Queen's Economics Department Working Paper No. 1399

Asymptotic Theory and Wild Bootstrap Inference with Clustered Errors

Antoine A. Djogbenou
York University

James G. MacKinnon
Queen's University

Morten Ørregaard Nielsen
Queen's University

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

4-2019

Asymptotic Theory and Wild Bootstrap Inference with Clustered Errors*

Antoine A. Djogbenou
York University
daa@yorku.ca

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

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

April 18, 2019

Abstract

We study inference based on cluster-robust variance estimators for regression models with clustered errors, focusing on the wild cluster bootstrap. We state conditions under which asymptotic and bootstrap tests and confidence intervals are asymptotically valid. These conditions put limits on the rates at which the cluster sizes can increase as the number of clusters tends to infinity. We also derive Edgeworth expansions for the asymptotic and bootstrap test statistics. Simulation experiments illustrate the theoretical results and suggest that alternative variants of the wild cluster bootstrap may perform quite differently. The Edgeworth expansions explain the overrejection of asymptotic tests and shed light on the choice of auxiliary distribution and whether to use restricted or unrestricted estimates in the bootstrap data-generating process.

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

JEL Codes: C15, C21, C23.

*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, Mikkel Sølvsten, and seminar participants at NY Camp Econometrics XII, 2017 CEA Annual Meeting, 2017 CESG Annual Meeting, 2017 Midwest Econometrics Meeting, 2018 Society of Labor Economics Meeting, 2018 Canadian Economics Association Meeting, 2018 Société Canadienne de Science Economique Meeting, 2018 Statistical Society of Canada Meeting, U.C. San Diego, U. Wisconsin-Madison, Binghamton U., Australian National U., and McGill University 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.

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). 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 with critical values from a t distribution. Although it typically does provide more accurate inferences, it too can fail in certain cases; see MacKinnon and Webb (2017a). 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-term and two-term Edgeworth expansions under somewhat stronger assumptions. These expansions explain the overrejection of the asymptotic test found in simulations. The expansions are also used to discuss the choice of auxiliary distribution, whether to use restricted or unrestricted residuals in the bootstrap DGP, and the conditions under which the wild cluster bootstrap may provide an asymptotic refinement.

Conditions for asymptotic validity of CRVE-based inference are given by White (1984, Chap. 6), Liang and Zeger (1986), Hansen (2007), Carter, Schnepel, and Steigerwald (2017), and Hansen and Lee (2019), among others. All but the last two of these works assume that clusters are equal-sized. Carter, Schnepel, and Steigerwald (2017) considers linear regression with a cluster structure and studies the effects of heterogeneity across clusters, but under stronger assumptions than ours. Hansen and Lee (2019) derives a law of large numbers and a central limit theorem for clustered samples under conditions that are very similar to ours and applies the results to several different estimation problems, including regression. However, we are not aware of any previous work on the asymptotic validity of wild bootstrap methods for clustered errors.

An alternative to the wild cluster bootstrap is the pairs cluster bootstrap, in which the bootstrap samples are constructed by resampling the regressors and regressand together at the cluster level. 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 (2017b). 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 the regressor matrix, 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 regressor matrix not to have full column rank.

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, and in Section 5, we discuss higher-order asymptotic theory. Based on all our results, we give some general practical guidance in Section 6, and Section 7 concludes. Appendices A, B, and C, provided in the supplementary material, contain preliminary lemmas, proofs of main results, and additional simulation evidence, respectively.

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. It is assumed that the $k \times 1$ score vectors, $\boldsymbol{s}_g = \boldsymbol{X}_g^\top \boldsymbol{u}_g$, satisfy $\mathrm{E}(\boldsymbol{s}_g) = \boldsymbol{0}$ for all g and

$$\mathrm{E}(\boldsymbol{s}_g \boldsymbol{s}_h^\top) = \mathbb{I}(g = h)\boldsymbol{\Sigma}_g, \quad g, h = 1, \ldots, G, \tag{2}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function and $\boldsymbol{\Sigma}_g$ is a $k \times k$ matrix. In cross-sectional regressions, it is common to make the stronger assumption that the regressors are exogenous. Under such a condition, it would be assumed that $\mathrm{E}(\boldsymbol{u}_g | \boldsymbol{X}) = \boldsymbol{0}$ and $\mathrm{E}(\boldsymbol{u}_g \boldsymbol{u}_h^\top | \boldsymbol{X}) = \mathbb{I}(g = h)\boldsymbol{\Omega}_g$ for all g, h, where $\boldsymbol{\Omega}_g$ is an $N_g \times N_g$ variance matrix. However, this condition is not necessary for the first-order asymptotic analysis, so we maintain only the weaker assumption in (2).

When $N_g = 1$ for all g, the model (1)-(2) reduces to the well-known linear regression model with heteroskedasticity of unknown form and predetermined regressors. Hence, as a special case, our results cover that model as well.

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

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

Let $\boldsymbol{Q} = N^{-1} \boldsymbol{X}^\top \boldsymbol{X}$ and $\boldsymbol{\Gamma} = N^{-2} \sum_{g=1}^{G} \boldsymbol{\Sigma}_g$. The usual sandwich form of the variance matrix of $\hat{\boldsymbol{\beta}}$ is given by

$$\boldsymbol{V} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \Big(\sum_{g=1}^{G} \boldsymbol{\Sigma}_g \Big) (\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \boldsymbol{Q}^{-1} \boldsymbol{\Gamma} \boldsymbol{Q}^{-1}. \tag{4}$$

We note that, under the assumption that the regressors are exogenous, $\boldsymbol{V}$ is the conditional variance of $\hat{\boldsymbol{\beta}}$ given $\boldsymbol{X}$. We now define the cluster-robust estimator of $\boldsymbol{V}$, i.e. the CRVE, as

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

where $\hat{\boldsymbol{\Gamma}} = N^{-2}\sum_{g=1}^{G}\hat{\boldsymbol{s}}_g\hat{\boldsymbol{s}}_g^\top = N^{-2}\sum_{g=1}^{G}\boldsymbol{X}_g^\top\hat{\boldsymbol{u}}_g\hat{\boldsymbol{u}}_g^\top\boldsymbol{X}_g$, using $\hat{\boldsymbol{s}}_g = \boldsymbol{X}_g^\top\hat{\boldsymbol{u}}_g$. The multiplicative factor d in (5) is a finite-sample correction that depends on N and/or G, and it is assumed throughout that $d \to 1$ as $G \to \infty$. For example, Hansen (2007) suggests using $d = G/(G-1)$, and Stata has implemented $d = G(N-1)/((G-1)(N-k))$ as the default.

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.

We let $\boldsymbol{\beta}_0$ denote the true value of $\boldsymbol{\beta}$. For concreteness and simplicity, we 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. Of course, it would not be difficult to extend our results to Wald tests of $r \geq 1$ linear restrictions.

2.1 Assumptions

To obtain the asymptotic limit theory for t_a, we need the following conditions, where, for any matrix $\boldsymbol{M}$, $\|\boldsymbol{M}\| = (\mathrm{Tr}(\boldsymbol{M}^\top\boldsymbol{M}))^{1/2}$ denotes the Euclidean norm.

Assumption 1. The sequence $\{\boldsymbol{s}_g\}$ is independent across g and satisfies, for all $g \in \mathbb{N}$, that $\mathrm{E}(\boldsymbol{s}_g) = \boldsymbol{0}$ and $\mathrm{E}(\boldsymbol{s}_g\boldsymbol{s}_g^\top) = \boldsymbol{\Sigma}_g$. In addition, for some $\lambda > 0$,

$$\sup_{i,g\in\mathbb{N}} \mathrm{E}\|\boldsymbol{s}_{ig}\|^{2+\lambda} < \infty,$$

where $\boldsymbol{s}_{ig} = \boldsymbol{X}_{ig}^\top u_{ig}$ denotes the score contribution of the i^{th} observation within cluster g, while $\boldsymbol{X}_{ig}$ and u_{ig} denote the i^{th} rows of $\boldsymbol{X}_g$ and $\boldsymbol{u}_g$, respectively.

Assumption 2. The regressor matrix $\boldsymbol{X}$ satisfies $\boldsymbol{Q} \xrightarrow{P} \boldsymbol{\Xi}_0$, where $\boldsymbol{\Xi}_0$ is finite and positive definite, and

$$\sup_{i,g\in\mathbb{N}} \mathrm{E}\|\boldsymbol{X}_{ig}^\top\boldsymbol{X}_{ig}\|^{2+\lambda} < \infty,$$

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

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

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

Assumption 1 imposes the conditions that $\{\boldsymbol{s}_g\} = \{\boldsymbol{X}_g^\top\boldsymbol{u}_g\}$ is independent across clusters, with finite $2+\lambda$ moments, and that $\boldsymbol{s}_g$ has zero mean and constant, but possibly heterogeneous, 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 variance of $\boldsymbol{a}^\top\hat{\boldsymbol{\beta}}$, as measured by $\boldsymbol{a}^\top\boldsymbol{V}\boldsymbol{a}$ with $\boldsymbol{V}$ defined in (4), multiplied by a non-random sequence $\{\mu_N\}$, converges to a finite, non-zero limit. Thus, μ_N can be interpreted as the rate at which information accumulates. 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 $\boldsymbol{a}^\top\boldsymbol{V}\boldsymbol{a}$ as represented by (the inverse of) the sequence μ_N, and to the moment condition in Assumption 1. Thus, Assumption 3 represents a trade-off between the extent of heterogeneity of cluster sizes allowed and the number of moments assumed to exist.

To analyze the role of μ_N, we investigate two extreme cases, with all other cases lying in between: (i) $\boldsymbol{s}_{ig}$ is uncorrelated across i and (ii) $\boldsymbol{s}_{ig}$ is correlated with $\boldsymbol{s}_{jg}$ for all i, j. Case (i) would be obtained, for example, if the regressors were exogenous and the errors uncorrelated. In case (i), it straightforwardly holds that $\boldsymbol{\Sigma}_g = \sum_{i=1}^{N_g} \mathrm{E}(\boldsymbol{s}_{ig}\boldsymbol{s}_{ig}^\top) = O(N_g)$ such that

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

Thus, in particular, $\hat{\boldsymbol{\beta}}$ converges at rate $O_P(N^{-1/2})$ in this case. On the other hand, in case (ii) we find that

$$\boldsymbol{\Sigma}_g = \mathrm{E}\Big(\sum_{i,j=1}^{N_g} \boldsymbol{s}_{ig}\boldsymbol{s}_{jg}^\top\Big) = O(N_g^2), \tag{8}$$

and it follows that

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

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), 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. For example, if four moments are assumed to exist (i.e. $\lambda = 2$ in Assumption 1), as in much related work, then 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,

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

The exponent in (12) is increasing in λ, and with four moments, for example, 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., $N_g = c_g N^{\alpha}$, 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 4 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 when four moments exist. However, in case (ii) above, where $\boldsymbol{\Sigma}_g = O(N_g^2)$, 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.

2.2 Including Fixed Effects

If the model includes cluster fixed effects, the matrix $\hat{\boldsymbol{V}}$ defined in (5) is singular. The rank of $\hat{\boldsymbol{V}}$ cannot exceed G, but the number of parameters must be greater than G whenever there are G fixed effects (and perhaps in other circumstances as well). With cluster fixed effects, 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 $\boldsymbol{V}$ to have zero diagonal elements for the coefficients of the fixed effects.

The presence of cluster fixed effects does not prevent us from using (5) to make inferences about the remaining elements of $\boldsymbol{\beta}$. Instead of doing so directly, it is both computationally faster and theoretically simpler to project the regressand and all the other regressors off the fixed effects so that all variables are expressed as deviations from cluster means; see Pustejovsky and Tipton (2018). We need to change Assumptions 1–3 slightly in order to allow for this.

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 $\boldsymbol{D}_g$ vertically. 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. The g^{th} diagonal block of $\boldsymbol{M_D}$ is the $N_g \times N_g$ matrix $\boldsymbol{M}_g = \mathbf{I} - \boldsymbol{\iota}(\boldsymbol{\iota}^\top\boldsymbol{\iota})^{-1}\boldsymbol{\iota}^\top$, where $\boldsymbol{\iota}$ is a vector of N_g ones. In model (1), we can replace $(\boldsymbol{y}, \boldsymbol{X})$ by $(\boldsymbol{M_D y}, \boldsymbol{M_D X})$ so as to partial out the fixed effects. For cluster g, this means replacing $(\boldsymbol{y}_g, \boldsymbol{X}_g)$ by $(\boldsymbol{M}_g\boldsymbol{y}_g, \boldsymbol{M}_g\boldsymbol{X}_g)$. Of course, demeaning the data within each cluster necessarily affects the pattern of intra-cluster correlation, introducing correlations even where there were none originally.

The assumptions for the model involving the data demeaned within each cluster are as follows, where $\check{\boldsymbol{s}}_g = \boldsymbol{X}_g^\top \boldsymbol{M}_g \boldsymbol{u}_g$ denotes the scores of the model for the demeaned data.

Assumption 1′. The sequence $\{\check{\boldsymbol{s}}_g\}$ is independent across g and satisfies, for all $g \in \mathbb{N}$, that $\mathrm{E}(\check{\boldsymbol{s}}_g) = \boldsymbol{0}$ and $\mathrm{E}(\check{\boldsymbol{s}}_g \check{\boldsymbol{s}}_g^\top) = \check{\boldsymbol{\Sigma}}_g$. In addition, for some $\lambda > 0$,

$$\sup_{i,g \in \mathbb{N}} \mathrm{E}\|\check{\boldsymbol{s}}_{ig}\|^{2+\lambda} < \infty.$$

Assumption 2′. The regressor matrix $\boldsymbol{X}$ satisfies $N^{-1}\boldsymbol{X}^\top \boldsymbol{M_D} \boldsymbol{X} \xrightarrow{P} \boldsymbol{\Xi}_M$, where $\boldsymbol{\Xi}_M$ is finite and positive definite, and

$$\sup_{i,g \in \mathbb{N}} \mathrm{E}\|(\boldsymbol{X}_{ig} - \bar{\boldsymbol{X}}_g)^\top (\boldsymbol{X}_{ig} - \bar{\boldsymbol{X}}_g)\|^{2+\lambda} < \infty,$$

where $\bar{\boldsymbol{X}}_g$ is the mean of the $\boldsymbol{X}_{ig}$ over cluster g and $\lambda > 0$ is defined in Assumption 1′. Furthermore, there exist a non-random sequence $\{\check{\mu}_N\}$ and a non-random, finite scalar $\check{v}_a > 0$ such that $\check{\mu}_N \to \infty$ and $\check{\mu}_N \boldsymbol{a}^\top \check{\boldsymbol{V}} \boldsymbol{a} \xrightarrow{P} \check{v}_a$, where

$$\check{\boldsymbol{V}} = d(\boldsymbol{X}^\top \boldsymbol{M_D} \boldsymbol{X})^{-1} \Big(\sum_{g=1}^{G} \check{\boldsymbol{\Sigma}}_g \Big) (\boldsymbol{X}^\top \boldsymbol{M_D} \boldsymbol{X})^{-1}. \tag{14}$$

Assumption 3′. For λ defined in Assumption 1′ and $\check{\mu}_N$ defined in Assumption 2′,

$$G \to \infty \quad \text{and} \quad \check{\mu}_N^{\frac{2+\lambda}{2+2\lambda}} \sup_{g \in \mathbb{N}} \frac{N_g}{N} \to 0.$$

To a great extent, Assumptions 1′–3′ are in fact implied by Assumptions 1–3. For example, because $\check{\boldsymbol{s}}_g = (\boldsymbol{X}_g^\top \otimes \boldsymbol{u}_g^\top) \operatorname{vec}(\boldsymbol{M}_g)$, the independence condition in Assumption 1′ is implied by independence of $\{\boldsymbol{X}_g \otimes \boldsymbol{u}_g\}$ across g. Moreover, it is trivial to see that $\mathrm{E}(\check{\boldsymbol{s}}_g) = \boldsymbol{0}$ is implied by Assumption 1, while the moment condition in Assumption 1′ is implied (using the c_r inequality) by the simple additional condition that $\sup_{i,j,g \in \mathbb{N}} \mathrm{E}\|\boldsymbol{X}_{jg} u_{ig}\|^{2+\lambda} < \infty$. Similarly, the moment condition in Assumption 2′ is trivially implied by the moment condition in Assumption 2. Thus, together with Assumptions 1–3, the following assumption is sufficient for the model with cluster fixed effects.

Assumption FE.

1. The sequence $\{\boldsymbol{X}_g \otimes \boldsymbol{u}_g\}$ is independent across g and $\sup_{i,j,g \in \mathbb{N}} \mathrm{E}\|\boldsymbol{X}_{jg} u_{ig}\|^{2+\lambda} < \infty$.
2. The regressor matrix $\boldsymbol{X}$ satisfies $N^{-1}\boldsymbol{X}^\top \boldsymbol{M_D} \boldsymbol{X} \xrightarrow{P} \boldsymbol{\Xi}_M$, where $\boldsymbol{\Xi}_M$ is finite and positive definite, and there exist a non-random sequence $\{\check{\mu}_N\}$ and a non-random, finite scalar $\check{v}_a > 0$ such that $\check{\mu}_N \to \infty$ and $\check{\mu}_N \boldsymbol{a}^\top \check{\boldsymbol{V}} \boldsymbol{a} \xrightarrow{P} \check{v}_a$.
3. Assumption 3 holds with μ_N replaced by $\check{\mu}_N$ defined in part 2.

The first two parts of Assumption FE strengthen Assumptions 1 and 2, but they are quite mild. The key thing they rule out is regressors that only vary at the cluster level, because $\boldsymbol{M_D} \boldsymbol{x} = \boldsymbol{0}$ if $\boldsymbol{x}$ is such a regressor, and this would violate part 2 of Assumption FE. Whenever a model involves cluster fixed effects, we will assume that Assumptions 1′–3′ hold. As argued above, an alternative is to assume that, in addition to Assumptions 1–3 for $\boldsymbol{y}$ and $\boldsymbol{X}$, the conditions in Assumption FE hold.

Of course, the variance matrix (14) for the model with fixed effects will be numerically different from the variance matrix (5) for the model without fixed effects. The former will often yield larger standard errors, but not always. The fact that $\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{X}^\top \boldsymbol{M_D} \boldsymbol{X}$ is positive semidefinite would make the diagonal elements of (14) larger than those of (5), but this can be offset by the fact that $\sum_{g=1}^{G} \check{\boldsymbol{\Sigma}}_g$ will generally be smaller than $\sum_{g=1}^{G} \boldsymbol{\Sigma}_g$, because projecting the error terms off the fixed effects reduces the amount of intra-cluster correlation.

2.3 Theory for Asymptotic Tests

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, Chap. 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, that paper's Assumption 1.b. requires that the fourth-order intra-cluster cross-moments of $\boldsymbol{u}_g$, i.e. $\mathrm{E}(u_{ig}u_{jg}u_{kg}u_{lg})$ with i,j,k,l not all the same, are identical to those of a multivariate normal distribution. It seems likely that this condition rules out a great deal of empirical data in economics and related disciplines. Moreover, that paper imposes very high-level assumptions to restrict cluster-size heterogeneity, and 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, and we only assume existence of $2+\lambda$ moments for some $\lambda > 0$. Very recently (indeed after the first draft of the present paper was written), Hansen and Lee (2019) derives a law of large numbers and a central limit theorem for clustered samples under conditions very similar to ours. The paper applies those results to several different estimation problems, including regression, but does 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, $\boldsymbol{\Sigma}_g$ or $\boldsymbol{\Omega}_g$, 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 $(\boldsymbol{a}^\top \boldsymbol{V}\boldsymbol{a})^{-1/2}\boldsymbol{a}^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, $(\boldsymbol{a}^\top \boldsymbol{V}\boldsymbol{a})^{-1}\boldsymbol{a}^\top \hat{\boldsymbol{V}}\boldsymbol{a}$, and t_a. See, e.g., Hansen (2007, Theorem 2) or Carter, Schnepel, and Steigerwald (2017) for related arguments.

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

$$\boldsymbol{a}^\top(\boldsymbol{\beta}_N - \boldsymbol{\beta}_0) = (\boldsymbol{a}^\top \boldsymbol{V}\boldsymbol{a})^{1/2}\delta, \tag{15}$$

which is often referred to as "Pitman drift." Under (15), the DGP is characterized by a drifting sequence of true values of the parameter vector $\boldsymbol{\beta}$ indexed by N 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 (15). It then holds that*

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

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

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

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 (16), which is $(\boldsymbol{a}^\top \boldsymbol{V}\boldsymbol{a})^{1/2}$, 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 $z_g = v_a^{-1/2}\mu_N^{1/2}N^{-1}\boldsymbol{a}^\top\boldsymbol{Q}^{-1}\boldsymbol{s}_g$, 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 (18), 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 (15) is given by the asymptotic local power function

$$1 - \Phi(z_{1-\alpha/2} - \delta) + \Phi(-z_{1-\alpha/2} - \delta), \tag{19}$$

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 (19) 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, the $\boldsymbol{\Sigma}_g$, and $\boldsymbol{X}$ via the quantity $(\boldsymbol{a}^\top\boldsymbol{V}\boldsymbol{a})^{1/2}$ that appears in (15). The interpretation of δ implicitly changes whenever $(\boldsymbol{a}^\top\boldsymbol{V}\boldsymbol{a})^{1/2}$ changes.

Recalling the definition of $\boldsymbol{V}$ 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 of the scores, i.e. $\boldsymbol{\Sigma}_g$, and is also confounded with the influence of the regressors $\boldsymbol{X}$. In general, the effects of the N_g, the $\boldsymbol{\Sigma}_g$, 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 $\hat{\boldsymbol{\beta}}$ defined in (3), unrestricted residuals $\hat{\boldsymbol{u}}$, and $\hat{\boldsymbol{V}}$ defined in (5). 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 the v_g^{*b} are 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
   $$\hat{\boldsymbol{V}}^{*b} = d(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}_g^\top\hat{\boldsymbol{u}}_g^{*b}\hat{\boldsymbol{u}}_g^{*b\top}\boldsymbol{X}_g\right)(\boldsymbol{X}^\top\boldsymbol{X})^{-1};$$
   (d) calculate the bootstrap t-statistic
   $$t_a^{*b} = \frac{\boldsymbol{a}^\top(\hat{\boldsymbol{\beta}}^{*b} - \ddot{\boldsymbol{\beta}})}{\sqrt{\boldsymbol{a}^\top\hat{\boldsymbol{V}}^{*b}\boldsymbol{a}}}.$$
4. Depending on whether the alternative hypothesis is H_L: $\boldsymbol{a}^\top\boldsymbol{\beta} < \boldsymbol{a}^\top\boldsymbol{\beta}_0$, H_R: $\boldsymbol{a}^\top\boldsymbol{\beta} > \boldsymbol{a}^\top\boldsymbol{\beta}_0$, or H_2: $\boldsymbol{a}^\top\boldsymbol{\beta} \neq \boldsymbol{a}^\top\boldsymbol{\beta}_0$, compute one of the following bootstrap P values:
   $$\hat{P}_L^* = \frac{1}{B}\sum_{b=1}^{B}\mathbb{I}(t_a^{*b} < t_a), \quad \hat{P}_R^* = \frac{1}{B}\sum_{b=1}^{B}\mathbb{I}(t_a^{*b} > t_a), \quad \text{or} \quad \hat{P}_S^* = \frac{1}{B}\sum_{b=1}^{B}\mathbb{I}(|t_a^{*b}| > |t_a|).$$
   If the alternative hypothesis is H_2, then the symmetric P value $\hat{P}_S^*$ could be replaced by the equal-tail P value, which is simply $2\min(\hat{P}_L^*, \hat{P}_R^*)$.

