}W_{g_1}^2)(\mathrm{E}W_{g_2}^2)(\mathrm{E}Z_{g_1}(A)Z_{g_2}(A)) \\ & + \frac{9}{4}\gamma_N^2 G^{-1}\sum_{g=1}^{G}\left(\mathrm{E}(W_g^4) - (\mathrm{E}(W_g^2))^2\right) - 6G^{-1}\sum_{g_1,g_2=1}^{G}(\mathrm{E}W_{g_1}^3 Z_{g_2}(A))(\mathrm{E}W_{g_2}^2) \\ & - 3\gamma_N G^{-1}\sum_{g=1}^{G}\left(\mathrm{E}(W_g^5) - \mathrm{E}(W_g^3)\mathrm{E}(W_g^2)\right) + 9\gamma_N G^{-1}\sum_{g_1,g_2=1}^{G}(\mathrm{E}W_{g_1}^2 Z_{g_2}(A))(\mathrm{E}W_{g_2}^2), \quad \text{(B.39)} \end{aligned}$$

which is finite by Lemma A.3 because Assumption 6 is satisfied with $\lambda = 2$. To verify Lyapunov's condition we find, using the c_r-inequality, that $\sum_{g=1}^{G}\mathrm{E}|z_g|^{2+\delta} \leq 3^{1+\delta}\sum_{j=1}^{3}\sum_{g=1}^{G}\mathrm{E}|z_{jg}|^{2+\delta}$. Here, using again the c_r-inequality,

$$\sum_{g=1}^{G}\mathrm{E}|z_{1g}|^{2+\delta} \leq 2^{1+\delta}G^{-1-\delta/2}\sum_{g=1}^{G}\mathrm{E}|W_g|^{6+3\delta} + 2^{1+\delta}G^{-1-\delta/2}\sum_{g=1}^{G}|\mathrm{E}(W_g^3)|^{2+\delta} \to 0,$$

by Lemma A.3 choosing $0 < \delta < (\lambda - 2)/3$, which is possible because for this result we have assumed $\lambda > 2$. By an identical argument, $\sum_{g=1}^{G}\mathrm{E}|z_{jg}|^{2+\delta} \to 0$ for $j = 2, 3$, and it follows that $\omega_N^{-1}\sum_{g=1}^{G} z_g \xrightarrow{d} \mathrm{N}(0,1)$.

B.6 Proof of Theorem 5.3

First, as in the proof of Theorem 5.2, we find that

$$\ddot{\xi}_N = \mathrm{E}^*(v^{*4})\frac{1}{G}\sum_{g=1}^{G}(\boldsymbol{a}^\top \ddot{\boldsymbol{V}}\boldsymbol{a})^{-2}\left(\boldsymbol{a}^\top \boldsymbol{Q}_N^{-1}\frac{\sqrt{G}}{N}\boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g\right)^4 = \mathrm{E}^*(v^{*4})(\xi_N + C_{1N} + C_{2N} + C_{3N} + C_{4N}),$$

where C_{iN}, for $i = 1, \ldots, 4$, are given by the same expressions as B_{iN}, for $i = 1, \ldots, 4$, replacing the powers $-3/2$ and 3 in B_{iN} by -2 and 4, respectively. Consequently, the proofs that $C_{iN} = o_P(1)$,

for $i = 1, \ldots, 4$, are nearly identical to those for the corresponding B_{iN} in the proof of Theorem 5.2, although the proofs here are simpler because only $o_P(1)$ is needed, and not a more refined limit as in Theorem 5.2. Hence, the proofs for C_{iN}, for $i = 1, \ldots, 4$, are omitted.

Next, using $\mathrm{E}^*(v_g^{*2}) = 1$ for all g, the proofs for $\ddot{\tau}_{jN} - \tau_{jN}$, for $j = 1, 2, 3$, follow in exactly the same way, but are simpler because fewer moments are involved and only $o_P(1)$ is needed. We therefore omit these proofs.

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*, 698–709.

Davidson, J., A. Monticini, and D. Peel (2007). Implementing the wild bootstrap using a two-point distribution. *Economics Letters 96*, 309–315.

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.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag.

Hansen, B. E. and S. Lee (2017). Asymptotic theory for clustered samples. Working paper, University of Wisconsin, Madison.

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.

Kline, P. and A. Santos (2012). Higher order properties of the wild bootstrap under misspecification. *Journal of Econometrics 171*, 54–70.

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

Liu, R. Y. (1988). Bootstrap procedures under some non-I.I.D. models. *Annals of Statistics 16*, 1696–1708.

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*, 649–665.

MacKinnon, J. G. and M. D. Webb (2017a). Pitfalls when estimating treatment effects using clustered data. *The Political Methodologist 24*, 20–31.

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 M. D. Webb (2018). The wild bootstrap for few (treated) clusters. *Econometrics Journal 21*, to appear.

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

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Annals of Statistics 21*, 255–285.

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

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Annals of Statistics 14*, 1261–1295.