We note that all the bootstrap P values defined in step 4 of the wild cluster bootstrap algorithm are exactly invariant to the multiplicative factor d in the CRVE, as long as the same factor is used for the original test statistic and the bootstrap test statistics. The factor d scales both t_a and t_a^{*b} by the same amount, leaving the indicator functions in step 4 unchanged.

Our next result demonstrates the validity of the WCB. Let the cumulative distribution function (CDF) of t_a under 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^* denote the corresponding expectation conditional on a given sample.

[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. However, when an efficient algorithm for computing WCR bootstrap P values is used, it is also easy to construct confidence intervals by inverting WCR bootstrap tests; see Roodman, MacKinnon, Nielsen, and Webb (2019).

Theorem 3.1. *Suppose Assumptions 1–3 are satisfied and* $\mathrm{E}^*|v^*|^{2+\lambda} < \infty$ *for some* $\lambda > 0$*, and that the true value of* $\boldsymbol{\beta}$ *is given by (15). 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 H_0 is true, that is, when $\delta = 0$ in (15), 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 (15), the bootstrap distribution $P^*(t_a^* \leq x)$ coincides with that of the original t-statistic under the null hypothesis 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 (19) 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 $\hat{\boldsymbol{V}}^*$. Also note that, as with the WCB algorithm, the WB algorithm is exactly invariant to the scaling factor d in the CRVE.

Theorem 3.2. *Suppose Assumptions 1–3 are satisfied and* $\mathrm{E}^*|v^*|^{2+\lambda} < \infty$ *for some* $\lambda > 0$*, and that the true value of* $\boldsymbol{\beta}$ *is given by (15). 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 (15), 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 $\boldsymbol{\Sigma}_g$ or $\boldsymbol{\Omega}_g$ 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 $\hat{\boldsymbol{\beta}}^*$ has a different asymptotic variance matrix (conditional on the original sample) than do both the actual estimator $\hat{\boldsymbol{\beta}}$ and the WCB estimator (conditional on the original sample); cf. (16) 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 $\boldsymbol{\Omega}_g$. We study the finite-sample performance of the WB and the WCB in the next section.

4 Simulation Experiments

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

Our experiments are based on the DGP

$$\boldsymbol{y}_g = \beta_1 + \beta_2 \boldsymbol{x}_g + \boldsymbol{u}_g, \quad \mathrm{E}(\boldsymbol{u}_g \boldsymbol{u}_g^\top) = \boldsymbol{\Omega}_g, \quad g = 1, \ldots, G, \tag{20}$$

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 all the experiments that we report, we set $\rho = 0.10$, because rejection frequencies for all versions of the WCB appear to be almost totally insensitive to the value of ρ, provided it is greater than about 0.05. We provide evidence on this point in Appendix C.

The null hypothesis is that $\beta_2 = 0$; this is equivalent to setting $\boldsymbol{a} = [0 \;\; 1]^\top$. To avoid the possibly excessive symmetry of normal errors, the errors are generated by a normal mixture model with skewness 1 and excess kurtosis 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 for u_{ig}, in particular $\rho = 0.10$, we use $p = 0.1967$, $\mu_1 = 0.7693$, $\mu_2 = -0.1884$, $\sigma_1 = 1.5734$, $\sigma_2 = 0.6770$, and $\rho_1 = 0.2556$.

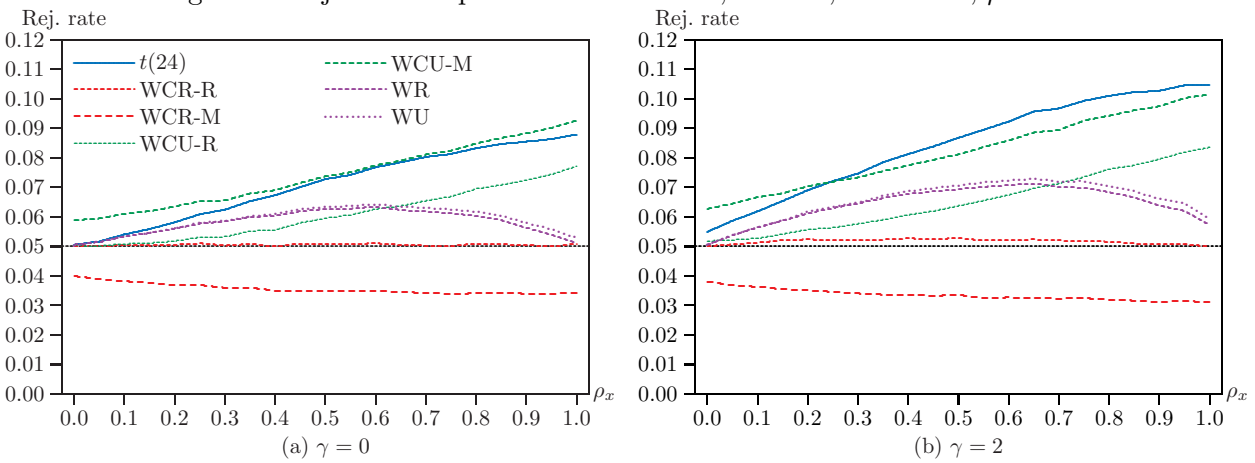
The regressor x_{ig} is a weighted sum of two $\chi^2(8)$ random variables, $x_{1,ig}$ and $x_{2,g}$, both of them recentered and rescaled to have mean 0 and variance 1, where $x_{1,ig}$ is independent across observations, and $x_{2,g}$ varies across but not within clusters. The weights are chosen so that the correlation of x_{ig} with x_{jg} for $i \neq j$ is ρ_x. Specifically, $x_{ig} = (1-\rho_x)^{1/2} x_{1,ig} + \rho_x^{1/2} x_{2,g}$. Thus the regressor is both skewed and leptokurtic, and it can display any amount of intra-cluster correlation. We also experimented with other ways of generating the x_{ig}; see Appendix C. The amount of skewness has very little effect on the rejection frequencies for WCR, but the ones for $t(G-1)$ and WCU do increase noticeably as the regressor becomes more skewed.

Since we have to impose conditions like Assumption 3 on the cluster sizes, we expect inference to be more difficult when cluster sizes differ; see MacKinnon and Webb (2017a) for some evidence on this point. In order to allow cluster sizes to vary systematically, we allocate N observations among G clusters using the equation

$$N_g = \left[\frac{N \exp(\gamma g/G)}{\sum_{j=1}^{G} \exp(\gamma j/G)} \right], \quad \text{for } g = 1, \ldots, G-1, \tag{21}$$

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.

Figure 1: Rejection frequencies at 0.05 level, $G = 25$, $N = 2500$, $\rho = 0.10$



In some of our experiments, $\gamma = 2$. This implies that, when $G = 25$, the largest cluster size is 7.3 times the smallest. As expected, the performance of all methods deteriorates when we increase γ from 0 to 2. Additional evidence on the effect of γ is provided in Appendix C.

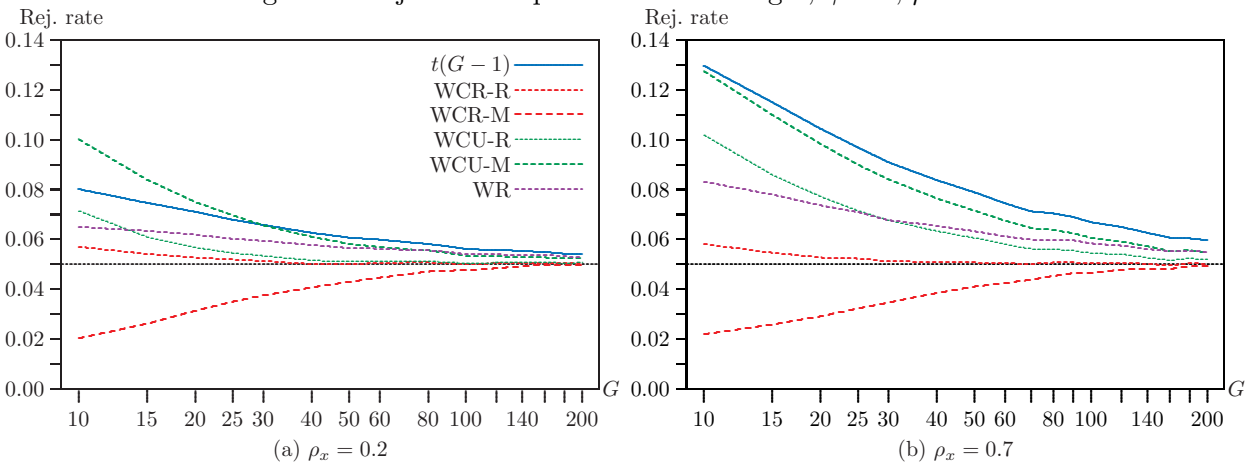
Every experiment has 400,000 replications. We use such a large number because, in many cases, the rejection frequencies for alternative methods are quite similar, and we need a large number to estimate them accurately. Fortunately, a recently developed algorithm for computing wild cluster (but not wild) bootstrap P values makes it remarkably inexpensive to perform such a large number of replications for all the WCB methods; see Roodman, MacKinnon, Nielsen, and Webb (2019).

In the first set of experiments, the model is as described above, with $G = 25$, $N = 2500$, and either $\gamma = 0$ (100 observations per cluster) or $\gamma = 2$ (between 32 and 234 observations per cluster). The two panels of Figure 1 show rejection frequencies for seven tests at the 0.05 level. The horizontal axis shows ρ_x, which varies from 0.0 to 1.0 by increments of 0.05. 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, which in this case is $t(24)$, instead of standard normal ones, as advocated by Bester, Conley, and Hansen (2011), and the CRVE is the one in (5) with the factor $d = G(N-1)/((G-1)(N-k))$. Without this factor, or if we had used the standard normal distribution instead of the $t(G-1)$ distribution, the overrejection by the t-test 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. For the ordinary wild bootstrap (WR and WU), the v^* are drawn from the Rademacher distribution. For the wild cluster bootstrap, they are drawn either from the Rademacher distribution (WCR-R and WCU-R) or from the two-point Mammen (1993) auxiliary distribution (WCR-M and WCU-M).

Both the cluster-robust t-test and the two variants of the WCU bootstrap test always overreject when $\rho_x > 0$, and they do so more severely as ρ_x increases. Interestingly, using the Mammen distribution leads to considerably more severe overrejection than using the Rademacher. In some cases, WCU-M actually rejects more often than $t(24)$. In contrast, the WCR-R bootstrap works very well in all cases, and the WCR-M bootstrap underrejects quite severely. The reasons for the excellent performance of WCR-R, the underrejection by WCR-M, and the overrejection by WCU-R

Figure 2: Rejection frequencies as G changes, $\gamma = 2$, $\rho = 0.10$



and WCU-M are analyzed in Section 5 using higher-order asymptotic theory.

The two ordinary wild bootstrap methods (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. Because WR and WU are always very similar, and are much more expensive to compute than all versions of the WCB, we do not consider WU in later experiments.

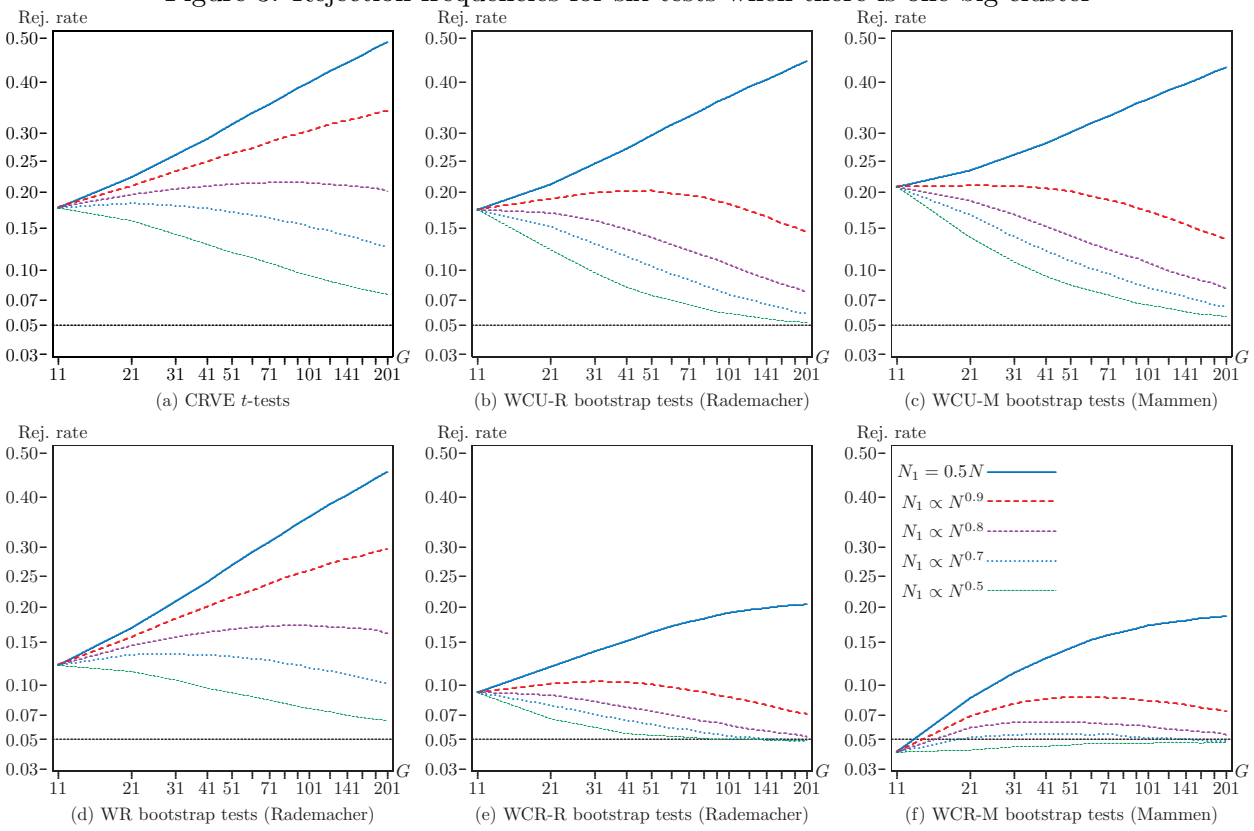
Since our focus is on the bootstrap, the only non-bootstrap procedure for which we report results is the t-test implemented in Stata. Carter, Schnepel, and Steigerwald (2017) has proposed a diagnostic called the "effective number of clusters" and denoted G^{*A}. When it is small, tests based on $t(G-1)$ are prone to overreject. For the experiments of Panel (a) in Figure 1, the average value of G^{*A} declines from 9.55 to 8.51 as ρ_x increases from 0.0 to 1.0. For the experiments of Panel (b), it declines from 8.23 to 6.01. Thus the value of G^{*A} correctly predicts that the t-test will perform better in Panel (a) than in Panel (b) and that its performance will deteriorate as ρ_x increases, more strongly in Panel (b). However, in both panels, the average values of G^{*A} decline sharply between $\rho_x = 0$ and $\rho_x = 0.15$ and very little thereafter. Thus using G^{*A} as a diagnostic fails to predict the very substantial increase in rejection frequencies for t-tests as ρ_x increases beyond 0.15.[2]

In the next two experiments, we vary the number of clusters G and the sample size together. The results are shown in Figure 2. In Panel (a), we fix ρ_x at 0.2, where WCR and WCU work quite well. In Panel (b), we fix it at 0.7, where all procedures except WCR-R work much less well. In both panels, we vary G from 10 to 30 by 5, then from 40 to 90 by 10, and finally from 100 to 200 by 20. The value of γ is 2, so that cluster sizes change as G, and therefore N, increase. However, the way in which they vary is essentially the same as G increases. The smallest sample size is 1000, and the largest is 20,000.

There are four striking results in Figure 2. The first is that all the wild cluster bootstrap tests improve rapidly as G increases. This is true even for WCU-M, which rejects more often than $t(G-1)$ in Panel (a) when G is small. The second is that WCR-R performs very much better than WCR-M and both variants of WCU, and indeed is easily the best performing test in Figure 2. Its rejection frequency is essentially 0.05 for $G \geq 30$ in Panel (a) and for $G \geq 40$ in Panel (b). This is consistent with the results of Davidson and MacKinnon (1999) and Davidson and Flachaire (2008) and reflects the fact that the restricted residuals are better estimators of the error terms

[2]Carter, Schnepel, and Steigerwald (2017, p. 707) discusses the use of critical values based on the $t(G^{*A})$ distribution as an alternative to $t(G-1)$, but advises against this procedure and presents no Monte Carlo evidence.

Figure 3: Rejection frequencies for six tests when there is one big cluster



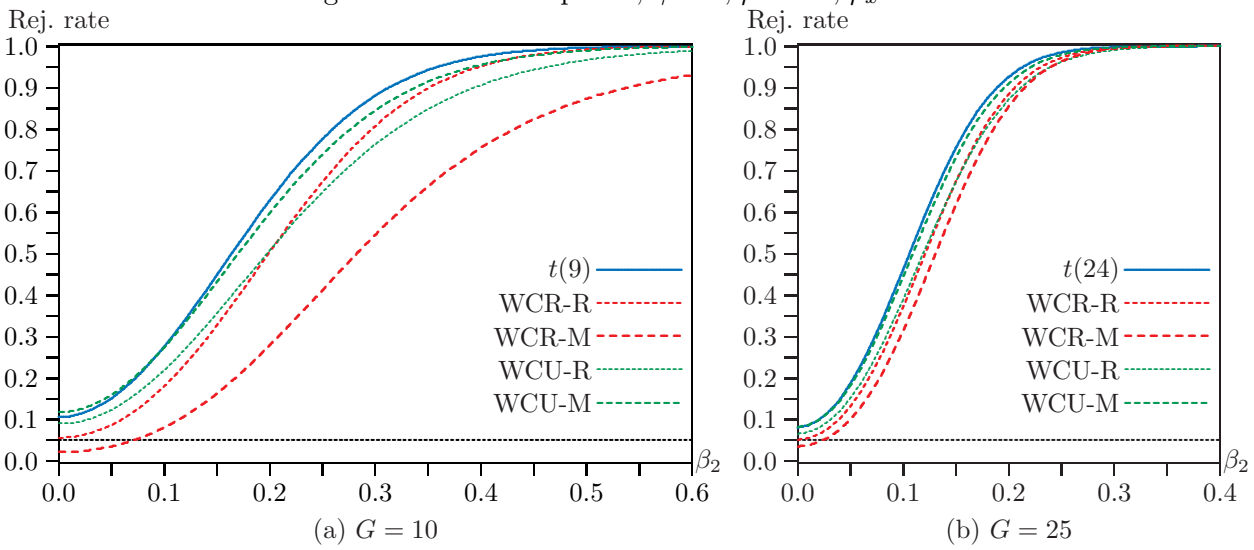
than the unrestricted ones, especially for high-leverage observations where the regressor happens to be particularly large. The third result is that WCR-M underrejects severely when G is small, but the underrejection essentially disappears by the time $G = 200$.

The fourth striking result in Figure 2 is that WR, the restricted ordinary wild bootstrap test, improves less rapidly than all the wild cluster bootstrap tests as G increases. In both panels, it is the second-best test, after WCR-R, when G is small. But it is clearly the worst bootstrap test for $G \geq 90$ in Panel (a), and it is the worst bootstrap test for $G = 200$ in Panel (b). This is exactly what we expect to see based on the discussion following Theorem 3.2.

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. Both the regressor and the error terms are distributed in the same way as in the experiments of Figures 1 and 2, with $\rho = 0.10$ and $\rho_x = 0.70$. We set $N = 200(G-1)$ and $N_1 = 500(N/1000)^\alpha$ 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 $\alpha = 0.9$, $N_1/N = 0.371$ for $G = 201$, and when $\alpha = 0.5$, $N_1/N = 0.112$ for $G = 201$. In contrast, for the experiments of Figure 2, the largest cluster constitutes 21.3% of the sample for $G = 10$ but only 1.6% for $G = 200$.

The six panels of Figure 3 show rejection frequencies for CRVE t-tests, ordinary wild bootstrap tests, and four wild cluster bootstrap tests for five values of α between 0.5 and 1. Since this experimental design violates the rate condition given in Assumption 3 when $\alpha = 1$, it is not

Figure 4: Simulated power, $\gamma = 0$, $\rho = 0.1$, $\rho_x = 0.7$



surprising that the rejection frequency for the CRVE t-test, in Panel (a), increases steadily with G. This is also true when $\alpha = 0.9$. For smaller values of α, rejection frequencies clearly drop as G increases beyond some threshold value, which varies with α. However, even for the smallest values of α, G would evidently have to be very large for t-tests to yield reliable inferences.

Interestingly, the rejection frequencies for the WR bootstrap, in Panel (d), look quite similar to the ones for the t-tests in Panel (a). They are somewhat smaller for each value of G, but for all values of α they vary with G in roughly the same way and always exceed 0.05.

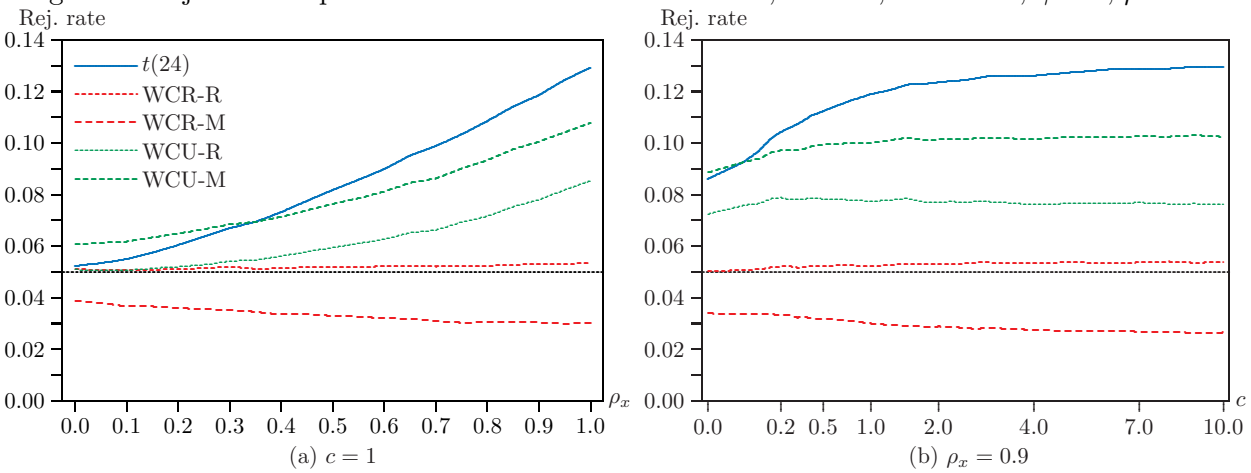
Panels (b) and (c) show rejection frequencies for the WCU-R and WCU-M bootstraps, respectively. For $\alpha = 1$, they look very similar to the ones for the t-test, overrejecting more and more severely as G increases. However, for smaller values of α, they either rise at first and then drop (for $\alpha = 0.9$), or they drop monotonically as G increases. For $G = 201$, both of the unrestricted WCB methods perform very much better than the t-test or the WR bootstrap.

Rejection frequencies for the WCR-R and WCR-M bootstraps, shown in Panels (e) and (f), are much smaller than for any of the other methods. In fact, the WCR-M tests actually underreject for $G = 11$, as they did for all values of G in Figures 1 and 2. However, except for $\alpha = 0.5$, they then overreject for all or most larger values of G. For both variants of the restricted WCB, the rejection frequencies appear to be converging to 0.05 as G becomes large except for $\alpha = 1$ and possibly $\alpha = 0.9$. For $\alpha = 0.5$ and $\alpha = 0.7$, both tests work almost perfectly for large values of G.

Up to this point, we have only studied test size. Figure 4 investigates the power of alternative tests for the case of equal-sized clusters with $N_g = 100$. The horizontal axis shows the true value of β_2 for tests of $\beta_2 = 0$. In Panel (a) there are 10 clusters, and in Panel (b) there are 25. We use a very small number in Panel (a) to make it easy to distinguish the five curves.[3] Results for $\gamma > 0$ (not reported) look similar to those in Figure 4, but, for any sample size and value of β_2, power diminishes as γ increases. This happens because the information content of a sample with clustered error terms (and the same pattern of intra-cluster correlation for all clusters) is maximized when

[3]We did not calculate the power of the WR bootstrap because it would have been very expensive, and because both our theoretical results (see the discussion following Theorem 3.2) and our simulation results (see in particular Figure 2) suggest that it will rarely be the procedure of choice. For the same reasons, we also omit this method in the last two sets of experiments, which we report below.

Figure 5: Rejection frequencies with heteroskedastic errors, $G = 25$, $N = 2500$, $\gamma = 0$, $\rho = 0.10$



all clusters are the same size.

In both panels of Figure 4, using the $t(G-1)$ distribution leads to substantial overrejection under the null hypothesis and therefore to apparently high (but meaningless) power. Interestingly, however, both WCU-R and WCU-M overreject about as severely as the t-test under the null but have less power for large values of β_2. WCR-R performs extremely well under the null and therefore has meaningful power. For larger values of β_2, it even has more power than some of the oversized tests. 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 $G = 25$ it has noticeably lower power.

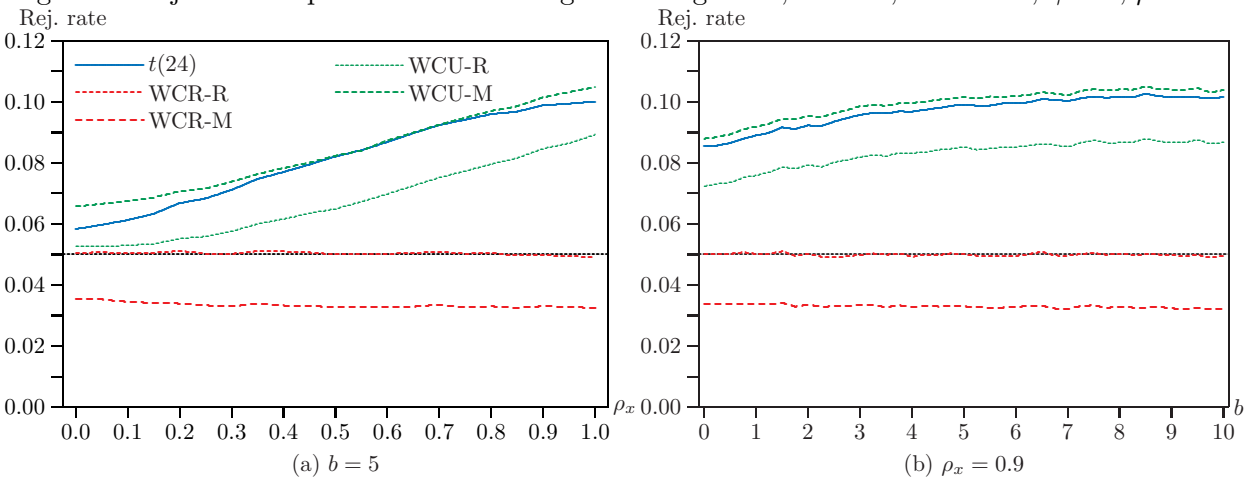
Overall, these results strongly favor WCR-R, the wild cluster restricted bootstrap using the Rademacher distribution. Moreover, the fact that all the tests seem to be converging to similar power functions as G increases from 10 to 25, which continues (in results that are not reported) as G increases further, 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. However, those results were for difference-in-differences regressions. Here we investigate the effects of heteroskedasticity in the model (20). 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.

Panel (a) of Figure 5 is comparable to Panel (a) of Figure 1. In both cases, there are 25 clusters, each with 100 observations. However, in Figure 5, the value of c is 1, which implies that there is substantial heteroskedasticity. When $\rho_x = 0$, so that the heteroskedasticity is solely at the individual level, all procedures perform very similarly in Figure 5 and 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 more striking. For larger values of ρ_x, the conventional procedure based on $t(24)$ critical values overrejects much more severely than it did before. So do the WCU-R and WCU-M bootstraps, although they now perform better relative to the conventional procedure. The restricted wild cluster bootstraps are hardly affected by this form of heteroskedasticity. Making the errors heteroskedastic has not changed the relationships among the five methods.

Panel (b) of Figure 5 demonstrates that these results are not a consequence of the choice to

Figure 6: Rejection frequencies with heterogeneous regressor, $G = 25$, $N = 2500$, $\gamma = 0$, $\rho = 0.10$



set $c = 1$. It shows rejection frequencies as a function of c for $\rho_x = 0.9$. Note that the horizontal axis has been subjected to a square 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. However, the effects of increasing c diminish rapidly as c increases beyond about 0.5.

In all the experiments reported so far, the regressor has the same distribution for all clusters. In Figure 6, we relax this assumption by allowing for heterogeneity across clusters. We introduce a parameter $b \geq 0$ which is used to generate the elements x_{ig} of the vector $\boldsymbol{x}_g$ in (20) as

$$x_{ig} = \Big(1 + b\frac{g-1}{G-1}\Big)w_{ig}, \tag{22}$$

where the w_{ig} are generated as weighted sums of $\chi^2(8)$ random variables in exactly the same way as the x_{ig} were generated for Figures 1–4. For (22), the variance of the x_{ig} increases with g for $b > 0$. There is no effect on the first cluster, and the effect is largest for the G^{th} one. Even for modest values of b, there is substantial heterogeneity across clusters. In practice, we would be surprised to encounter heterogeneity as extreme as that for the larger values of b in our experiments.

Panel (a) of Figure 6 is comparable to Panel (a) of Figure 1, and it looks very similar. The main effect of setting $b = 5$ instead of $b = 0$ is that the t-test and the two unrestricted WCB tests overreject noticeably more often for larger values of ρ_x. The two restricted WCB tests are not affected. Panel (b) of Figure 6 shows rejection frequencies as functions of b when $\rho_x = 0.9$. As b increases, the tendency of the t-test and the two unrestricted WCB tests to overreject becomes more pronounced, but again there is no effect on the two restricted WCB tests. There is a simple explanation for this phenomenon. Heterogeneity causes the observations associated with some clusters to have much higher leverage than the ones associated with other clusters. This necessarily affects the unrestricted residuals but not the restricted ones, because the latter are simply deviations from sample means.[4]

Several striking regularities emerge from Figures 1–6. The conventional t-test always overrejects, and it does so more severely for higher values of γ or ρ_x. The two unrestricted WCB tests always

[4]This may incorrectly suggest that our simulation results depend on the fact that (20) has only one regressor. In Appendix C, we therefore add up to 8 additional regressors. Doing so often causes the rejection frequencies to increase noticeably, but it does not change the ordering of the various methods.

overreject, with WCU-M doing so more severely than WCU-R. The overrejection always becomes more severe as ρ_x increases. These results do not carry over to the two restricted WCB tests. Except in Figure 3, where there is one big cluster, WCR-M always underrejects, and WCR-R has very accurate size. When G is allowed to vary, in Figure 2, the rejection frequencies for all four WCB tests approach 0.05 more rapidly than those of the t-test or the ordinary wild bootstrap test. In the next section, we explain these results using higher-order asymptotic theory.

5 Higher-Order Asymptotic Theory

In this section, we first derive Edgeworth expansions for the CDFs of the sample t-statistic and the WCB t-statistic. We apply these expansions to investigate several findings from the simulations in the previous section, such as the overrejection by the asymptotic t-test, the choice of auxiliary distribution in the WCB, and the choice between restricted (WCR) and unrestricted (WCU) bootstrap DGPs. We also discuss 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 improved to $o_P(G^{-m/2})$ for $m = 1, 2$, uniformly in x.

5.1 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. We consider both one-term and two-term Edgeworth expansions. The 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 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.

For two-sided symmetric tests, we have the 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 derive these expansions, we need to strengthen the conditions on the moments in Assumptions 1 and 2 and the conditions on the cluster sizes in Assumption 3. In particular, the latter is replaced by the following:

Assumption 4. 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 cluster sizes are bounded, which appears to be necessary to keep the theory tractable. We note that, under Assumption 4, 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,

$$\frac{\mu_N}{N} \to c_1, \quad \frac{G}{N} \to c_2, \quad \frac{G}{\mu_N} \to c_3, \quad \sqrt{G}(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = O_P(1); \tag{26}$$

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.

In what follows, we also need to assume the existence of more moments than in Assumptions 1 and 2. In particular, to derive the Edgeworth expansion of the CDF of t_a in (23), we apply the so-called "smooth function model", e.g. Bhattacharya and Ghosh (1978) and Skovgaard (1981). In the proof, see (B.21), we write t_a as a function of

$$\boldsymbol{W}_g = (\boldsymbol{W}_{1g}^\top, \operatorname{vech}(\boldsymbol{W}_{2g})^\top, \operatorname{vech}(\boldsymbol{W}_{3g})^\top, \operatorname{vech}(\boldsymbol{W}_{4g})^\top, \operatorname{vec}(\boldsymbol{W}_{5g})^\top)^\top, \tag{27}$$

where, for a symmetric matrix $\boldsymbol{A}$, $\operatorname{vech}(\boldsymbol{A})$ denotes the vector composed of the unique elements of $\boldsymbol{A}$, and we define

$$\boldsymbol{W}_{1g} = \boldsymbol{X}_g^\top \boldsymbol{u}_g, \quad \boldsymbol{W}_{2g} = \boldsymbol{W}_{1g}\boldsymbol{W}_{1g}^\top, \quad \boldsymbol{W}_{3g} = \boldsymbol{X}_g^\top \boldsymbol{X}_g, \quad \boldsymbol{W}_{4g} = \boldsymbol{W}_{3g} \otimes \boldsymbol{W}_{3g}, \quad \boldsymbol{W}_{5g} = \boldsymbol{W}_{3g} \otimes \boldsymbol{W}_{1g}. \tag{28}$$

We strengthen Assumptions 1 and 2 by imposing the following conditions. To allow for a constant term in the notation, we define $\breve{\boldsymbol{W}}_g$ to be equal to $\boldsymbol{W}_g$ with any non-random variables and redundant variables removed.

Assumption 5. The sequence $\{\breve{\boldsymbol{W}}_g\}$ is independent across g and $\sup_{g \in \mathbb{N}} \mathrm{E}\|\breve{\boldsymbol{W}}_g\|^{2+m+\lambda} < \infty$ for some $\lambda > 0$. In addition, $\mathrm{E}(\boldsymbol{u}_g|\boldsymbol{X}_g) = \boldsymbol{0}$, and the smallest eigenvalues of $G^{-1}\sum_{g=1}^G \operatorname{Var}(\breve{\boldsymbol{W}}_g)$ and $G^{-1}\sum_{g=1}^G \mathrm{E}(\boldsymbol{W}_{3g})$ are bounded away from zero for G sufficiently large.

The conditions in Assumption 5 imply and strengthen those in Assumptions 1 and 2. In particular, the existence of additional moments in Assumption 5 is required to derive cumulants of (an approximation to) the t-statistic, which are needed to derive the Edgeworth expansions. The predeterminedness condition, $\mathrm{E}(\boldsymbol{u}_g|\boldsymbol{X}_g) = \boldsymbol{0}$, is quite standard in linear regression, and the condition on $G^{-1}\sum_{g=1}^G \mathrm{E}(\boldsymbol{W}_{3g})$ is equivalent to the positive definiteness condition on $\boldsymbol{\Xi}_0$ in Assumption 2. More generally, Assumptions 4 and 5 imply Assumptions 1–3.

Validity of Edgeworth expansions requires further regularity conditions. In particular, we use "Cramér's condition", which is given as follows; see, e.g., Bhattacharya and Rao (1976, Thm. 20.6).

Assumption 6. The characteristic function $\chi_g(\boldsymbol{t})$ satisfies $\limsup_{g\to\infty} \limsup_{||\boldsymbol{t}||\to\infty} |\chi_g(\boldsymbol{t})| < 1$.

For the validity of the Edgeworth expansion of the CDF of t_a in (23), we will apply Assumption 6 with $\chi_g(\boldsymbol{t})$ denoting the characteristic function of $\breve{\boldsymbol{W}}_g$. This condition will be satisfied if the distribution of $\breve{\boldsymbol{W}}_g$ is sufficiently smooth (has a non-degenerate absolutely continuous component), which is the reason for the introduction of the notation $\breve{\boldsymbol{W}}_g$.

A similar condition is required on the characteristic function of the wild bootstrap auxiliary random variables v_g^* to prove validity of the Edgeworth expansion of the bootstrap distribution

in (24). This is, of course, theoretically appealing, but it rules out all commonly applied discrete distributions for v_g^*, including the Rademacher and Mammen distributions. See, for example, Liu (1988) or Kline and Santos (2012) for discussions in a non-cluster context.

In the next theorem, we show that, under Assumption 6, the Edgeworth expansions in (23)–(25) are valid and derive the functions q_j and $\ddot{q}_j$. If Assumption 6 does not hold, then we call (23)–(25) the "formal" Edgeworth expansions. It is quite common in the bootstrap literature, e.g. Mammen (1993), to analyze the formal Edgeworth expansions even in cases where Assumption 6 is not imposed and may not even hold. In those cases, the formal Edgeworth expansions are nevertheless often used to explain finite-sample simulation findings, such as the overrejection of the asymptotic test and superiority of the bootstrap, and also to shed light on the choice of the distribution of the auxiliary random variables, v_g^*, as well as the difference between the restricted and unrestricted versions of the bootstrap. We partly follow this approach in some of the subsequent analysis.

Theorem 5.1. *Suppose Assumptions 4 and 5 are satisfied for some $\lambda > 0$, that Assumption 6 is satisfied for the characteristic function of $\breve{\boldsymbol{W}}_g$, and that H_0 is true, where $\breve{\boldsymbol{W}}_g$ is defined in (27) and the paragraph following it. Then the m-term Edgeworth expansion of the CDF of t_a is given by (23) for $m = 1, 2$, while that of $|t_a|$ is given by (25) for $m = 2$ with*

$$
\begin{aligned}
q_1(x) &= \frac{1}{2d^{1/2}}\boldsymbol{a}^\top\boldsymbol{\gamma}_{1,1}\boldsymbol{a} + \frac{1}{3d^{3/2}}\boldsymbol{a}^\top\boldsymbol{\gamma}_{1,1}\boldsymbol{a}(x^2-1) \quad \textit{and} \\
q_2(x) &= -\frac{1}{d}\Big((\boldsymbol{a}^\top\boldsymbol{\gamma}_{1,1}\boldsymbol{a})^2 - \frac{1}{2}\operatorname{Tr}\{\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,1}\} + \operatorname{Tr}\{\boldsymbol{\gamma}_{1,3}\} - \boldsymbol{a}^\top\boldsymbol{\xi}_{1,1}\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,1}\boldsymbol{a} + 2\boldsymbol{a}^\top\boldsymbol{\xi}_{1,1}\boldsymbol{\xi}_{3,2}\Big)x \\
&\quad - \frac{1}{12d^2}\Big(8(\boldsymbol{a}^\top\boldsymbol{\gamma}_{1,1}\boldsymbol{a})^2 - \xi_{2,2} - 6\boldsymbol{a}^\top\boldsymbol{\xi}_{1,1}\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,1}\boldsymbol{a} + 12\boldsymbol{a}^\top\boldsymbol{\xi}_{1,1}\boldsymbol{\xi}_{3,2}\Big)(x^3-3x) \\
&\quad - \frac{1}{18d^3}(\boldsymbol{a}^\top\boldsymbol{\gamma}_{1,1}\boldsymbol{a})^2(x^5-10x^3+15x),
\end{aligned}
$$

where

$$
\boldsymbol{\gamma}_{m,n} = \frac{1}{G}\sum_{g=1}^{G}\mathrm{E}(\boldsymbol{a}^\top\boldsymbol{Z}_{1g}\boldsymbol{Z}_{mg}\boldsymbol{Z}_{ng}^\top), \quad \boldsymbol{\xi}_{m,n} = \frac{1}{G}\sum_{g=1}^{G}\mathrm{E}(\boldsymbol{Z}_{mg}\boldsymbol{Z}_{ng}^\top), \tag{29}
$$

and

$$
\boldsymbol{Z}_{1g} = \nu_a^{-1/2}\boldsymbol{\Xi}^{-1}\boldsymbol{W}_{1g}, \quad Z_{2g} = (\boldsymbol{a}^\top\boldsymbol{Z}_{1g})^2, \quad \boldsymbol{Z}_{3g} = \boldsymbol{W}_{3g}\boldsymbol{\Xi}^{-1}\boldsymbol{a}, \tag{30}
$$

with

$$
\boldsymbol{\Xi} = \frac{1}{G}\sum_{g=1}^{G}\mathrm{E}(\boldsymbol{X}_g^\top\boldsymbol{X}_g) \quad \textit{and} \quad \nu_a = \boldsymbol{a}^\top\boldsymbol{\Xi}^{-1}\frac{N^2}{G}\boldsymbol{\Gamma}\boldsymbol{\Xi}^{-1}\boldsymbol{a}. \tag{31}
$$

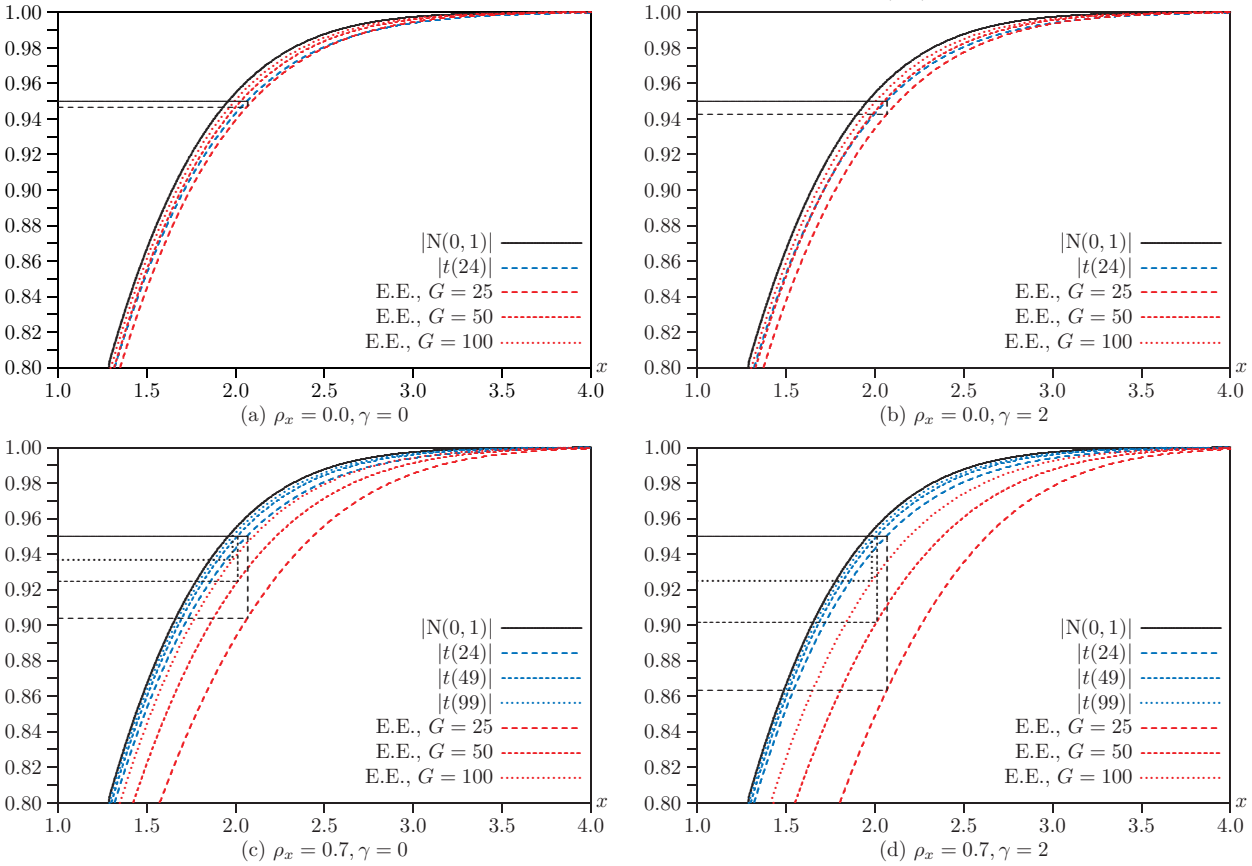
If, in addition, $\mathrm{E}^*|v^*|^{2+2m+\lambda} < \infty$ for some $\lambda > 0$ and Assumption 6 holds for the characteristic function of $(v_g^*, v_g^{*2})^\top$, then the Edgeworth expansions of the CDFs of t_a^* and $|t_a^*|$ are given by the same expressions as those of t_a, but with $\ddot{q}_j$ 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 $\boldsymbol{Z}_{ig}$ by $\boldsymbol{Z}_{ig}^*$, where

$$
\boldsymbol{Z}_{1g}^* = \ddot{\nu}_a^{-1/2}\ddot{\boldsymbol{\Xi}}^{-1}\boldsymbol{W}_{1g}^*, \qquad Z_{2g}^* = (\boldsymbol{a}^\top\boldsymbol{Z}_{1g}^*)^2, \qquad \boldsymbol{Z}_{3g}^* = \boldsymbol{W}_{3g}\ddot{\boldsymbol{\Xi}}^{-1}\boldsymbol{a}, \tag{32}
$$

$$
\boldsymbol{W}_{1g}^* = \boldsymbol{X}_g^\top\boldsymbol{u}_g^*, \qquad \ddot{\boldsymbol{\Xi}} = \bar{\boldsymbol{W}}_3 = \frac{1}{G}\sum_{g=1}^{G}\boldsymbol{W}_{3g}, \qquad \ddot{\nu}_a = \boldsymbol{a}^\top\bar{\boldsymbol{W}}_3^{-1}\frac{N^2}{G}\ddot{\boldsymbol{\Gamma}}\bar{\boldsymbol{W}}_3^{-1}\boldsymbol{a}. \tag{33}
$$

With the exception of the DGP for the simulation results reported in Figure 3 (one big cluster), which violates even the conditions for the first-order theory, all the DGPs considered in Section 4 that satisfy the null hypothesis also satisfy the regularity conditions for the expansions of the CDF

Figure 7: Edgeworth expansions of CDF of $|t_a|$



of $|t_a|$. In particular, they satisfy Assumption 6. Hence, we can explain the observed behavior of the t-test in Section 4 by analyzing the parameters $\boldsymbol{\gamma}_{m,n}$ and $\boldsymbol{\xi}_{m,n}$ in Theorem 5.1 and the functions $q_j(x)$. To this end, we now proceed to draw Edgeworth expansions of the CDF of $|t_a|$ for some of the DGPs considered in Section 4.

In each panel of Figure 7, we plot the right tail of the CDF of the absolute value of a standard normal random variable, i.e. $2\Phi(x) - 1$, along with CDFs of the absolute value of a random variable that follows the $t(G-1)$ distribution for $G = 25$, and also for $G = 50$ and $G = 100$ in the bottom two panels where there is less overlap with other curves. In addition, and this is the point of the figure, we plot the two-term Edgeworth expansions given in (25) and Theorem 5.1 for $G = 25$, $G = 50$, and $G = 100$. For the Edgeworth expansions, we calculate the parameters $\boldsymbol{\gamma}_{m,n}$ and $\boldsymbol{\xi}_{m,n}$ in Theorem 5.1, and hence the functions $q_j(x)$, using the true values from the DGP, and then plot the right-hand side of (25) against x, omitting the $o(G^{-1})$ term.

The four panels of Figure 7 correspond to different setups from Section 4. In each case, the DGP is (20) with $\rho = 0.1$, and the four panels differ in terms of ρ_x (either 0.0 or 0.7) and γ (either 0 or 2). These parameter values generate quite different behavior for the t-test, as documented in Figures 1 and 2. In each panel of Figure 7, and most obviously Panels (c) and (d), it is evident that the Edgeworth expansions provide a very substantial improvement over the CDFs of both the reference normal approximation and the $t(G-1)$ distribution. One way to see this is to follow the 0.95 percentile horizontally across a panel until reaching the CDF of the $t(G-1)$ distribution. This, of course, is the critical value used for the t-test as implemented, e.g., in Stata. Following

then a vertical line to the Edgeworth expansion, and after that a horizontal line across the panel, will give one minus the size of the t-test as predicted by the Edgeworth expansion. For reference, these lines have been drawn for $G = 25$ in all panels of Figure 7, and additionally for $G = 50$ and $G = 100$ in Panels (c) and (d).

In Panels (a) and (b) of Figure 7, where $\gamma = 0$ and $\gamma = 2$, respectively, with $\rho_x = 0.0$ in both cases, the Edgeworth expansions for $G = 25$ show that the t-test is oversized only very slightly. On the other hand, in Panels (c) and (d), where $\rho_x = 0.7$, size distortions are much larger, although they decrease as G increases. Overall, this corresponds well with the simulation results for the t-test presented in Figures 1 and 2, from which it appears that ρ_x, and to a slightly lesser extent also γ, play an important role in the size distortion observed for the t-test.

More generally, Figure 7 shows how the Edgeworth expansions, via the parameters $\boldsymbol{\gamma}_{m,n}$ and $\boldsymbol{\xi}_{m,n}$ in Theorem 5.1, can be used to analyze and "explain" the simulation findings for the t-test presented in Section 4. We next turn our attention to the different variants of the WCB and use the higher-order theory to shed light on the simulation results for them.

5.2 Refinements and Choice of Bootstrap

The bootstrap errors for one-sided and two-sided tests are given, uniformly in x, by

$$P^*(t_a^* \leq x) - P(t_a \leq x) = \sum_{j=1}^{m} G^{-j/2}(\ddot{q}_j(x) - q_j(x))\phi(x) + o_P(G^{-m/2}), \quad m = 1, 2, \tag{34}$$

$$P^*(|t_a^*| \leq x) - P(|t_a| \leq x) = 2G^{-1}(\ddot{q}_2(x) - q_2(x))\phi(x) + o_P(G^{-1}). \tag{35}$$

Based on the Edgeworth expansions in Theorem 5.1, (34) and (35) can be analyzed to discuss the possibility of refinements, i.e. conditions under which a bootstrap procedure is able to eliminate the leading term(s) on the right-hand side of either (34) or (35). We note that this carries the caveat of existence, c.f. the discussion immediately following Assumption 6 above, to which we return below.

The next theorem gives results for the skewness and kurtosis terms that appear on the right-hand sides of equations (34) and (35) after the functions $\ddot{q}_j(x)$ and $q_j(x)$ defined in Theorem 5.1 are substituted into them.

Theorem 5.2. *Suppose Assumptions 4 and 5 are satisfied for some $\lambda > 0$ and that $\mathrm{E}^*(v^{*4}) < \infty$. Then it holds that*

$$\boldsymbol{a}^\top \ddot{\boldsymbol{\gamma}}_{1,1}\boldsymbol{a} - \boldsymbol{a}^\top \boldsymbol{\gamma}_{1,1}\boldsymbol{a} = \boldsymbol{a}^\top \boldsymbol{\gamma}_{1,1}\boldsymbol{a}(\mathrm{E}^*(v^{*3}) - 1) + o_P(1),$$
$$\ddot{\xi}_{2,2} - \xi_{2,2} = \xi_{2,2}(\mathrm{E}^*(v^{*4}) - 1) + o_P(1).$$

In Theorem 5.2 we give results for the skewness correction term, $\boldsymbol{a}^\top \ddot{\boldsymbol{\gamma}}_{1,1}\boldsymbol{a} - \boldsymbol{a}^\top \boldsymbol{\gamma}_{1,1}\boldsymbol{a}$, and the kurtosis correction term, $\ddot{\xi}_{2,2} - \xi_{2,2}$. The names arise because $\boldsymbol{a}^\top \boldsymbol{\gamma}_{1,1}\boldsymbol{a} = G^{-1}\sum_{g=1}^{G} \mathrm{E}(\boldsymbol{a}^\top \boldsymbol{Z}_{1g})^3$ and $\xi_{2,2} = G^{-1}\sum_{g=1}^{G} \mathrm{E}(\boldsymbol{a}^\top \boldsymbol{Z}_{1g})^4$ can be interpreted as measures of skewness and kurtosis, respectively, of $\boldsymbol{a}^\top \hat{\boldsymbol{\beta}}$ (specifically of the random variable $\boldsymbol{a}^\top \boldsymbol{Z}_{1g}$). Although not presented in Theorem 5.2, the remaining parameters appearing in the function $\ddot{q}_2(x)$ in Theorem 5.1, i.e. $\ddot{\xi}_{3,3}$, etc., can be shown to be consistent for their non-bootstrap counterparts. We do not present these results because they do not depend on the moments of the auxiliary random variable, v^*, other than $\mathrm{E}^*(v^{*2}) = 1$.

The expansions in Theorem 5.1 show that the first term on the right-hand side of (34) is the skewness correction term, $\boldsymbol{a}^\top \ddot{\boldsymbol{\gamma}}_{1,1}\boldsymbol{a} - \boldsymbol{a}^\top \boldsymbol{\gamma}_{1,1}\boldsymbol{a}$. Theorem 5.2 shows that this is zero under either of two circumstances. The first is when the skewness term, $\boldsymbol{a}^\top \gamma_{1,1}\boldsymbol{a}$, is itself zero. 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 in which $N_g = 1$ for all g.

In other words, Theorems 5.1 and 5.2 show that the WCB can achieve a refinement in the sense of skewness correction, where the right-hand side of (34) is $o_P(G^{-1/2})$, uniformly in x, instead of $o_P(1)$, uniformly in x, as in Theorem 3.1, under either of the two conditions mentioned in the previous paragraph. To be rigorous, this would require Assumption 6 to hold for the distribution of the auxiliary random variable, v^*. Although there exist distributions with third moment of one for which Assumption 6 holds (e.g., Liu 1988), they are rarely used in practice. The main reason is that the fourth moment of such random variables is much greater than one.

The second result in Theorem 5.2 shows how a large fourth moment for the auxiliary distribution is undesirable because that would inflate, rather than eliminate, the kurtosis correction term, $\ddot{\xi}_{2,2} - \xi_{2,2}$, even asymptotically. Indeed, Theorem 5.2 shows that only an auxiliary distribution with fourth moment equal to one, i.e. $\mathrm{E}^*(v^{*4}) = 1$, can eliminate the kurtosis correction term. Although it does not satisfy Assumption 6, the Rademacher distribution is the only distribution with zero mean and second and fourth moments equal to one. Therefore, only it can eliminate the kurtosis term in the formal bootstrap errors given by equations (34) and (35).

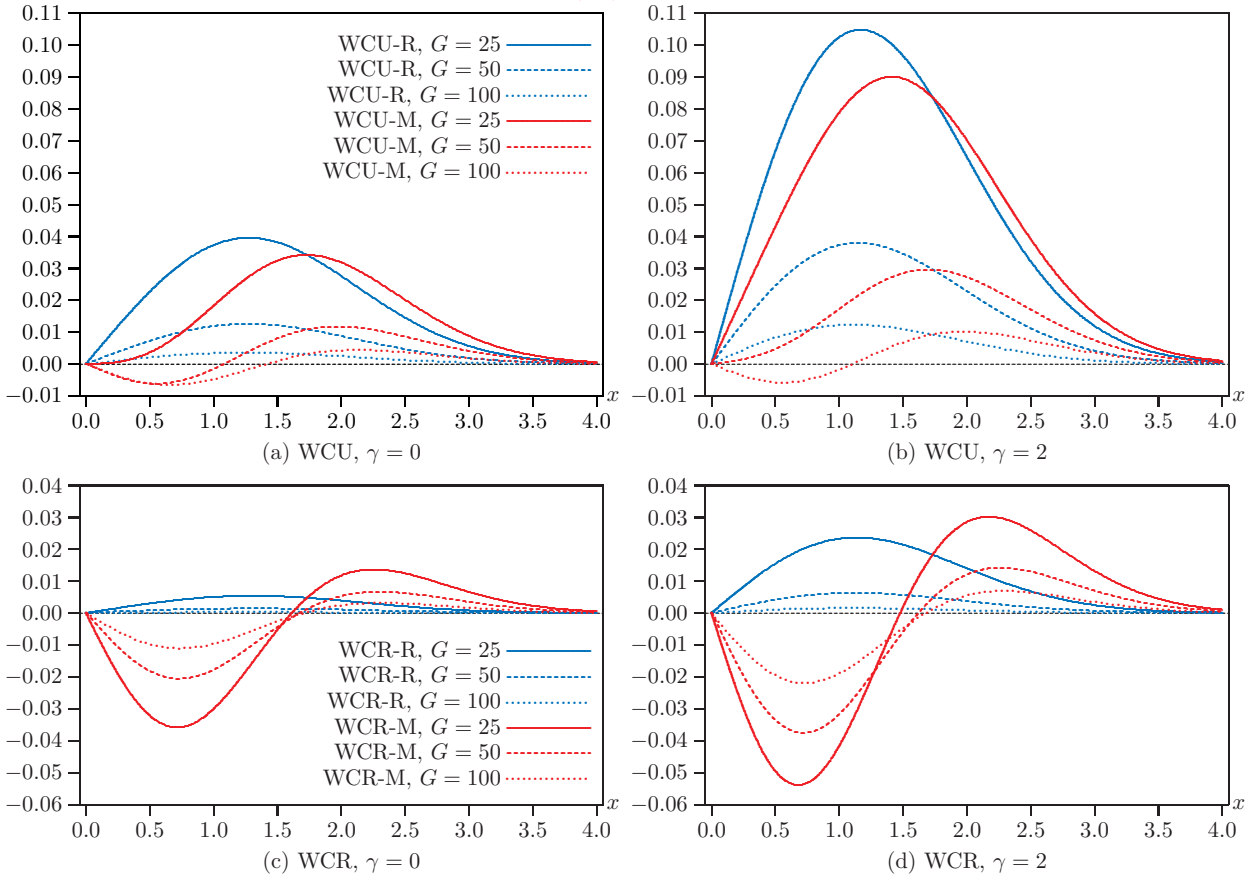
This analysis reveals a trade-off between skewness correction and kurtosis correction, that is, between the relative importance of the third and fourth moments. The Mammen (1993) distribution is the most commonly applied auxiliary distribution that provides a skewness correction, while the Rademacher distribution is the only one that provides a kurtosis correction. Consequently, while the expansions in Theorem 5.1 can only be considered formal expansions for these choices of auxiliary distributions, because neither of them satisfies Assumption 6, we can use the formal expansions to inform the choice of auxiliary distribution.

We also note from Theorem 5.1 that the bootstrap errors in (34) and (35) depend on whether restricted or unrestricted parameter estimates are used in the bootstrap DGP, via the estimates $\ddot{\boldsymbol{\gamma}}_{m,n}$ and $\ddot{\boldsymbol{\xi}}_{m,n}$ in Theorem 5.1. To avoid excessive reliance on theoretical analysis that is hindered by the issue of existence and the conditions of Assumption 6 for the auxiliary random variables, we instead take a graphical approach and plot the formal bootstrap errors in (34)–(35). We then use these to inform both the choice of auxiliary distribution and the choice of restricted or unrestricted estimates in the bootstrap DGP.

In Figure 8, we plot the formal bootstrap error for the two-sided case, i.e. the right-hand side of (35), ignoring the $o_P(G^{-1})$ term, for two common choices of auxiliary distribution, namely, the Rademacher and Mammen distributions. As in Figure 7, the setup is the same as that in Figure 1 (with $\rho_x = 0.7$) and Panel (b) of Figure 2. In particular, therefore, the errors are skewed, suggesting that the Mammen distribution may have an advantage in this case, or at least that there is a nontrivial trade-off between skewness correction and kurtosis correction.

In Panels (a) and (b) of Figure 8 we plot the bootstrap error for the WCU bootstrap for $\gamma = 0$ (equal-sized clusters) and $\gamma = 2$ (unbalanced clusters), respectively. The plots are shown for both the Rademacher and Mammen auxiliary distributions and for $G = 25$, $G = 50$, and $G = 100$. Three features emerge from these two panels. First, the maximum bootstrap error for the Rademacher and Mammen auxiliary distributions is similar, although it occurs further to the right for the latter. Second, for both auxiliary distributions, the WCU bootstrap has too much mass near the left, and particularly in the center of the distribution, which results in the bootstrap distribution having too little mass in the right tail. This implies that the WCU bootstrap will overreject. Third, while the tendency to overreject is evident in both Panels (a) and (b), it is most noticeable in Panel (b), suggesting that the WCU bootstrap will overreject more for higher values of γ. Indeed, these three features for the WCU bootstrap correspond exactly to what was found in our simulations in Section 4, and especially Figure 1.

Figure 8: Bootstrap error (35) based on two-term expansions



(a) WCU, $\gamma = 0$ (b) WCU, $\gamma = 2$

(c) WCR, $\gamma = 0$ (d) WCR, $\gamma = 2$

The corresponding plots for the WCR bootstrap are shown in Panels (c) and (d) of Figure 8. Again, there are three noticeable features. First, the magnitudes of the bootstrap errors for the WCR bootstrap are lower than for the WCU bootstrap, suggesting that the former should have better size control than the latter. Second, the bootstrap error for the WCR bootstrap with the Mammen distribution shows a very clear shift of mass from left to right, implying that the bootstrap distribution will have too much mass in the right tail, leading to negative size distortion (under-rejection). Third, comparing the Rademacher and Mammen auxiliary distributions, it is clear that the former has much smaller bootstrap errors than the latter. Again, these three features correspond exactly to our simulation findings in Section 4.

The WCR bootstrap with the Rademacher auxiliary distribution clearly has the smallest bootstrap errors among all four bootstrap methods in Panels (a) and (c), and also among all four in Panels (b) and (d). This may be surprising, because the error terms are skewed, which should favor the Mammen distribution. Moreover, it is striking that, upon closer examination, all the curves in Figure 8 seem to be approaching the zero line at rate G^{-1}, except for the WCR-R ones, which are clearly approaching it faster than that. This suggests that the skewness correction is less important than the kurtosis correction in this case, which results in the smaller bootstrap errors for the Rademacher auxiliary distribution. Of course, the finding that the Rademacher distribution is superior to the Mammen distribution in the context of the wild bootstrap is not particularly surprising. There is, in fact, a good deal of simulation evidence that, for the ordinary wild bootstrap

without clustering, using a v^* with third moment of one often does not work particularly well; see, e.g., Davidson, Monticini, and Peel (2007) and Davidson and Flachaire (2008).

Finally, when the bootstrap error is very small in Figure 8, the bootstrap achieves the same rejection frequency as the Edgeworth CDFs in Figure 7. Thus, the very small bootstrap error for WCR-R in Panels (c) and (d) of Figure 8 gives a theoretical explanation for the excellent finite-sample size control for WCR-R tests observed in Section 4.

6 Practical Guidance

The first-order asymptotic theory of Sections 2 and 3, the simulation evidence of Section 4 and Appendix C, and the higher-order theory of Section 5 together provide a good deal of practical guidance for making cluster-robust inferences.

Because Theorem 2.1 justifies the use of the CRVE $\hat{\boldsymbol{V}}$ in (5) when the number of clusters G is large, it might seem that we simply have to count clusters. However, this is not sufficient and can be very misleading. For asymptotic inference based on cluster-robust standard errors and the $t(G-1)$ distribution to be reliable when G is not very large, the clusters cannot be too heterogeneous, in terms of either the cluster sizes N_g or the matrices $\boldsymbol{X}_g^\top \boldsymbol{X}_g$ and $\boldsymbol{\Sigma}_g$. In addition, the extent to which regressors vary between rather than within clusters can matter greatly. Thus, when attempting to make inferences using a CRVE, it is advisable to keep the following points in mind:

- Under ideal circumstances, asymptotic inference is probably reliable with $G \geq 50$ clusters. However, circumstances are rarely ideal.
- The more heterogeneity there is across clusters, the larger is the number of clusters needed for reliable inference. With extreme heterogeneity, it may be impossible (see Figure 3).
- The more the key regressor varies between rather than within clusters, the larger is the value of G needed for reliable inference. If there are additional regressors that vary principally between clusters, the number is probably even larger; see Appendix C.
- The "effective number of clusters" proposed in Carter, Schnepel, and Steigerwald (2017) provides a useful diagnostic. When it is much smaller than G, and especially when it is small in absolute value (say, less than 30), asymptotic inference cannot be expected to work well.
- The best version of the wild cluster bootstrap provides much more reliable inferences than the t-test with $G-1$ degrees of freedom. This is WCR-R, which employs restricted estimates and the Rademacher distribution.
- The two variants of the wild cluster bootstrap that use unrestricted residuals (WCU-R and WCU-M) generally reject more often than WCR-R, and their performance is closer to that of the t-test; WCU-M is particularly unreliable. In contrast, WCR-M often underrejects. Thus, both unrestricted residuals and the Mammen distribution should normally be avoided.
- Because the wild cluster bootstrap is remarkably inexpensive to implement when G is not too large, even when N is very large (Roodman et al. 2019), we recommend using WCR-R whenever there is any reason to believe that asymptotic inference may not be reliable.
- The ordinary wild bootstrap (WB) should rarely be used. In our simulations, it never seems to perform as well as WCR-R. Moreover, since it offers no asymptotic refinement, its performance improves more slowly with G than that of the WCB. However, MacKinnon and Webb (2018) discusses certain special cases where it may be desirable to use the WB.

7 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 does not assume that the regressors are exogenous, and 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, Theorem 5.1 provides higher-order asymptotic theory that explains the sometimes poor performance of asymptotic tests, and Theorem 5.2 gives conditions under which the WCB may attain a higher-order asymptotic refinement. It also sheds light on the choice of auxiliary distribution in the WCB and the choice between restricted and unrestricted residuals in the bootstrap DGP. Both simulation evidence and higher-order theory suggest that the restricted WCB using the Rademacher auxiliary distribution is generally the best choice.

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.

Bhattacharya, R. N., and J. K. Ghosh. 1978. On the validity of the formal Edgeworth expansion. *Annals of Statistics* 6:434–451.

Bhattacharya, R. N., and R. R. Rao. 1976. *Normal Approximation and Asymptotic Expansions*. Philadelphia: SIAM.

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. 2004. *Econometric Theory and Methods*. New York: Oxford University Press.

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

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

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

Hansen, B. E., and S. Lee. 2019. Asymptotic theory for clustered samples. *Journal of Econometrics* to appear.

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. 2018. The wild bootstrap for few (treated) clusters. *Econometrics Journal* 21:114–135.

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

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

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* 36:672–683.

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

Skovgaard, I. B. 1981. Transformation of an Edgeworth expansion by a sequence of smooth functions. *Scandinavian Journal of Statistics* 8:207–217.

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

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

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

Supplementary Material for "Asymptotic Theory and Wild Bootstrap Inference with Clustered Errors"

Antoine A. Djogbenou
York University
daa@yorku.ca

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

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

April 18, 2019

Abstract

This supplementary material to Djogbenou, MacKinnon, and Nielsen (2019) contains three appendices. Appendix A states and proves some preliminary lemmas. These are used in the proofs of the main results, which are given in Appendix B. Appendix C presents some additional simulation evidence to complement the simulation results in Section 4 of the paper.

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,*

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

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

$$\mathrm{E}\|\boldsymbol{s}_g\|^\theta = \mathrm{E}\Big\|\sum_{i=1}^{N_g} \boldsymbol{s}_{ig}\Big\|^\theta \leq \mathrm{E}\Big(\sum_{i=1}^{N_g} \|\boldsymbol{s}_{ig}\|\Big)^\theta \leq N_g^{\theta-1} \sum_{i=1}^{N_g} \mathrm{E}\|\boldsymbol{s}_{ig}\|^\theta. \tag{A.1}$$

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

We next give three lemmas that will be used to derive the required moments for the higher-order theory of Section 5.

Lemma A.3. *Suppose Assumptions 4 and 5 are satisfied. Let $\boldsymbol{Z}_{mg}$ be given by (30) and (B.38)–(B.46), and let $\boldsymbol{\mu}_{mg} = \mathrm{E}(\boldsymbol{Z}_{mg})$. For any integer $k \geq 2$, for which the following moment exists, it holds that*

$$\mathrm{E}\Big(\prod_{j=1}^{k} (\bar{\boldsymbol{Z}}_{m_j} - \bar{\boldsymbol{\mu}}_{m_j})\Big) = \begin{cases} O(G^{-k/2}) & \text{if } k \text{ is even,} \\ O(G^{-(k+1)/2}) & \text{if } k \text{ is odd.} \end{cases}$$

Proof. The left-hand side is

$$\mathrm{E}\Big(\prod_{j=1}^{k} (\bar{\boldsymbol{Z}}_{m_j} - \bar{\boldsymbol{\mu}}_{m_j})\Big) = G^{-k} \sum_{g_1,\ldots,g_k=1}^{G} \mathrm{E}\Big(\prod_{j=1}^{k} (\boldsymbol{Z}_{m_j,g_j} - \boldsymbol{\mu}_{m_j,g_j})\Big),$$

where the summation indexes $g_1, \ldots, g_k$ must be equal at least in pairs because $\mathrm{E}(\boldsymbol{Z}_{m_j,g_j} - \boldsymbol{\mu}_{m_j,g_j}) = 0$ for $j = 1, \ldots, k$. The result then follows directly because the normalization is G^{-k}. □

Lemma A.4. *Suppose Assumptions 4 and 5 are satisfied. Let $\boldsymbol{Z}_{mg}$ be given by (30) and (B.38)–(B.46), and let $\boldsymbol{\mu}_{mg} = \mathrm{E}(\boldsymbol{Z}_{mg})$. Then, when the following moments exist, it holds that*

$$\begin{aligned}
&\mathrm{E}(\bar{Z}_6^2) = G^{-1}, && \mathrm{E}(\bar{Z}_6^3(\bar{Z}_2 - \bar{\mu}_2)) = 3G^{-2}\gamma_{6,6} + O(G^{-3}),\\
&\mathrm{E}(\bar{Z}_6^3) = G^{-2}\gamma_{6,6}, && \mathrm{E}(\bar{Z}_6^4(\bar{Z}_2 - \bar{\mu}_2)) = 6G^{-3}(\xi_{2,2} - \bar{\xi}_{2,2}) + 4G^{-3}\gamma_{6,6}^2 + O(G^{-4}),\\
&\mathrm{E}(\bar{Z}_6^4) = 3G^{-2} + G^{-3}(\xi_{2,2} - 3\bar{\xi}_{2,2}), && \mathrm{E}(\bar{Z}_6^2(\bar{Z}_2 - \bar{\mu}_2)^2) = G^{-2}(\xi_{2,2} - \bar{\xi}_{2,2}) + 2G^{-2}\gamma_{6,6}^2 + O(G^{-3}),\\
&\mathrm{E}(\bar{Z}_6^6) = 15G^{-3} + O(G^{-4}), && \mathrm{E}(\bar{Z}_6^4(\bar{Z}_2 - \bar{\mu}_2)^2) = 3G^{-3}(\xi_{2,2} - \bar{\xi}_{2,2}) + 12G^{-3}\gamma_{6,6}^2 + O(G^{-4}),\\
&\mathrm{E}(\bar{Z}_6(\bar{Z}_2 - \bar{\mu}_2)) = G^{-1}\gamma_{6,6}, && \mathrm{E}(\bar{Z}_6^2 \bar{\boldsymbol{Z}}_1^\top \bar{\boldsymbol{Z}}_5) = G^{-2}(\mathrm{Tr}\{\boldsymbol{\gamma}_{1,3}\} + 2\boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,2}) + O(G^{-3}),\\
&\mathrm{E}(\bar{Z}_6^2(\bar{Z}_2 - \bar{\mu}_2)) = G^{-2}(\xi_{2,2} - \bar{\xi}_{2,2}), && \mathrm{E}(\bar{Z}_6^4 \bar{\boldsymbol{Z}}_1^\top \bar{\boldsymbol{Z}}_5) = G^{-3}(3\,\mathrm{Tr}\{\boldsymbol{\gamma}_{1,3}\} + 12\boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,2}) + O(G^{-4}),
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{E}(((\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top \bar{\boldsymbol{Z}}_1)^2) = G^{-2}\,\mathrm{Tr}\{(\boldsymbol{\xi}_{3,3} - \bar{\boldsymbol{\xi}}_{3,3})\boldsymbol{\xi}_{1,1}\} + O(G^{-3}),\\
&\mathrm{E}(\bar{Z}_6^2((\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top \bar{\boldsymbol{Z}}_1)^2) = G^{-3}\,\mathrm{Tr}\{(\boldsymbol{\xi}_{3,3} - \bar{\boldsymbol{\xi}}_{3,3})\boldsymbol{\xi}_{1,1}\} + 2G^{-3}\boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,6} - 2G^{-3}\boldsymbol{\xi}_{6,1}\bar{\boldsymbol{\xi}}_{3,3}\boldsymbol{\xi}_{1,6} + O(G^{-4}),\\
&\mathrm{E}(\bar{Z}_6^2 \bar{\boldsymbol{Z}}_1^\top \bar{\boldsymbol{\mu}}_4 \bar{\boldsymbol{Z}}_1) = G^{-2}\,\mathrm{Tr}\{\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,1}\} + 2G^{-2}\boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,6} + O(G^{-3}),\\
&\mathrm{E}(\bar{Z}_6^4 \bar{\boldsymbol{Z}}_1^\top \bar{\boldsymbol{\mu}}_4 \bar{\boldsymbol{Z}}_1) = 3G^{-3}\,\mathrm{Tr}\{\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,1}\} + 12G^{-3}\boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,6} + O(G^{-4}),
\end{aligned}$$

where $\bar{\boldsymbol{\xi}}_{m,n} = G^{-1}\sum_{g=1}^{G} \boldsymbol{\mu}_{mg}\boldsymbol{\mu}_{ng}^\top$.

Furthermore, for $m_1 \in \{1,5,6\}$, $m_2, m_3 \in \{3,4,9,10,11,12\}$, *and* $m_4 \in \{2,8\}$,

$$\begin{aligned}
&\mathrm{E}(\bar{Z}_6 \bar{\boldsymbol{Z}}_{m_1}^\top (\bar{\boldsymbol{Z}}_{m_2} - \bar{\boldsymbol{\mu}}_{m_2})) = G^{-2} \operatorname{Tr}\{\boldsymbol{\gamma}_{m_1,m_2} - \bar{\boldsymbol{\gamma}}_{m_1,m_2}\},\\
&\mathrm{E}(\bar{Z}_6^3 \bar{\boldsymbol{Z}}_{m_1}^\top (\bar{\boldsymbol{Z}}_{m_2} - \bar{\boldsymbol{\mu}}_{m_2})) = 3G^{-3} \operatorname{Tr}\{\boldsymbol{\gamma}_{m_1,m_2} - \bar{\boldsymbol{\gamma}}_{m_1,m_2}\} + 3G^{-3} \boldsymbol{\xi}_{6,m_1} (\boldsymbol{\xi}_{m_2,2} - \bar{\boldsymbol{\xi}}_{m_2,2}) + O(G^{-4}),\\
&\mathrm{E}(\bar{Z}_6 \bar{\boldsymbol{Z}}_{m_1}^\top (\bar{\boldsymbol{Z}}_{m_2} - \bar{\boldsymbol{\mu}}_{m_2})^\top (\bar{\boldsymbol{Z}}_{m_4} - \bar{\boldsymbol{\mu}}_{m_4})) = G^{-2} \boldsymbol{\xi}_{6,m_1} (\boldsymbol{\zeta}_{m_2,m_4} - \bar{\boldsymbol{\zeta}}_{m_2,m_4}) + O(G^{-3}),\\
&\mathrm{E}(\bar{Z}_6^3 \bar{\boldsymbol{Z}}_{m_1}^\top (\bar{\boldsymbol{Z}}_{m_2} - \bar{\boldsymbol{\mu}}_{m_2})^\top (\bar{\boldsymbol{Z}}_{m_4} - \bar{\boldsymbol{\mu}}_{m_4})) = 3G^{-3} \boldsymbol{\xi}_{6,m_1} (\boldsymbol{\zeta}_{m_2,m_4} - \bar{\boldsymbol{\zeta}}_{m_2,m_4}) + O(G^{-4}),\\
&\mathrm{E}(\bar{Z}_6 \bar{\boldsymbol{Z}}_{m_1}^\top (\bar{\boldsymbol{Z}}_{m_2} - \bar{\boldsymbol{\mu}}_{m_2})^\top (\bar{\boldsymbol{Z}}_{m_3} - \bar{\boldsymbol{\mu}}_{m_3})) = G^{-2} \boldsymbol{\xi}_{6,m_1} (\boldsymbol{\zeta}_{m_2,m_3} - \bar{\boldsymbol{\zeta}}_{m_2,m_3}) + O(G^{-3}),\\
&\mathrm{E}(\bar{Z}_6^3 \bar{\boldsymbol{Z}}_{m_1}^\top (\bar{\boldsymbol{Z}}_{m_2} - \bar{\boldsymbol{\mu}}_{m_2})^\top (\bar{\boldsymbol{Z}}_{m_3} - \bar{\boldsymbol{\mu}}_{m_3})) = 3G^{-3} \boldsymbol{\xi}_{6,m_1} (\boldsymbol{\zeta}_{m_2,m_3} - \bar{\boldsymbol{\zeta}}_{m_2,m_3}) + O(G^{-4}),
\end{aligned}$$

where

$$\bar{\boldsymbol{\gamma}}_{m,n} = \frac{1}{G} \sum_{g=1}^{G} \mathrm{E}(Z_{6g} \boldsymbol{Z}_{mg}) \boldsymbol{\mu}_{ng}^\top \qquad \boldsymbol{\zeta}_{m,n} = \frac{1}{G} \sum_{g=1}^{G} \mathrm{E}(\boldsymbol{Z}_{mg}^\top \boldsymbol{Z}_{ng}), \qquad \bar{\boldsymbol{\zeta}}_{m,n} = \frac{1}{G} \sum_{g=1}^{G} \boldsymbol{\mu}_{mg}^\top \boldsymbol{\mu}_{ng}.$$

Proof. First, note that the $\boldsymbol{Z}_{mg}$ only appear in deviations from the mean, i.e., all summations are over products of zero-mean random variables. The implication is that, in all summations, the indexes must be equal at least in pairs; see Lemma A.3.

For the moments of $\bar{Z}_6$ we find by Assumption 5 that

$$\begin{aligned}
&\mathrm{E}(\bar{Z}_6^2) = G^{-2} \sum_{g=1}^{G} \mathrm{E}(Z_{6g}^2) = G^{-2} \sum_{g=1}^{G} \mathrm{E}(Z_{2g}) = G^{-1} \bar{\mu}_2 = G^{-1},\\
&\mathrm{E}(\bar{Z}_6^3) = G^{-3} \sum_{g=1}^{G} \mathrm{E}(Z_{6g}^3) = G^{-2} \gamma_{6,6},\\
&\mathrm{E}(\bar{Z}_6^4) = G^{-3} \xi_{2,2} + 3G^{-4} \sum_{g,h=1}^{G} \mathrm{E}(Z_{6g}^2) \mathrm{E}(Z_{6h}^2) - 3G^{-4} \sum_{g=1}^{G} \left(\mathrm{E}(Z_{6g}^2)\right)^2 = G^{-3} \xi_{2,2} + 3G^{-2} - 3G^{-3} \bar{\xi}_{2,2},\\
&\mathrm{E}(\bar{Z}_6^6) = 15G^{-6} \sum_{g,h,i=1}^{G} \mathrm{E}(Z_{6g}^2) \mathrm{E}(Z_{6h}^2) \mathrm{E}(Z_{6i}^2) + O(G^{-4}) = 15G^{-3} + O(G^{-4}).
\end{aligned}$$

For the cross-moments of $\bar{Z}_6$, $\bar{\boldsymbol{Z}}_1$, and $\bar{\boldsymbol{Z}}_5$, we note that $\boldsymbol{Z}_{5g} = \boldsymbol{Z}_{3g} Z_{6g}$ and find

$$\begin{aligned}
\mathrm{E}(\bar{Z}_6^2 \bar{\boldsymbol{Z}}_1^\top \bar{\boldsymbol{Z}}_5) &= G^{-4} \sum_{g_1,\ldots,g_4=1}^{G} \mathrm{E}(Z_{6g_1} Z_{6g_2} \boldsymbol{Z}_{1g_3}^\top \boldsymbol{Z}_{5g_4})\\
&= G^{-4} \sum_{g,h=1}^{G} \mathrm{E}(Z_{6g}^2) \mathrm{E}(\boldsymbol{Z}_{1h}^\top \boldsymbol{Z}_{5h}) + 2G^{-4} \sum_{g,h=1}^{G} \mathrm{E}(Z_{6g} \boldsymbol{Z}_{1g}^\top) \mathrm{E}(Z_{6h} \boldsymbol{Z}_{5h}) + O(G^{-3})\\
&= G^{-3} \sum_{g=1}^{G} \mathrm{E}(Z_{6g} \boldsymbol{Z}_{1g}^\top \boldsymbol{Z}_{3g}) + 2G^{-4} \sum_{g,h=1}^{G} \mathrm{E}(Z_{6g} \boldsymbol{Z}_{1g}^\top) \mathrm{E}(Z_{6h}^2 \boldsymbol{Z}_{3h}) + O(G^{-3})\\
&= G^{-2} \operatorname{Tr}\{\boldsymbol{\gamma}_{1,3}\} + 2G^{-2} \boldsymbol{\xi}_{6,1} \boldsymbol{\xi}_{3,2} + O(G^{-3}),\\
\mathrm{E}(\bar{Z}_6^4 \bar{\boldsymbol{Z}}_1^\top \bar{\boldsymbol{Z}}_5) &= G^{-6} \sum_{g_1,\ldots,g_6=1}^{G} \mathrm{E}(Z_{6g_1} Z_{6g_2} Z_{6g_3} Z_{6g_4} \boldsymbol{Z}_{1g_5}^\top \boldsymbol{Z}_{5g_6})\\
&= 3G^{-6} \sum_{g,h,i=1}^{G} \mathrm{E}(Z_{6g}^2) \mathrm{E}(Z_{6h}^2) \mathrm{E}(\boldsymbol{Z}_{1i}^\top \boldsymbol{Z}_{5i})
\end{aligned}$$

$$+ 12G^{-6} \sum_{g,h,i=1}^{G} \mathrm{E}(Z_{6g}^2)\mathrm{E}(Z_{6h}\boldsymbol{Z}_{1h}^\top)\mathrm{E}(Z_{6i}\boldsymbol{Z}_{5i}) + O(G^{-4})$$
$$= 3G^{-4}\sum_{g=1}^{G} \mathrm{E}(Z_{6g}\boldsymbol{Z}_{1g}^\top\boldsymbol{Z}_{3g}) + 12G^{-5}\sum_{g,h=1}^{G} \mathrm{E}(Z_{6g}\boldsymbol{Z}_{1g}^\top)\mathrm{E}(Z_{6h}^2\boldsymbol{Z}_{3h}) + O(G^{-4})$$
$$= 3G^{-3}\,\mathrm{Tr}\{\boldsymbol{\gamma}_{1,3}\} + 12G^{-3}\boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,2} + O(G^{-4}).$$

Next, the cross-moments of $\bar{Z}_6$ and $\bar{Z}_2$ are

$$\mathrm{E}(\bar{Z}_6(\bar{Z}_2 - \bar{\mu}_2)) = G^{-2}\sum_{g=1}^{G} \mathrm{E}(Z_{6g}(Z_{2g} - \mu_{2g})) = G^{-2}\sum_{g=1}^{G} \mathrm{E}(Z_{6g}Z_{2g}) = G^{-1}\gamma_{6,6},$$
$$\mathrm{E}(\bar{Z}_6^2(\bar{Z}_2 - \bar{\mu}_2)) = G^{-3}\sum_{g=1}^{G} \mathrm{E}(Z_{6g}^2(Z_{2g} - \mu_{2g})) = G^{-2}(\xi_{2,2} - \bar{\xi}_{2,2}),$$
$$\mathrm{E}(\bar{Z}_6^3(\bar{Z}_2 - \bar{\mu}_2)) = 3G^{-4}\sum_{g,h=1}^{G} \mathrm{E}(Z_{6g}^2)\mathrm{E}(Z_{6h}(Z_{2h} - \mu_{2h})) + O(G^{-3})$$
$$= 3G^{-3}\sum_{g=1}^{G} \mathrm{E}(Z_{6g}Z_{2g}) + O(G^{-3}) = 3G^{-2}\gamma_{6,6} + O(G^{-3}),$$
$$\mathrm{E}(\bar{Z}_6^4(\bar{Z}_2 - \bar{\mu}_2)) = 6G^{-5}\sum_{g,h=1}^{G} \mathrm{E}(Z_{6g}^2)\mathrm{E}(Z_{6h}^2(Z_{2h} - \mu_{2h})) + 4G^{-5}\sum_{g,h=1}^{G} \mathrm{E}(Z_{6g}^3)\mathrm{E}(Z_{6h}(Z_{2h} - \mu_{2h})) + O(G^{-4})$$
$$= 6G^{-3}(\xi_{2,2} - \bar{\xi}_{2,2}) + 4G^{-3}\gamma_{6,6}^2 + O(G^{-4}),$$

and

$$\mathrm{E}(\bar{Z}_6^2(\bar{Z}_2 - \bar{\mu}_2)^2) = G^{-4}\sum_{g,h=1}^{G} \mathrm{E}(Z_{6g}^2)\mathrm{E}((Z_{2h} - \mu_{2h})^2) + 2G^{-4}\Big(\sum_{g=1}^{G} \mathrm{E}(Z_{6g}(Z_{2g} - \mu_{2g}))\Big)^2 + O(G^{-3})$$
$$= G^{-2}(\xi_{2,2} - \bar{\xi}_{2,2}) + 2G^{-2}\gamma_{6,6}^2 + O(G^{-3}),$$
$$\mathrm{E}(\bar{Z}_6^4(\bar{Z}_2 - \bar{\mu}_2)^2) = 3G^{-6}\sum_{g,h,i=1}^{G} \mathrm{E}(Z_{6g}^2)\mathrm{E}(Z_{6h}^2)\mathrm{E}((Z_{2i} - \mu_{2i})^2)$$
$$+ 12G^{-6}\sum_{g=1}^{G} \mathrm{E}(Z_{6g}^2)\Big(\sum_{h=1}^{G} \mathrm{E}(Z_{6h}(Z_{2h} - \mu_{2h}))\Big)^2 + O(G^{-4})$$
$$= 3G^{-3}(\xi_{2,2} - \bar{\xi}_{2,2}) + 12G^{-3}\gamma_{6,6}^2 + O(G^{-4}).$$

Next, using the law of iterated expectations,

$$\mathrm{E}(((\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top\bar{\boldsymbol{Z}}_1)^2) = G^{-4}\sum_{g_1,\ldots,g_4=1}^{G} \mathrm{Tr}\left\{\mathrm{E}((\boldsymbol{Z}_{3g_1} - \boldsymbol{\mu}_{3g_1})^\top\boldsymbol{Z}_{1g_2}(\boldsymbol{Z}_{3g_3} - \boldsymbol{\mu}_{3g_3})^\top\boldsymbol{Z}_{1g_4})\right\}$$
$$= G^{-4}\sum_{g,h=1}^{G} \mathrm{Tr}\left\{\mathrm{E}((\boldsymbol{Z}_{3g} - \boldsymbol{\mu}_{3g})(\boldsymbol{Z}_{3g} - \boldsymbol{\mu}_{3g})^\top)\mathrm{E}(\boldsymbol{Z}_{1h}\boldsymbol{Z}_{1h}^\top)\right\} + O(G^{-3})$$
$$= G^{-4}\sum_{g,h=1}^{G} \mathrm{Tr}\left\{\mathrm{E}(\boldsymbol{Z}_{3g}\boldsymbol{Z}_{3g}^\top)\mathrm{E}(\boldsymbol{Z}_{1h}\boldsymbol{Z}_{1h}^\top)\right\} - G^{-4}\sum_{g,h=1}^{G} \boldsymbol{\mu}_{3g}^\top\mathrm{E}(\boldsymbol{Z}_{1h}\boldsymbol{Z}_{1h}^\top)\boldsymbol{\mu}_{3g} + O(G^{-3})$$
$$= G^{-2}\,\mathrm{Tr}\left\{(\boldsymbol{\xi}_{3,3} - \bar{\boldsymbol{\xi}}_{3,3})\boldsymbol{\xi}_{1,1}\right\} + O(G^{-3}),\ \text{and}$$

$$\begin{aligned}
\mathrm{E}(\bar{Z}_6^2((\bar{\boldsymbol{Z}}_3-\bar{\boldsymbol{\mu}}_3)^\top\bar{\boldsymbol{Z}}_1)^2) &= G^{-6}\sum_{g_1,\ldots,g_6=1}^{G}\mathrm{Tr}\left\{\mathrm{E}(Z_{6g_1}Z_{6g_2}(\boldsymbol{Z}_{3g_3}-\boldsymbol{\mu}_{3g_3})^\top\boldsymbol{Z}_{1g_4}(\boldsymbol{Z}_{3g_5}-\boldsymbol{\mu}_{3g_5})^\top\boldsymbol{Z}_{1g_6})\right\}\\
&= G^{-6}\sum_{g,h,i=1}^{G}\mathrm{E}(Z_{6g}^2)\,\mathrm{Tr}\left\{\mathrm{E}((\boldsymbol{Z}_{3h}-\boldsymbol{\mu}_{3h})(\boldsymbol{Z}_{3h}-\boldsymbol{\mu}_{3h})^\top)\mathrm{E}(\boldsymbol{Z}_{1i}\boldsymbol{Z}_{1i}^\top)\right\}\\
&\quad + 2G^{-6}\sum_{g,h,i=1}^{G}\mathrm{Tr}\left\{\mathrm{E}((\boldsymbol{Z}_{3g}-\boldsymbol{\mu}_{3g})(\boldsymbol{Z}_{3g}-\boldsymbol{\mu}_{3g})^\top)\mathrm{E}(Z_{6h}\boldsymbol{Z}_{1h})\mathrm{E}(Z_{6i}\boldsymbol{Z}_{1i}^\top)\right\}+O(G^{-4})\\
&= G^{-3}\,\mathrm{Tr}\left\{(\boldsymbol{\xi}_{3,3}-\bar{\boldsymbol{\xi}}_{3,3})\boldsymbol{\xi}_{1,1}\right\}+2G^{-3}\boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,6}-2G^{-3}\boldsymbol{\xi}_{6,1}\bar{\boldsymbol{\xi}}_{3,3}\boldsymbol{\xi}_{1,6}+O(G^{-4}).
\end{aligned}$$

Next, using $\bar{\boldsymbol{\mu}}_4=\boldsymbol{\xi}_{3,3}$,

$$\begin{aligned}
\mathrm{E}(\bar{Z}_6^2\bar{\boldsymbol{Z}}_1^\top\bar{\boldsymbol{\mu}}_4\bar{\boldsymbol{Z}}_1) &= G^{-4}\sum_{g_1,\ldots,g_4=1}^{G}\mathrm{Tr}\left\{\mathrm{E}(Z_{6g_1}Z_{6g_2}\boldsymbol{Z}_{1g_3}^\top\boldsymbol{\xi}_{3,3}\boldsymbol{Z}_{1g_4})\right\}\\
&= G^{-4}\sum_{g,h=1}^{G}\mathrm{E}(Z_{6g}^2)\,\mathrm{Tr}\left\{\boldsymbol{\xi}_{3,3}\mathrm{E}(\boldsymbol{Z}_{1h}\boldsymbol{Z}_{1h}^\top)\right\}+2G^{-4}\sum_{g,h=1}^{G}\mathrm{E}(Z_{6g}\boldsymbol{Z}_{1g}^\top)\boldsymbol{\xi}_{3,3}\mathrm{E}(Z_{6h}\boldsymbol{Z}_{1h})+O(G^{-3})\\
&= G^{-2}\,\mathrm{Tr}\left\{\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,1}\right\}+2G^{-2}\boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,6}+O(G^{-3}),\ \text{and}\\
\mathrm{E}(\bar{Z}_6^4\bar{\boldsymbol{Z}}_1^\top\bar{\boldsymbol{\mu}}_4\bar{\boldsymbol{Z}}_1) &= 3G^{-6}\sum_{g,h,i=1}^{G}\mathrm{E}(Z_{6g}^2)\mathrm{E}(Z_{6h}^2)\,\mathrm{Tr}\left\{\boldsymbol{\xi}_{3,3}\mathrm{E}(\boldsymbol{Z}_{1i}\boldsymbol{Z}_{1h}^\top)\right\}\\
&\quad + 12G^{-6}\sum_{g,h,i=1}^{G}\mathrm{E}(Z_{6g}^2)\mathrm{E}(Z_{6h}\boldsymbol{Z}_{1h}^\top)\boldsymbol{\xi}_{3,3}\mathrm{E}(Z_{6i}\boldsymbol{Z}_{1i})+O(G^{-4})\\
&= 3G^{-3}\,\mathrm{Tr}\left\{\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,1}\right\}+12G^{-3}\boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,6}+O(G^{-4}).
\end{aligned}$$

For $m_1\in\{1,5,6\}$ and $m_2\in\{3,4,9,10,11,12\}$ we have $\mathrm{E}(\bar{Z}_6\bar{\boldsymbol{Z}}_{m_1}^\top(\bar{\boldsymbol{Z}}_{m_2}-\bar{\boldsymbol{\mu}}_{m_2}))=G^{-2}\,\mathrm{Tr}\{\boldsymbol{\gamma}_{m_1,m_2}-\bar{\boldsymbol{\gamma}}_{m_1,m_2}\}$ by definition and

$$\begin{aligned}
\mathrm{E}(\bar{Z}_6^3\bar{\boldsymbol{Z}}_{m_1}^\top(\bar{\boldsymbol{Z}}_{m_2}-\bar{\boldsymbol{\mu}}_{m_2})) &= G^{-5}\sum_{g_1,\ldots,g_5=1}^{G}\mathrm{E}(Z_{6g_1}Z_{6g_2}Z_{6g_3}\boldsymbol{Z}_{m_1g_5}^\top(\boldsymbol{Z}_{m_2g_4}-\boldsymbol{\mu}_{m_2g_4}))\\
&= 3G^{-5}\sum_{g,h=1}^{G}\mathrm{E}(Z_{6g}^2)\mathrm{E}(Z_{6h}\boldsymbol{Z}_{m_1h}^\top(\boldsymbol{Z}_{m_2h}-\boldsymbol{\mu}_{m_2h}))\\
&\quad + 3G^{-5}\sum_{g,h=1}^{G}\mathrm{E}(Z_{6g}\boldsymbol{Z}_{m_1g}^\top)\mathrm{E}(Z_{6h}^2(\boldsymbol{Z}_{m_2h}-\boldsymbol{\mu}_{m_2h}))+O(G^{-4})\\
&= 3G^{-4}\sum_{g=1}^{G}\mathrm{E}(Z_{6g}\boldsymbol{Z}_{m_1g}^\top(\boldsymbol{Z}_{m_2g}-\boldsymbol{\mu}_{m_2g}))\\
&\quad + 3G^{-5}\sum_{g,h=1}^{G}\mathrm{E}(Z_{6g}\boldsymbol{Z}_{m_1g}^\top)(\mathrm{E}(Z_{2h}\boldsymbol{Z}_{m_2h})-\mu_{2h}\boldsymbol{\mu}_{m_2h})+O(G^{-4})\\
&= 3G^{-3}\,\mathrm{Tr}\{\boldsymbol{\gamma}_{m_1,m_2}-\bar{\boldsymbol{\gamma}}_{m_1,m_2}\}+3G^{-3}\boldsymbol{\xi}_{6,m_1}(\boldsymbol{\xi}_{m_2,2}-\bar{\boldsymbol{\xi}}_{m_2,2})+O(G^{-4}).
\end{aligned}$$

If also $m_4\in\{2,8\}$ then, using the law of iterated expectations,

$$\mathrm{E}(\bar{Z}_6\bar{\boldsymbol{Z}}_{m_1}^\top(\bar{\boldsymbol{Z}}_{m_2}-\bar{\boldsymbol{\mu}}_{m_2})^\top(\bar{\boldsymbol{Z}}_{m_4}-\bar{\boldsymbol{\mu}}_{m_4}))$$

$$= G^{-4} \sum_{g_1,\ldots,g_4=1}^{G} \mathrm{E}(Z_{6g_1} \boldsymbol{Z}_{m_1g_2}^\top (\boldsymbol{Z}_{m_2g_3} - \boldsymbol{\mu}_{m_2g_3})^\top (\boldsymbol{Z}_{m_4g_4} - \boldsymbol{\mu}_{m_4g_4}))$$

$$= G^{-4} \sum_{g,h=1}^{G} \mathrm{E}(Z_{6g} \boldsymbol{Z}_{m_1g}^\top) \mathrm{E}((\boldsymbol{Z}_{m_2h} - \boldsymbol{\mu}_{m_2h})^\top (\boldsymbol{Z}_{m_4h} - \boldsymbol{\mu}_{m_4h})) + O(G^{-3})$$

$$= G^{-2} \boldsymbol{\xi}_{6,m_1} (\boldsymbol{\zeta}_{m_2,m_4} - \bar{\boldsymbol{\zeta}}_{m_2,m_4}) + O(G^{-3}),$$

$$\mathrm{E}(\bar{Z}_6^3 \bar{\boldsymbol{Z}}_{m_1}^\top (\bar{\boldsymbol{Z}}_{m_2} - \bar{\boldsymbol{\mu}}_{m_2})^\top (\bar{\boldsymbol{Z}}_{m_4} - \bar{\boldsymbol{\mu}}_{m_4}))$$

$$= G^{-6} \sum_{g_1,\ldots,g_6=1}^{G} \mathrm{E}(Z_{6g_1} Z_{6g_2} Z_{6g_3} \boldsymbol{Z}_{m_1g_4}^\top (\boldsymbol{Z}_{m_2g_5} - \boldsymbol{\mu}_{m_2g_5})^\top (\boldsymbol{Z}_{m_4g_6} - \boldsymbol{\mu}_{m_4g_6}))$$

$$= 3G^{-6} \sum_{g,h,i=1}^{G} \mathrm{E}(Z_{6g}^2) \mathrm{E}(Z_{6h} \boldsymbol{Z}_{m_1h}^\top) \mathrm{E}((\boldsymbol{Z}_{m_2i} - \boldsymbol{\mu}_{m_2i})^\top (\boldsymbol{Z}_{m_4i} - \boldsymbol{\mu}_{m_4i})) + O(G^{-4})$$

$$= 3G^{-3} \boldsymbol{\xi}_{6,m_1} (\boldsymbol{\zeta}_{m_2,m_4} - \bar{\boldsymbol{\zeta}}_{m_2,m_4}) + O(G^{-4}).$$

Finally, for $m_1 \in \{1, 5, 6\}$ and $m_2, m_3 \in \{3, 4, 9, 10, 11, 12\}$,

$$\mathrm{E}(\bar{Z}_6 \bar{\boldsymbol{Z}}_{m_1}^\top (\bar{\boldsymbol{Z}}_{m_2} - \bar{\boldsymbol{\mu}}_{m_2})^\top (\bar{\boldsymbol{Z}}_{m_3} - \bar{\boldsymbol{\mu}}_{m_3}))$$

$$= G^{-4} \sum_{g_1,\ldots,g_4=1}^{G} \mathrm{E}(Z_{6g_1} \boldsymbol{Z}_{m_1g_2}^\top (\boldsymbol{Z}_{m_2g_3} - \boldsymbol{\mu}_{m_2g_3})^\top (\boldsymbol{Z}_{m_3g_4} - \boldsymbol{\mu}_{m_3g_4}))$$

$$= G^{-4} \sum_{g,h=1}^{G} \mathrm{E}(Z_{6g} \boldsymbol{Z}_{m_1g}^\top) \mathrm{E}((\boldsymbol{Z}_{m_2h} - \boldsymbol{\mu}_{m_2h})^\top (\boldsymbol{Z}_{m_3h} - \boldsymbol{\mu}_{m_3h})) + O(G^{-3})$$

$$= G^{-4} \sum_{g,h=1}^{G} \mathrm{E}(Z_{6g} \boldsymbol{Z}_{m_1g}^\top) \mathrm{E}(\boldsymbol{Z}_{m_2h}^\top \boldsymbol{Z}_{m_3h}) - G^{-4} \sum_{g,h=1}^{G} \mathrm{E}(Z_{6g} \boldsymbol{Z}_{m_1g}^\top) \boldsymbol{\mu}_{m_2h}^\top \boldsymbol{\mu}_{m_3h} + O(G^{-3})$$

$$= G^{-2} \boldsymbol{\xi}_{6,m_1} (\boldsymbol{\zeta}_{m_2,m_3} - \bar{\boldsymbol{\zeta}}_{m_2,m_3}) + O(G^{-3}), \text{ and}$$

$$\mathrm{E}(\bar{Z}_6^3 \bar{\boldsymbol{Z}}_{m_1}^\top (\bar{\boldsymbol{Z}}_{m_2} - \bar{\boldsymbol{\mu}}_{m_2})^\top (\bar{\boldsymbol{Z}}_{m_3} - \bar{\boldsymbol{\mu}}_{m_3}))$$

$$= 3G^{-6} \sum_{g,h,i=1}^{G} \mathrm{E}(Z_{6g}^2) \mathrm{E}(Z_{6h} \boldsymbol{Z}_{m_1h}^\top) \mathrm{E}((\boldsymbol{Z}_{m_2h} - \boldsymbol{\mu}_{m_2h})^\top (\boldsymbol{Z}_{m_3h} - \boldsymbol{\mu}_{m_3h})) + O(G^{-4})$$

$$= 3G^{-5} \sum_{g,h=1}^{G} \mathrm{E}(Z_{6g} \boldsymbol{Z}_{m_1g}^\top) \mathrm{E}(\boldsymbol{Z}_{m_2h}^\top \boldsymbol{Z}_{m_3h}) - 3G^{-5} \sum_{g,h=1}^{G} \mathrm{E}(Z_{6g} \boldsymbol{Z}_{m_1g}^\top) \boldsymbol{\mu}_{m_2h}^\top \boldsymbol{\mu}_{m_3h} + O(G^{-4})$$

$$= 3G^{-3} \boldsymbol{\xi}_{6,m_1} (\boldsymbol{\zeta}_{m_2,m_3} - \bar{\boldsymbol{\zeta}}_{m_2,m_3}) + O(G^{-4}).$$

□

Lemma A.5. *Under the conditions of Lemma A.4 it holds that*

$$\gamma_{6,11} = \boldsymbol{\xi}_{2,3} \boldsymbol{\xi}_{1,6}, \qquad \bar{\gamma}_{6,11} = \bar{\boldsymbol{\xi}}_{2,3} \boldsymbol{\xi}_{1,6} \tag{A.2}$$

$$\zeta_{11,11} = \boldsymbol{\xi}_{6,1} \boldsymbol{\xi}_{3,3} \boldsymbol{\xi}_{1,6}, \qquad \bar{\zeta}_{11,11} = \boldsymbol{\xi}_{6,1} \bar{\boldsymbol{\xi}}_{3,3} \boldsymbol{\xi}_{1,6}, \tag{A.3}$$

$$\boldsymbol{\zeta}_{11,3} = \boldsymbol{\xi}_{3,3} \boldsymbol{\xi}_{1,6}, \qquad \bar{\boldsymbol{\zeta}}_{11,3} = \bar{\boldsymbol{\xi}}_{3,3} \boldsymbol{\xi}_{1,6}, \tag{A.4}$$

$$\zeta_{3,12} = \boldsymbol{\zeta}_{3,9} \boldsymbol{\xi}_{1,6}, \qquad \bar{\zeta}_{3,12} = \bar{\boldsymbol{\zeta}}_{3,9} \boldsymbol{\xi}_{1,6}, \tag{A.5}$$

$$\zeta_{3,10} = \mathrm{Tr}\left\{\boldsymbol{\xi}_{3,3} \boldsymbol{\xi}_{1,1}\right\}, \qquad \bar{\zeta}_{3,10} = \mathrm{Tr}\left\{\bar{\boldsymbol{\xi}}_{3,3} \boldsymbol{\xi}_{1,1}\right\}, \tag{A.6}$$

$$\zeta_{11,2} = \xi_{11,2} = \boldsymbol{\xi}_{6,1} \boldsymbol{\xi}_{3,2}, \qquad \bar{\zeta}_{11,2} = \bar{\xi}_{11,2} = \boldsymbol{\xi}_{6,1} \bar{\boldsymbol{\xi}}_{3,2}, \tag{A.7}$$

$$\zeta_{3,8} = \mathrm{Tr}\{\boldsymbol{\gamma}_{1,3}\}, \qquad \bar{\zeta}_{3,8} = \mathrm{Tr}\{\bar{\boldsymbol{\gamma}}_{1,3}\}. \tag{A.8}$$

Proof. The proofs of the results in the second column are identical to those of the results in the first column and are therefore omitted. Using $Z_{6g}\boldsymbol{Z}_{1g}^\top = \boldsymbol{a}^\top \boldsymbol{Z}_{1g}\boldsymbol{Z}_{1g}^\top$, we find that $\boldsymbol{a}^\top \boldsymbol{\xi}_{1,1} = \boldsymbol{\xi}_{6,1}$ and hence $Z_{11g} = \boldsymbol{\xi}_{6,1}\boldsymbol{Z}_{3g}$. It follows that

$$\gamma_{6,11} = G^{-1}\sum_{g=1}^{G} \mathrm{E}((\boldsymbol{a}^\top \boldsymbol{Z}_{1g})^2 \boldsymbol{Z}_{3g})\boldsymbol{\xi}_{1,6} = \boldsymbol{\xi}_{2,3}\boldsymbol{\xi}_{1,6},$$

$$\zeta_{11,11} = G^{-1}\sum_{g=1}^{G} \boldsymbol{\xi}_{6,1}\mathrm{E}(\boldsymbol{Z}_{3g}\boldsymbol{Z}_{3g}^\top)\boldsymbol{\xi}_{1,6} = \boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,6},$$

$$\boldsymbol{\zeta}_{11,3} = G^{-1}\sum_{g=1}^{G} \mathrm{E}(\boldsymbol{Z}_{3g}^\top \boldsymbol{\xi}_{1,6}\boldsymbol{Z}_{3g}) = G^{-1}\sum_{g=1}^{G} \mathrm{E}(\boldsymbol{Z}_{3g}\boldsymbol{Z}_{3g}^\top)\boldsymbol{\xi}_{1,6} = \boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,6}, \text{ and}$$

$$\xi_{11,2} = \boldsymbol{\xi}_{6,1} G^{-1}\sum_{g=1}^{G} \mathrm{E}(\boldsymbol{Z}_{3g}Z_{2g}) = \boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,2}.$$

We also notice that $\boldsymbol{Z}_{12g} = \boldsymbol{Z}_{9g}\boldsymbol{\xi}_{1,1}\boldsymbol{a} = \boldsymbol{Z}_{9g}\boldsymbol{\xi}_{1,6}$ and $\boldsymbol{Z}_{10g} = \boldsymbol{\xi}_{1,1}\boldsymbol{Z}_{3g}$, so that

$$\zeta_{3,12} = G^{-1}\sum_{g=1}^{G} \mathrm{E}(\boldsymbol{Z}_{3g}^\top \boldsymbol{Z}_{12g}) = \zeta_{3,9}\boldsymbol{\xi}_{1,6}, \text{ and}$$

$$\zeta_{3,10} = G^{-1}\sum_{g=1}^{G} \mathrm{Tr}\left\{\mathrm{E}(\boldsymbol{Z}_{3g}^\top \boldsymbol{\xi}_{1,1}\boldsymbol{Z}_{3g})\right\} = \mathrm{Tr}\left\{\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,1}\right\}.$$

Finally, $\boldsymbol{Z}_{8g} = \boldsymbol{Z}_{7g}\boldsymbol{a} = \boldsymbol{Z}_{1g}\boldsymbol{Z}_{1g}^\top\boldsymbol{a}$, which implies (A.8) by the definition of $\boldsymbol{\gamma}_{m,n}$. □

Appendix B: Proofs of Main Results

B.1 Proof of Theorem 2.1

Proof of (16). The left-hand side of (16) is

$$(\boldsymbol{a}^\top \boldsymbol{V}\boldsymbol{a})^{-1/2}\boldsymbol{a}^\top \boldsymbol{Q}^{-1}N^{-1}\sum_{g=1}^{G} \boldsymbol{s}_g = v_a^{-1/2}\mu_N^{1/2}\boldsymbol{a}^\top \boldsymbol{\Xi}_0^{-1}N^{-1}\sum_{g=1}^{G} \boldsymbol{s}_g(1+o_P(1))$$

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

$$v_a^{-1/2}\mu_N^{1/2}\boldsymbol{a}^\top \boldsymbol{\Xi}_0^{-1}\frac{1}{N}\sum_{g=1}^{G} \boldsymbol{s}_g \xrightarrow{d} \mathrm{N}(0,1). \tag{B.1}$$

We define $z_g = v_a^{-1/2}\mu_N^{1/2}N^{-1}\boldsymbol{a}^\top \boldsymbol{\Xi}_0^{-1}\boldsymbol{s}_g$, which, by Assumption 1, is an independent sequence with mean zero and variance given by $\mathrm{E}(z_g^2) = v_a^{-1}\mu_N N^{-2}\boldsymbol{a}^\top \boldsymbol{\Xi}_0^{-1}\boldsymbol{\Sigma}_g\boldsymbol{\Xi}_0^{-1}\boldsymbol{a}$. By Assumption 2, $\sum_{g=1}^{G}\mathrm{E}(z_g^2) \to 1$, and 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

$$\begin{aligned}\sum_{g=1}^{G}\mathrm{E}|z_g|^{2+\xi} &\le v_a^{-1-\xi/2}\mu_N^{1+\xi/2}\|\boldsymbol{a}^\top \boldsymbol{\Xi}_0^{-1}\|^{2+\xi}N^{-2-\xi}\sum_{g=1}^{G}\mathrm{E}\|\boldsymbol{s}_g\|^{2+\xi} \\ &\le C\mu_N^{1+\xi/2}N^{-2-\xi}\sum_{g=1}^{G} N_g^{2+\xi} \le C\mu_N^{1+\xi/2}N^{-1-\xi}\sup_{g\in\mathbb{N}} N_g^{1+\xi} \to 0,\end{aligned} \tag{B.2}$$

where the second inequality is due to positive definiteness of $\boldsymbol{\Xi}_0$ (Assumption 2) and Lemma A.2 (with $\theta = \xi + 2$), and the convergence is due to Assumption 3 setting $\xi = \lambda$.

Proof of (17). Because $d \to 1$ as $G \to \infty$, we can proceed as if $d = 1$ in this proof without any loss of generality. We first define $\boldsymbol{V}_0 = \boldsymbol{\Xi}_0^{-1} N^{-2} \sum_{g=1}^G \boldsymbol{\Sigma}_g \boldsymbol{\Xi}_0^{-1}$ and apply the decomposition

$$\frac{\boldsymbol{a}^\top \hat{\boldsymbol{V}} \boldsymbol{a}}{\boldsymbol{a}^\top \boldsymbol{V} \boldsymbol{a}} - 1 = (\boldsymbol{a}^\top \boldsymbol{V} \boldsymbol{a})^{-1} \boldsymbol{a}^\top (\hat{\boldsymbol{V}} - \boldsymbol{V}) \boldsymbol{a} = (\boldsymbol{a}^\top \boldsymbol{V}_0 \boldsymbol{a})^{-1} \boldsymbol{a}^\top \Big(\boldsymbol{A}_1 - \boldsymbol{A}_2 - \boldsymbol{A}_2^\top + \boldsymbol{A}_3 \Big) \boldsymbol{a} (1 + o_P(1)),$$

where we used Assumption 2 and

$$\begin{aligned}
\boldsymbol{A}_1 &= \frac{1}{N^2} \boldsymbol{\Xi}_0^{-1} \sum_{g=1}^G \boldsymbol{s}_g \boldsymbol{s}_g^\top \boldsymbol{\Xi}_0^{-1} - \frac{1}{N^2} \boldsymbol{\Xi}_0^{-1} \sum_{g=1}^G \boldsymbol{\Sigma}_g \boldsymbol{\Xi}_0^{-1}, \\
\boldsymbol{A}_2 &= \frac{1}{N^2} \boldsymbol{\Xi}_0^{-1} \sum_{g=1}^G \boldsymbol{s}_g (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)^\top \boldsymbol{X}_g^\top \boldsymbol{X}_g \boldsymbol{\Xi}_0^{-1}, \text{ and} \\
\boldsymbol{A}_3 &= \frac{1}{N^2} \boldsymbol{\Xi}_0^{-1} \sum_{g=1}^G \boldsymbol{X}_g^\top \boldsymbol{X}_g (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)^\top \boldsymbol{X}_g^\top \boldsymbol{X}_g \boldsymbol{\Xi}_0^{-1}.
\end{aligned}$$

Thus, we need to show that $(\boldsymbol{a}^\top \boldsymbol{V}_0 \boldsymbol{a})^{-1} \boldsymbol{a}^\top \boldsymbol{A}_m \boldsymbol{a} \xrightarrow{P} 0$, or equivalently $\mu_N \boldsymbol{a}^\top \boldsymbol{A}_m \boldsymbol{a} \xrightarrow{P} 0$, for $m = 1, 2, 3$.

To prove the result for $m = 1$, we use a truncation argument. Let $r_g = N^{-1} (\boldsymbol{a}^\top \boldsymbol{V}_0 \boldsymbol{a})^{-1/2} \boldsymbol{a}^\top \boldsymbol{\Xi}_0^{-1} \boldsymbol{s}_g$, such that $(\boldsymbol{a}^\top \boldsymbol{V}_0 \boldsymbol{a})^{-1} \boldsymbol{a}^\top \boldsymbol{A}_1 \boldsymbol{a} = \sum_{g=1}^G r_g^2 - 1$ has mean zero. Also define the truncated variable $q_g = r_g \mathbb{I}(|r_g| \leq \epsilon)$ such that $r_g^2 = q_g^2 + r_g^2 \mathbb{I}(|r_g| > \epsilon)$. By the triangle inequality,

$$\mathrm{E} \Big| \sum_{g=1}^G r_g^2 - 1 \Big| \leq \mathrm{E} \Big| \sum_{g=1}^G \Big(q_g^2 - \mathrm{E}(q_g^2) \Big) \Big| + \mathrm{E} \Big| \sum_{g=1}^G \Big(r_g^2 \mathbb{I}(|r_g| > \epsilon) - \mathrm{E}(r_g^2 \mathbb{I}(|r_g| > \epsilon)) \Big) \Big|. \tag{B.3}$$

The second term satisfies

$$\begin{aligned}
\mathrm{E} \Big| \sum_{g=1}^G \Big(r_g^2 \mathbb{I}(|r_g| > \epsilon) - \mathrm{E}(r_g^2 \mathbb{I}(|r_g| > \epsilon)) \Big) \Big| &\leq 2 \sum_{g=1}^G \mathrm{E} \Big(|r_g|^{2+\lambda} |r_g|^{-\lambda} \mathbb{I}(|r_g| > \epsilon) \Big) \\
&\leq 2 \epsilon^{-\lambda} \sum_{g=1}^G \mathrm{E} |r_g|^{2+\lambda} \leq C \mu_N^{1+\lambda/2} N^{-1-\lambda} \sup_{g \in \mathbb{N}} N_g^{1+\lambda} \to 0,
\end{aligned}$$

where the last inequality uses Assumption 2 and Lemma A.2, and the convergence is due to Assumption 3. To show that the first term of (B.3) is negligible, we use Jensen's inequality and show convergence in mean-square, noting that the truncated variable q_g has all moments finite. That is,

$$\mathrm{Var} \left(\sum_{g=1}^G q_g^2 \right) = \sum_{g=1}^G \mathrm{Var}(q_g^2) \leq \epsilon^2 \sum_{g=1}^G \mathrm{Var}(|q_g|) \leq \epsilon^2 \sum_{g=1}^G \mathrm{E}(q_g^2) \leq \epsilon^2 \sum_{g=1}^G \mathrm{E}(r_g^2) = \epsilon^2,$$

where the last inequality is because $\mathrm{E}(q_g^2) = \mathrm{E}(r_g^2 \mathbb{I}(|r_g| \leq \epsilon) \leq \mathrm{E}(r_g^2)$ and the last equality holds because $\sum_{g=1}^G \mathrm{E}(r_g^2) = 1$. This proves the result for $m = 1$ since ϵ is arbitrary.

Next, we analyze the case $m = 2$. By the Cauchy-Schwarz inequality,

$$\begin{aligned}
\Big| (\boldsymbol{a}^\top \boldsymbol{V}_0 \boldsymbol{a})^{-1} \boldsymbol{a}^\top \boldsymbol{A}_2 \boldsymbol{a} \Big| &\leq \left((\boldsymbol{a}^\top \boldsymbol{V}_0 \boldsymbol{a})^{-1} \frac{1}{N^2} \boldsymbol{a}^\top \boldsymbol{\Xi}_0^{-1} \sum_{g=1}^G \boldsymbol{s}_g \boldsymbol{s}_g^\top \boldsymbol{\Xi}_0^{-1} \boldsymbol{a} \right)^{1/2} \Big((\boldsymbol{a}^\top \boldsymbol{V}_0 \boldsymbol{a})^{-1} \boldsymbol{a}^\top \boldsymbol{A}_3 \boldsymbol{a} \Big)^{1/2} \\
&= \Big(1 + (\boldsymbol{a}^\top \boldsymbol{V}_0 \boldsymbol{a})^{-1} \boldsymbol{a}^\top \boldsymbol{A}_1 \boldsymbol{a} \Big)^{1/2} \Big((\boldsymbol{a}^\top \boldsymbol{V}_0 \boldsymbol{a})^{-1} \boldsymbol{a}^\top \boldsymbol{A}_3 \boldsymbol{a} \Big)^{1/2},
\end{aligned}$$

such that the result for $m = 2$ follows by proving the results for $m = 1$ and $m = 3$.

Finally, for $m = 3$ we first find the bound

$$\|\mu_N \boldsymbol{a}^\top \boldsymbol{A}_3 \boldsymbol{a}\| \leq \mu_N \frac{1}{N^2} \|\boldsymbol{\Xi}_0^{-1}\|^2 \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N\|^2 \sum_{g=1}^{G} \|\boldsymbol{X}_g^\top \boldsymbol{X}_g\|^2.$$

Here we note that $\sum_{g=1}^{G} \|\boldsymbol{X}_g^\top \boldsymbol{X}_g\|^2 = O_P(\sum_{g=1}^{G} N_g^2) = O_P(N \sup_{g\in\mathbb{N}} N_g)$ by Lemma A.2 and $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N\| = O_P(\|\boldsymbol{V}\|^{1/2}) = O_P(N^{-1/2} \sup_{g\in\mathbb{N}} N_g^{1/2})$; see (9). It follows that

$$\|\mu_N \boldsymbol{a}^\top \boldsymbol{A}_3 \boldsymbol{a}\| = O_P\Big(\mu_N N^{-2} \sup_{g\in\mathbb{N}} N_g^2\Big) = o_P(1).$$

Proof of (18). We use (15) to decompose the t-statistic (6) as

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

and the result then follows directly from (16), (17), and Slutsky's Theorem.

B.2 Proof of Theorem 3.1

Because the bootstrap is exactly invariant to the multiplicative factor d, we can, without loss of generality, set $d = 1$ in this proof. 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 $\epsilon > 0$, we want to show that

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

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

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

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 define the score vectors $\ddot{\boldsymbol{s}}_g = \boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g$, and let $\ddot{\boldsymbol{\Gamma}} = N^{-2} \sum_{g=1}^{G} \ddot{\boldsymbol{s}}_g \ddot{\boldsymbol{s}}_g^\top = N^{-2} \sum_{g=1}^{G} \boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g \ddot{\boldsymbol{u}}_g^\top \boldsymbol{X}_g$ and $\ddot{\boldsymbol{V}} = \boldsymbol{Q}^{-1} \ddot{\boldsymbol{\Gamma}} \boldsymbol{Q}^{-1}$ 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 (15),

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

It follows from (B.7) that $\boldsymbol{a}^\top(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_N) = O_P(\mu_N^{-1/2})$. 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 the bootstrap score vectors $\boldsymbol{s}_g^* = \boldsymbol{X}_g^\top \boldsymbol{u}_g^* = \boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g v_g^*$ and the scalar random variables $z_g^* = (\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-1/2} \boldsymbol{a}^\top \boldsymbol{Q}^{-1} N^{-1} \boldsymbol{s}_g^*$ 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 $x \in \mathbb{R}$,

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

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 $\sum_{g=1}^G \mathrm{E}^* |z_g^*|^{2+\xi} \xrightarrow{P} 0$ for some $\xi > 0$.

We first find that, because $\|\boldsymbol{X}_g^\top \boldsymbol{X}_g\| = O_P(N_g)$ and $\|\boldsymbol{s}_g\| = O_P(N_g)$ by Lemma A.2,

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

Note that (B.10) only requires that Assumptions 1 and 2 hold for some $\lambda \geq 0$, i.e. with two moments. 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,

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

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

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

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

Proof of (B.5). We note that $\hat{\boldsymbol{s}}_g^* = \boldsymbol{X}_g^\top \hat{\boldsymbol{u}}_g^* = \boldsymbol{s}_g^* - \boldsymbol{X}_g^\top \boldsymbol{X}_g (\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})$, which implies the decomposition

$$(\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-1} \boldsymbol{a}^\top (\hat{\boldsymbol{V}}^* - \ddot{\boldsymbol{V}}) \boldsymbol{a} = (\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-1} \boldsymbol{a}^\top \Big(\boldsymbol{B}_1^* - \boldsymbol{B}_2^* - \boldsymbol{B}_2^{*\top} + \boldsymbol{B}_3^* \Big) \boldsymbol{a},$$

where, using also $\boldsymbol{s}_g^* = \ddot{\boldsymbol{s}}_g v_g^*$,

$$\begin{aligned} \boldsymbol{B}_1^* &= \boldsymbol{Q}^{-1} \frac{1}{N^2} \sum_{g=1}^G \ddot{\boldsymbol{s}}_g \ddot{\boldsymbol{s}}_g^\top \boldsymbol{Q}^{-1} (v_g^{*2} - 1), \\ \boldsymbol{B}_2^* &= \boldsymbol{Q}^{-1} \frac{1}{N^2} \sum_{g=1}^G \boldsymbol{s}_g^* (\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})^\top \boldsymbol{X}_g^\top \boldsymbol{X}_g \boldsymbol{Q}^{-1}, \text{ and} \\ \boldsymbol{B}_3^* &= \boldsymbol{Q}^{-1} \frac{1}{N^2} \sum_{g=1}^G \boldsymbol{X}_g^\top \boldsymbol{X}_g (\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}}) (\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}})^\top \boldsymbol{X}_g^\top \boldsymbol{X}_g \boldsymbol{Q}^{-1}. \end{aligned}$$

With this decomposition, it suffices to prove that, for any $\epsilon > 0$, $P^*(|(\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-1} \boldsymbol{a}^\top \boldsymbol{B}_m^* \boldsymbol{a}| > \epsilon) \xrightarrow{P} 0$ for $m = 1, 2, 3$. The proofs for each term roughly follow those for the corresponding term in the proof of (17).

For $m = 1$, we use the truncation argument and define $r_g^* = N^{-1}(\boldsymbol{a}^\top \ddot{\boldsymbol{V}}\boldsymbol{a})^{-1/2}\boldsymbol{a}^\top \boldsymbol{Q}^{-1}\ddot{\boldsymbol{s}}_g v_g^*$ such that $(\boldsymbol{a}^\top \ddot{\boldsymbol{V}}\boldsymbol{a})^{-1}\boldsymbol{a}^\top \boldsymbol{B}_1^*\boldsymbol{a} = \sum_{g=1}^G r_g^{*2} - 1$ satisfies $\sum_{g=1}^G \mathrm{E}^*(r_g^{*2}) - 1 = 0$ because $\mathrm{E}^*(v_g^{*2}) = 1$. We then decompose $\mathrm{E}^*\big|\sum_{g=1}^G r_g^{*2} - 1\big|$ as in (B.3), where for each term we use the same arguments as for (B.3). For example, for the second term we apply the bound

$$\begin{aligned}\mathrm{E}^*\bigg|\sum_{g=1}^G \Big(r_g^{*2}\mathbb{I}(|r_g^*| > \epsilon) - \mathrm{E}(r_g^{*2}\mathbb{I}(|r_g^*| > \epsilon))\Big)\bigg| &\leq 2\epsilon^{-\lambda}\sum_{g=1}^G \mathrm{E}^*|r_g^*|^{2+\lambda} \\ &= O_P\Big(\mu_N^{1+\lambda/2}N^{-1-\lambda}\sup_{g\in\mathbb{N}} N_g^{1+\lambda}\Big) = o_P(1),\end{aligned}$$

using, in particular, (B.11) and Assumptions 2 and 3.

Next, for $m = 2$, we apply the Cauchy-Schwarz inequality to obtain the bound

$$\begin{aligned}\Big|(\boldsymbol{a}^\top \ddot{\boldsymbol{V}}\boldsymbol{a})^{-1}\boldsymbol{a}^\top \boldsymbol{B}_2^*\boldsymbol{a}\Big| &\leq \bigg((\boldsymbol{a}^\top \ddot{\boldsymbol{V}}\boldsymbol{a})^{-1}\boldsymbol{a}^\top \boldsymbol{Q}^{-1}\frac{1}{N^2}\sum_{g=1}^G \boldsymbol{s}_g^*\boldsymbol{s}_g^{*\top}\boldsymbol{Q}^{-1}\boldsymbol{a}\bigg)^{1/2}\Big((\boldsymbol{a}^\top \ddot{\boldsymbol{V}}\boldsymbol{a})^{-1}\boldsymbol{a}^\top \boldsymbol{B}_3^*\boldsymbol{a}\Big)^{1/2} \\ &= \Big(1 + (\boldsymbol{a}^\top \ddot{\boldsymbol{V}}\boldsymbol{a})^{-1}\boldsymbol{a}^\top \boldsymbol{B}_1^*\boldsymbol{a}\Big)^{1/2}\Big((\boldsymbol{a}^\top \ddot{\boldsymbol{V}}\boldsymbol{a})^{-1}\boldsymbol{a}^\top \boldsymbol{B}_3^*\boldsymbol{a}\Big)^{1/2}, \qquad \text{(B.13)}\end{aligned}$$

such that the result for $m = 2$ follows by proving the results for $m = 1$ and $m = 3$.

Finally, to prove the result for $m = 3$, we first note that $\mathrm{E}^*\|\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}}\|^2 = O_P(\|\ddot{\boldsymbol{V}}\|) = O_P(N^{-1}\sup_{g\in\mathbb{N}} N_g)$. Then we apply Markov's inequality,

$$\begin{aligned}P^*\Big(\big|(\boldsymbol{a}^\top \ddot{\boldsymbol{V}}\boldsymbol{a})^{-1}\boldsymbol{a}^\top \boldsymbol{B}_3^*\boldsymbol{a}\big| > \epsilon\Big) &\leq \epsilon^{-1}(\boldsymbol{a}^\top \ddot{\boldsymbol{V}}\boldsymbol{a})^{-1}\|\boldsymbol{Q}^{-1}\|^2\mathrm{E}^*\|\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}}\|^2\frac{1}{N^2}\sum_{g=1}^G \|\boldsymbol{X}_g^\top \boldsymbol{X}_g\|^2 \\ &= O_P\Big(\mu_N N^{-2}\sup_{g\in\mathbb{N}} N_g^2\Big) = o_P(1), \qquad \text{(B.14)}\end{aligned}$$

using also the last term of (B.8) together with Assumption 2 and Lemma A.2.

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

B.3 Proof of Theorem 3.2

Because the bootstrap is exactly invariant to the multiplicative factor d, we can, without loss of generality, set $d = 1$ in this proof. We first define some notation. Let $\bar{\boldsymbol{\Sigma}}_g = \sum_{i=1}^{N_g}\mathrm{E}(\boldsymbol{s}_{ig}\boldsymbol{s}_{ig}^\top)$ denote the variance matrix of the scores obtained by setting all the covariances between $\boldsymbol{s}_{ig}$ and $\boldsymbol{s}_{jg}$ to zero for $i \neq j$, let $\bar{\boldsymbol{\Gamma}} = N^{-2}\sum_{g=1}^G \bar{\boldsymbol{\Sigma}}_g$ and $\bar{\boldsymbol{V}} = \boldsymbol{Q}^{-1}\bar{\boldsymbol{\Gamma}}\boldsymbol{Q}^{-1}$; cf. (2), (4), and Assumption 2. Notice that, except in very special cases, $\bar{\boldsymbol{V}} \neq \boldsymbol{V}$. We also let $\ddot{\boldsymbol{V}} = \boldsymbol{Q}^{-1}\ddot{\boldsymbol{\Gamma}}\boldsymbol{Q}^{-1}$ and $\ddot{\boldsymbol{\Gamma}} = N^{-2}\sum_{g=1}^G\sum_{i=1}^{N_g}\ddot{\boldsymbol{s}}_{ig}\ddot{\boldsymbol{s}}_{ig}^\top$ 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 $\epsilon > 0$,

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

$$P^*\bigg(\bigg|\frac{\boldsymbol{a}^\top \hat{\boldsymbol{V}}^*\boldsymbol{a}}{\boldsymbol{a}^\top \bar{\boldsymbol{V}}\boldsymbol{a}} - 1\bigg| > \epsilon\bigg) \xrightarrow{P} 0, \qquad \text{(B.16)}$$

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

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

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

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

$$\boldsymbol{a}^\top (\ddot{\boldsymbol{V}} - \bar{\boldsymbol{V}}) \boldsymbol{a} = \boldsymbol{a}^\top \Big(\boldsymbol{C}_1 - \boldsymbol{C}_2 - \boldsymbol{C}_2^\top + \boldsymbol{C}_3 \Big) \boldsymbol{a},$$

where

$$\boldsymbol{C}_1 = \boldsymbol{Q}^{-1} \frac{1}{N^2} \sum_{g=1}^{G} \sum_{i=1}^{N_g} (\boldsymbol{s}_{ig} \boldsymbol{s}_{ig}^\top - \mathrm{E}(\boldsymbol{s}_{ig} \boldsymbol{s}_{ig}^\top)) \boldsymbol{Q}^{-1},$$

$$\boldsymbol{C}_2 = \boldsymbol{Q}^{-1} \frac{1}{N^2} \sum_{g=1}^{G} \sum_{i=1}^{N_g} \boldsymbol{s}_{ig} (\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)^\top \boldsymbol{X}_{ig}^\top \boldsymbol{X}_{ig} \boldsymbol{Q}^{-1}, \text{ and}$$

$$\boldsymbol{C}_3 = \boldsymbol{Q}^{-1} \frac{1}{N^2} \sum_{g=1}^{G} \sum_{i=1}^{N_g} \boldsymbol{X}_{ig}^\top \boldsymbol{X}_{ig} (\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)^\top \boldsymbol{X}_{ig}^\top \boldsymbol{X}_{ig} \boldsymbol{Q}^{-1},$$

and show that $(\boldsymbol{a}^\top \bar{\boldsymbol{V}} \boldsymbol{a})^{-1} \boldsymbol{a}^\top \boldsymbol{C}_m \boldsymbol{a} \xrightarrow{P} 0$ for $m = 1, \ldots, 3$. Equivalently, since $(\boldsymbol{a}^\top \bar{\boldsymbol{V}} \boldsymbol{a})^{-1} = O_P(N)$, we show that $N \boldsymbol{a}^\top \boldsymbol{C}_m \boldsymbol{a} \xrightarrow{P} 0$ for $m = 1, \ldots, 3$.

To prove the result for $m = 1$, for any conforming vector $\boldsymbol{b}$, let $w_{ig} = \boldsymbol{b}^\top (\boldsymbol{s}_{ig} \boldsymbol{s}_{ig}^\top - \mathrm{E}(\boldsymbol{s}_{ig} \boldsymbol{s}_{ig}^\top)) \boldsymbol{b}$, which is independent across g and mean zero with $\mathrm{E}|w_{ig}|^{1+\lambda/2} < \infty$. Hence, by Lemma A.1, $\sum_{g=1}^{G} \sum_{i=1}^{N_g} w_{ig} = O_P(N^{\max\{1/(1+\lambda/2), 1/2\}})$ such that $|N \boldsymbol{a}^\top \boldsymbol{C}_1 \boldsymbol{a}| = O_P(N^{\max\{1/(1+\lambda/2), 1/2\} - 1}) = o_P(1)$ by Assumption 2 and because $\lambda > 0$.

For $m = 2$, we apply the bound

$$|N \boldsymbol{a}^\top \boldsymbol{C}_2 \boldsymbol{a}| \leq N \|\boldsymbol{Q}^{-1}\|^2 \|\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_N\| \frac{1}{N^2} \sum_{g=1}^{G} \sum_{i=1}^{N_g} \|\boldsymbol{X}_{ig}^\top \boldsymbol{X}_{ig}\| \|\boldsymbol{s}_{ig}\| = O_P\Big(N^{-1/2} \sup_{g \in \mathbb{N}} N_g^{1/2} \Big) = o_P(1),$$

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

$$|N \boldsymbol{a}^\top \boldsymbol{C}_3 \boldsymbol{a}| \leq N \|\boldsymbol{Q}^{-1}\|^2 \|\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_N\|^2 \frac{1}{N^2} \sum_{g=1}^{G} \sum_{i=1}^{N_g} \|\boldsymbol{X}_{ig}^\top \boldsymbol{X}_{ig}\|^2 = O_P\Big(N^{-1} \sup_{g \in \mathbb{N}} N_g \Big) = o_P(1).$$

Proof of (B.15). We have $(\boldsymbol{a}^\top \bar{\boldsymbol{V}} \boldsymbol{a})^{-1/2} \boldsymbol{a}^\top (\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}}) = (\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-1/2} (1 + o_P(1)) \boldsymbol{a}^\top \boldsymbol{Q}^{-1} N^{-1} \sum_{g=1}^G \sum_{i=1}^{N_g} \boldsymbol{s}_{ig}^*$ by (B.19). Under the WB probability measure, $\boldsymbol{s}_{ig}^*$ is heteroskedastic, but independent across both i and g. Let $z_{ig}^* = (\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-1/2} \boldsymbol{a}^\top \boldsymbol{Q}^{-1} N^{-1} \boldsymbol{s}_{ig}^*$, 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 $z_{1ig}^* = (\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-1/2} \boldsymbol{a}^\top \boldsymbol{Q}^{-1} N^{-1} \boldsymbol{s}_{ig} v_{ig}^*$ and $z_{2ig}^* = (\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-1/2} \boldsymbol{a}^\top \boldsymbol{Q}^{-1} N^{-1} \boldsymbol{X}_{ig}^\top \boldsymbol{X}_{ig} (\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_N) v_{ig}^*$. We first obtain the bound

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

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

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

which is $O(N)$ by Assumption 1. It follows, using also (B.19) and Assumption 2, 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 $0 < \xi \leq \lambda$. Next, by (B.8) and (B.19),

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

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, 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.16) follows if, for any $\epsilon > 0$, $P^*(|(\boldsymbol{a}^\top \ddot{\boldsymbol{V}} \boldsymbol{a})^{-1} \boldsymbol{a}^\top \hat{\boldsymbol{V}}^* \boldsymbol{a} - 1| > \epsilon) \xrightarrow{P} 0$. To prove this, we apply the decomposition

$$\boldsymbol{a}^\top (\hat{\boldsymbol{V}}^* - \ddot{\boldsymbol{V}}) \boldsymbol{a} = \boldsymbol{a}^\top \Big(\boldsymbol{D}_1^* + \boldsymbol{D}_2^* - \boldsymbol{D}_3^* - \boldsymbol{D}_3^{*\top} + \boldsymbol{D}_4^* \Big) \boldsymbol{a},$$

where

$$\boldsymbol{D}_1^* = \boldsymbol{Q}^{-1}\frac{1}{N^2}\sum_{g=1}^{G}\sum_{i=1}^{N_g}\ddot{\boldsymbol{s}}_{ig}\ddot{\boldsymbol{s}}_{ig}^\top\boldsymbol{Q}^{-1}(v_{ig}^{*2}-1),$$

$$\boldsymbol{D}_2^* = \boldsymbol{Q}^{-1}\frac{1}{N^2}\sum_{g=1}^{G}\sum_{i\neq j=1}^{N_g}\ddot{\boldsymbol{s}}_{ig}\ddot{\boldsymbol{s}}_{jg}^\top\boldsymbol{Q}^{-1}v_{ig}^*v_{jg}^*,$$

$$\boldsymbol{D}_3^* = \boldsymbol{Q}^{-1}\frac{1}{N^2}\sum_{g=1}^{G}\boldsymbol{s}_g^*(\hat{\boldsymbol{\beta}}^*-\ddot{\boldsymbol{\beta}})^\top\boldsymbol{X}_g^\top\boldsymbol{X}_g\boldsymbol{Q}^{-1}, \text{ and}$$

$$\boldsymbol{D}_4^* = \boldsymbol{Q}^{-1}\frac{1}{N^2}\sum_{g=1}^{G}\boldsymbol{X}_g^\top\boldsymbol{X}_g(\hat{\boldsymbol{\beta}}^*-\ddot{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^*-\ddot{\boldsymbol{\beta}})^\top\boldsymbol{X}_g^\top\boldsymbol{X}_g\boldsymbol{Q}^{-1}.$$

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

To prove the result for $m=1$, we define $w_{ig}^*=z_{ig}^{*2}-\mathrm{E}^*(z_{ig}^{*2})$, where z_{ig}^* is defined in the proof of (B.15), such that $(\boldsymbol{a}^\top\ddot{\boldsymbol{V}}\boldsymbol{a})^{-1}\boldsymbol{a}^\top\boldsymbol{D}_1^*\boldsymbol{a}=\sum_{g=1}^{G}\sum_{i=1}^{N_g}w_{ig}^*$. We let $0<\xi\leq\max\{\lambda,2\}$ and apply the von Bahr-Esseen and Jensen inequalities,

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

which is $o_P(1)$ by the Lyapunov condition in the proof of (B.15). This proves the result for $m=1$.

For $m=2$, we note that $(\boldsymbol{a}^\top\ddot{\boldsymbol{V}}\boldsymbol{a})^{-1}\boldsymbol{a}^\top\boldsymbol{D}_2^*\boldsymbol{a}=\sum_{g=1}^{G}\sum_{i\neq j=1}^{N_g}z_{ig}^*z_{jg}^*$, and we prove convergence in mean-square. By independence of z_{ig}^* across both g and i (under the WB probability measure),

$$\mathrm{E}^*\big((\boldsymbol{a}^\top\ddot{\boldsymbol{V}}\boldsymbol{a})^{-1}\boldsymbol{a}^\top\boldsymbol{D}_2^*\boldsymbol{a}\big)^2\leq C\sum_{g=1}^{G}\sum_{i,j=1}^{N_g}\mathrm{E}^*(z_{ig}^{*2})\mathrm{E}^*(z_{jg}^{*2})=O_P(N^{-1}\sup_{g\in\mathbb{N}}N_g),$$

which is $o_P(1)$ by (10) and where we used again the Lyapunov condition from the proof of (B.15).

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

$$\Big|(\boldsymbol{a}^\top\ddot{\boldsymbol{V}}\boldsymbol{a})^{-1}\boldsymbol{a}^\top\boldsymbol{D}_3^*\boldsymbol{a}\Big|\leq\Big(1+(\boldsymbol{a}^\top\ddot{\boldsymbol{V}}\boldsymbol{a})^{-1}\boldsymbol{a}^\top\boldsymbol{D}_1^*\boldsymbol{a}+(\boldsymbol{a}^\top\ddot{\boldsymbol{V}}\boldsymbol{a})^{-1}\boldsymbol{a}^\top\boldsymbol{D}_2^*\boldsymbol{a}\Big)^{1/2}\Big((\boldsymbol{a}^\top\ddot{\boldsymbol{V}}\boldsymbol{a})^{-1}\boldsymbol{a}^\top\boldsymbol{D}_4^*\boldsymbol{a}\Big)^{1/2},$$

such that the result for $m=3$ follows by proving the results for $m=1,2,4$.

Finally, for $m=4$ we apply Markov's inequality as in (B.14) and find

$$P^*\Big(\big|N\boldsymbol{a}^\top\boldsymbol{D}_4^*\boldsymbol{a}\big|>\epsilon\Big)\leq\epsilon^{-1}N\mathrm{E}^*\|\hat{\boldsymbol{\beta}}^*-\ddot{\boldsymbol{\beta}}\|^2\|\boldsymbol{Q}^{-1}\|^2\frac{1}{N^2}\sum_{g=1}^{G}\|\boldsymbol{X}_g\boldsymbol{X}_g^\top\|^2=O_P\Big(N^{-1}\sup_{g\in\mathbb{N}}N_g\Big)$$

because $\mathrm{E}^*\|\hat{\boldsymbol{\beta}}^*-\ddot{\boldsymbol{\beta}}\|^2=O_P(1)$ under the WB probability measure and using Assumption 2 and Lemma A.2. The result for $m=4$ follows because $N^{-1}\sup_{g\in\mathbb{N}}N_g\to 0$ by (10).

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

B.4 Proof of Theorem 5.1

We apply the smooth function model of Bhattacharya and Ghosh (1978) and particularly Thm. 3.2 of Skovgaard (1981); see also Ch. 2 of Hall (1992) for a textbook treatment. We use (28) and $\hat{\boldsymbol{u}}_g = \boldsymbol{u}_g - \boldsymbol{X}_g(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ to write the sample t-statistic as

$$\begin{aligned} t_a &= d^{-1/2}\bigg(\frac{1}{G}\sum_{g=1}^{G} \boldsymbol{a}^\top \bar{\boldsymbol{W}}_3^{-1}(\boldsymbol{X}_g^\top \hat{\boldsymbol{u}}_g \hat{\boldsymbol{u}}_g^\top \boldsymbol{X}_g)\bar{\boldsymbol{W}}_3^{-1}\boldsymbol{a}\bigg)^{-1/2} \sqrt{G}\boldsymbol{a}^\top \bar{\boldsymbol{W}}_3^{-1}\bar{\boldsymbol{W}}_1 \\ &= d^{-1/2}\Big(\boldsymbol{a}^\top \bar{\boldsymbol{W}}_3^{-1}\bar{\boldsymbol{W}}_2\bar{\boldsymbol{W}}_3^{-1}\boldsymbol{a} + (\boldsymbol{a}^\top \bar{\boldsymbol{W}}_3^{-1} \otimes \boldsymbol{a}^\top \bar{\boldsymbol{W}}_3^{-1})\bar{\boldsymbol{W}}_4 \operatorname{vec}(\bar{\boldsymbol{W}}_3^{-1}\bar{\boldsymbol{W}}_1\bar{\boldsymbol{W}}_1^\top \bar{\boldsymbol{W}}_3^{-1}) \\ &\quad - 2(\boldsymbol{a}^\top \bar{\boldsymbol{W}}_3^{-1} \otimes \boldsymbol{a}^\top \bar{\boldsymbol{W}}_3^{-1})\bar{\boldsymbol{W}}_5\bar{\boldsymbol{W}}_3^{-1}\bar{\boldsymbol{W}}_1\Big)^{-1/2} \sqrt{G}\boldsymbol{a}^\top \bar{\boldsymbol{W}}_3^{-1}\bar{\boldsymbol{W}}_1, \end{aligned} \tag{B.21}$$

which, by Assumptions 4 and 5, is a smooth function of the sample average of the random vector $\breve{\boldsymbol{W}}_g$ for N (or equivalently G) sufficiently large. The existence of a valid Edgeworth expansion for the CDF of $\breve{\boldsymbol{W}}_g$ follows from Bhattacharya and Rao (1976, Thm. 20.6) because we have imposed Cramér's condition in Assumption 6 on the characteristic function $\chi_g(\boldsymbol{t})$ of $\breve{\boldsymbol{W}}_g$ and because the moment condition $\sup_{g\in\mathbb{N}} \mathrm{E}\|\breve{\boldsymbol{W}}_g\|^{2+m+\lambda} < \infty$ for $\lambda > 0$ implies the Lindeberg-type condition (20.54) in Bhattacharya and Rao (1976, Thm. 20.6).

Following the delta method as detailed in Remark 1.4 of Bhattacharya and Ghosh (1978) and p. 209 of Skovgaard (1981), we first derive an approximation

$$t_a = d^{-1/2}\tilde{t}_a + O_P(G^{-3/2}) \tag{B.22}$$

and define the approximate cumulants

$$\Pi_1(t_a) = d^{-1/2}\tilde{\mathrm{E}}(\tilde{t}_a), \tag{B.23}$$

$$\Pi_2(t_a) = d^{-1}(\tilde{\mathrm{E}}(\tilde{t}_a^2) - (\tilde{\mathrm{E}}(\tilde{t}_a))^2), \tag{B.24}$$

$$\Pi_3(t_a) = d^{-3/2}(\tilde{\mathrm{E}}(\tilde{t}_a^3) - 3\tilde{\mathrm{E}}(\tilde{t}_a^2)\tilde{\mathrm{E}}(\tilde{t}_a) + 2(\tilde{\mathrm{E}}(\tilde{t}_a))^3), \tag{B.25}$$

$$\Pi_4(t_a) = d^{-2}(\tilde{\mathrm{E}}(\tilde{t}_a^4) - 4\tilde{\mathrm{E}}(\tilde{t}_a^3)\tilde{\mathrm{E}}(\tilde{t}_a) - 3(\tilde{\mathrm{E}}(\tilde{t}_a^2))^2 + 12\tilde{\mathrm{E}}(\tilde{t}_a^2)(\tilde{\mathrm{E}}(\tilde{t}_a))^2 - 6(\tilde{\mathrm{E}}(\tilde{t}_a))^4), \tag{B.26}$$

where $\tilde{\mathrm{E}}(\tilde{t}_a^k)$ denotes the approximate moments of $\tilde{t}_a$. The latter are obtained by taking powers of $\tilde{t}_a$, dropping terms that are at most $O_P(G^{-(m+1)/2})$, and then taking expectations of the remaining terms.

As we will see, the approximate cumulants in (B.23)–(B.26) all have the structure

$$\Pi_j(t_a) = d^{-j/2}\sum_{i=0}^{m} G^{-i/2}\kappa_{ji} + O(G^{-(m+1)/2}), \quad j = 1, \ldots, 4, \tag{B.27}$$

for given constants κ_{ji} that satisfy $\kappa_{10} = \kappa_{21} = \kappa_{30} = \kappa_{40} = \kappa_{41} = 0$ and $\kappa_{20} = 1$. Inversion of the characteristic function yields the Edgeworth expansions (23) and (25) with the polynomials

$$q_1(x) = -\frac{1}{d^{1/2}}\kappa_{11} - \frac{1}{6d^{3/2}}\kappa_{31}(x^2 - 1), \tag{B.28}$$

$$q_2(x) = -\frac{1}{2d}(\kappa_{22} + \kappa_{11}^2)x - \frac{1}{24d^2}(\kappa_{42} + 4\kappa_{11}\kappa_{31})(x^3 - 3x) - \frac{1}{72d^3}\kappa_{31}^2(x^5 - 10x^3 + 15x). \tag{B.29}$$

We analogously define the corresponding bootstrap cumulants $\ddot{\Pi}_j(t_a^*)$ for $j = 1, \ldots, 4$, replacing the population mean $\mathrm{E}(\cdot)$ by the bootstrap analog $\mathrm{E}^*(\cdot)$, and deduce $\ddot{\kappa}_{ji}$, and hence $\ddot{q}_1$ and $\ddot{q}_2$, in the same way as κ_{ji}.

The remainder of the proof is divided into three parts. First, we derive the approximation (B.22) to the sample t-statistic. Then we find the approximate moments and cumulants as needed to determine the coefficients κ_{ji}, for $j = 1, \ldots, 4$. In the final part, we derive the corresponding results for (both versions of) the bootstrap t-statistic.

B.4.1 Derivation of the approximation (B.22)

We first note that

$$\boldsymbol{\Xi} \sim O(1) \quad \text{and} \quad \nu_a \sim O(1), \tag{B.30}$$

where "$\sim$" means exact rate in the sense that the right-hand side is not "small o". The first statement in (B.30) follows from Assumptions 4 and 5 and the second uses also (26).

There are two non-linearities in (B.21) that we will need to linearize, namely the inverse in $\bar{\boldsymbol{W}}_3^{-1}$ and the square-root in the denominator. First, we apply the expansion

$$\begin{aligned}
\bar{\boldsymbol{W}}_3^{-1} &= \boldsymbol{\Xi}^{-1} - \boldsymbol{\Xi}^{-1}(\bar{\boldsymbol{W}}_3 - \boldsymbol{\Xi})\bar{\boldsymbol{W}}_3^{-1} \\
&= \boldsymbol{\Xi}^{-1} - \boldsymbol{\Xi}^{-1}(\bar{\boldsymbol{W}}_3 - \boldsymbol{\Xi})\boldsymbol{\Xi}^{-1} + \boldsymbol{\Xi}^{-1}(\bar{\boldsymbol{W}}_3 - \boldsymbol{\Xi})\boldsymbol{\Xi}^{-1}(\bar{\boldsymbol{W}}_3 - \boldsymbol{\Xi})\bar{\boldsymbol{W}}_3^{-1} \\
&= \boldsymbol{\Xi}^{-1} - \boldsymbol{\Xi}^{-1}(\bar{\boldsymbol{W}}_3 - \boldsymbol{\Xi})\boldsymbol{\Xi}^{-1} + \boldsymbol{\Xi}^{-1}(\bar{\boldsymbol{W}}_3 - \boldsymbol{\Xi})\boldsymbol{\Xi}^{-1}(\bar{\boldsymbol{W}}_3 - \boldsymbol{\Xi})\boldsymbol{\Xi}^{-1} + O_P(G^{-3/2}),
\end{aligned} \tag{B.31}$$

where the order of the remainder follows from Lemma A.1 and (B.30). Using (B.31), the numerator of (B.21) is linearized as

$$\sqrt{G}\boldsymbol{a}^\top\boldsymbol{\Xi}^{-1}\bar{\boldsymbol{W}}_1 - \sqrt{G}\boldsymbol{a}^\top\boldsymbol{\Xi}^{-1}(\bar{\boldsymbol{W}}_3 - \boldsymbol{\Xi})\boldsymbol{\Xi}^{-1}\bar{\boldsymbol{W}}_1 + \sqrt{G}\boldsymbol{a}^\top\boldsymbol{\Xi}^{-1}(\bar{\boldsymbol{W}}_3 - \boldsymbol{\Xi})\boldsymbol{\Xi}^{-1}(\bar{\boldsymbol{W}}_3 - \boldsymbol{\Xi})\boldsymbol{\Xi}^{-1}\bar{\boldsymbol{W}}_1 + O_P(G^{-3/2}),$$

where the order of the remainder follows from (B.31) together with Lemma A.1 and Assumption 5. Using again (B.31), the denominator of (B.21) is (the square-root of)

$$\begin{aligned}
&\nu_a + \boldsymbol{a}^\top\boldsymbol{\Xi}^{-1}(\bar{\boldsymbol{W}}_2 - \mathrm{E}(\bar{\boldsymbol{W}}_2))\boldsymbol{\Xi}^{-1}\boldsymbol{a} - 2\boldsymbol{a}^\top\boldsymbol{\Xi}^{-1}(\bar{\boldsymbol{W}}_3 - \boldsymbol{\Xi})\boldsymbol{\Xi}^{-1}\mathrm{E}(\bar{\boldsymbol{W}}_2)\boldsymbol{\Xi}^{-1}\boldsymbol{a} \\
&\quad + \boldsymbol{a}^\top\boldsymbol{\Xi}^{-1}(\bar{\boldsymbol{W}}_3 - \boldsymbol{\Xi})\boldsymbol{\Xi}^{-1}\mathrm{E}(\bar{\boldsymbol{W}}_2)\boldsymbol{\Xi}^{-1}(\bar{\boldsymbol{W}}_3 - \boldsymbol{\Xi})\boldsymbol{\Xi}^{-1}\boldsymbol{a} \\
&\quad + 2\boldsymbol{a}^\top\boldsymbol{\Xi}^{-1}(\bar{\boldsymbol{W}}_3 - \boldsymbol{\Xi})\boldsymbol{\Xi}^{-1}(\bar{\boldsymbol{W}}_3 - \boldsymbol{\Xi})\boldsymbol{\Xi}^{-1}\mathrm{E}(\bar{\boldsymbol{W}}_2)\boldsymbol{\Xi}^{-1}\boldsymbol{a} - 2\boldsymbol{a}^\top\boldsymbol{\Xi}^{-1}(\bar{\boldsymbol{W}}_3 - \boldsymbol{\Xi})\boldsymbol{\Xi}^{-1}(\bar{\boldsymbol{W}}_2 - \mathrm{E}(\bar{\boldsymbol{W}}_2))\boldsymbol{\Xi}^{-1}\boldsymbol{a} \\
&\quad + (\boldsymbol{a}^\top\boldsymbol{\Xi}^{-1} \otimes \boldsymbol{a}^\top\boldsymbol{\Xi}^{-1})\bar{\boldsymbol{W}}_4\,\mathrm{vec}(\boldsymbol{\Xi}^{-1}\bar{\boldsymbol{W}}_1\bar{\boldsymbol{W}}_1^\top\boldsymbol{\Xi}^{-1}) - 2(\boldsymbol{a}^\top\boldsymbol{\Xi}^{-1} \otimes \boldsymbol{a}^\top\boldsymbol{\Xi}^{-1})\bar{\boldsymbol{W}}_5\boldsymbol{\Xi}^{-1}\bar{\boldsymbol{W}}_1 + O_P(G^{-3/2}).
\end{aligned}$$

Second, we apply a second-order Taylor-series expansion around $x = 0$,

$$(\nu_a + x)^{-1/2} = \nu_a^{-1/2} - \frac{1}{2}\nu_a^{-3/2}x + \frac{3}{8}\nu_a^{-5/2}x^2 + O(x^3),$$

which, together with the linearization of the numerator, yields the expansion (B.22) with

$$\tilde{t} = \sqrt{G}\bar{Z}_6 - \sqrt{G}(\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top\bar{\boldsymbol{Z}}_1 + \sqrt{G}(\bar{Z}_{11} - \bar{\mu}_{11})\bar{Z}_6 - \frac{1}{2}\sqrt{G}(\bar{Z}_2 - \bar{\mu}_2)\bar{Z}_6 \tag{B.32}$$

$$+ \frac{1}{2}\sqrt{G}(\bar{Z}_2 - \bar{\mu}_2)(\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top\bar{\boldsymbol{Z}}_1 + \sqrt{G}(\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top(\bar{\boldsymbol{Z}}_9 - \bar{\boldsymbol{\mu}}_9)\bar{\boldsymbol{Z}}_1 \tag{B.33}$$

$$- \sqrt{G}(\bar{Z}_{11} - \bar{\mu}_{11})(\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top\bar{\boldsymbol{Z}}_1 - \frac{1}{2}\sqrt{G}(\bar{\boldsymbol{Z}}_{10} - \bar{\boldsymbol{\mu}}_{10})^\top(\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)\bar{Z}_6 \tag{B.34}$$

$$- \sqrt{G}(\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top(\bar{\boldsymbol{Z}}_{12} - \bar{\boldsymbol{\mu}}_{12})\bar{Z}_6 + \sqrt{G}(\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top(\bar{\boldsymbol{Z}}_8 - \bar{\boldsymbol{\mu}}_8)\bar{Z}_6 \tag{B.35}$$

$$+ \sqrt{G}\bar{\boldsymbol{Z}}_1^\top\bar{\boldsymbol{Z}}_5\bar{Z}_6 - \frac{1}{2}\sqrt{G}\bar{\boldsymbol{Z}}_1^\top\bar{\boldsymbol{\mu}}_4\bar{\boldsymbol{Z}}_1\bar{Z}_6 + \frac{3}{2}\sqrt{G}(\bar{Z}_{11} - \bar{\mu}_{11})^2\bar{Z}_6 \tag{B.36}$$

$$+ \frac{3}{8}\sqrt{G}(\bar{Z}_2 - \bar{\mu}_2)^2\bar{Z}_6 - \frac{3}{2}\sqrt{G}(\bar{Z}_2 - \bar{\mu}_2)(\bar{Z}_{11} - \bar{\mu}_{11})\bar{Z}_6, \tag{B.37}$$

where $\boldsymbol{Z}_{1g}, Z_{2g}, \boldsymbol{Z}_{3g}$ are defined in (30), $\boldsymbol{\mu}_{jg} = \mathrm{E}(\boldsymbol{Z}_{jg})$, $\bar{\boldsymbol{\mu}}_j = G^{-1}\sum_{g=1}^{G}\boldsymbol{\mu}_{jg}$, and

$$\boldsymbol{Z}_{4g} = \boldsymbol{X}_g^\top \boldsymbol{X}_g \boldsymbol{\Xi}^{-1}\boldsymbol{a}\boldsymbol{a}^\top \boldsymbol{\Xi}^{-1}\boldsymbol{X}_g^\top \boldsymbol{X}_g, \tag{B.38}$$

$$\boldsymbol{Z}_{5g} = \boldsymbol{Z}_{3g}\boldsymbol{a}^\top \boldsymbol{Z}_{1g} = \nu_a^{-1/2}\boldsymbol{X}_g^\top \boldsymbol{X}_g \boldsymbol{\Xi}^{-1}\boldsymbol{a}\boldsymbol{a}^\top \boldsymbol{\Xi}^{-1}\boldsymbol{X}_g^\top \boldsymbol{u}_g, \tag{B.39}$$

$$Z_{6g} = \boldsymbol{a}^\top \boldsymbol{Z}_{1g} = \nu_a^{-1/2}\boldsymbol{a}^\top \boldsymbol{\Xi}^{-1}\boldsymbol{X}_g^\top \boldsymbol{u}_g, \tag{B.40}$$

$$\boldsymbol{Z}_{7g} = \boldsymbol{Z}_{1g}\boldsymbol{Z}_{1g}^\top = \nu_a^{-1}\boldsymbol{\Xi}^{-1}\boldsymbol{X}_g^\top \boldsymbol{u}_g \boldsymbol{u}_g^\top \boldsymbol{X}_g \boldsymbol{\Xi}^{-1}, \tag{B.41}$$

$$\boldsymbol{Z}_{8g} = \boldsymbol{Z}_{7g}\boldsymbol{a} = \nu_a^{-1}\boldsymbol{\Xi}^{-1}\boldsymbol{X}_g^\top \boldsymbol{u}_g \boldsymbol{u}_g^\top \boldsymbol{X}_g \boldsymbol{\Xi}^{-1}\boldsymbol{a}, \tag{B.42}$$

$$\boldsymbol{Z}_{9g} = \boldsymbol{\Xi}^{-1}\boldsymbol{X}_g^\top \boldsymbol{X}_g, \tag{B.43}$$

$$\boldsymbol{Z}_{10g} = \bar{\boldsymbol{\mu}}_7 \boldsymbol{Z}_{3g} = \bar{\boldsymbol{\mu}}_7 \boldsymbol{X}_g^\top \boldsymbol{X}_g \boldsymbol{\Xi}^{-1}\boldsymbol{a}, \tag{B.44}$$

$$Z_{11g} = \boldsymbol{a}^\top \bar{\boldsymbol{\mu}}_7 \boldsymbol{Z}_{3g} = \boldsymbol{a}^\top \bar{\boldsymbol{\mu}}_7 \boldsymbol{X}_g^\top \boldsymbol{X}_g \boldsymbol{\Xi}^{-1}\boldsymbol{a}, \tag{B.45}$$

$$\boldsymbol{Z}_{12g} = \boldsymbol{Z}_{9g}\bar{\boldsymbol{\mu}}_7 \boldsymbol{a} = \boldsymbol{\Xi}^{-1}\boldsymbol{X}_g^\top \boldsymbol{X}_g \bar{\boldsymbol{\mu}}_7 \boldsymbol{a}. \tag{B.46}$$

It follows easily from Lemma A.1 and Assumptions 4 and 5 that

$$\bar{\boldsymbol{Z}}_j - \bar{\boldsymbol{\mu}}_j = G^{-1}\sum_{g=1}^{G}(\boldsymbol{Z}_{jg} - \boldsymbol{\mu}_{jg}) = O_P(G^{-1/2}) \quad \text{for } j = 1, \ldots, 12. \tag{B.47}$$

It is also easy to see that $\bar{\boldsymbol{\mu}}_j = G^{-1}\sum_g \boldsymbol{\mu}_{jg} = O(1)$ (non-random) for $j = 1, \ldots, 12$. In $\tilde{t}$, it then follows straightforwardly from (B.47) that the first term on the right-hand side of (B.32) is $O_P(1)$, while the remaining terms on the right-hand side of (B.32) are $O_P(G^{-1/2})$. All the terms in (B.33)–(B.37) are $O_P(G^{-1})$.

B.4.2 Derivation of the cumulant expansion (B.27)

We first find the approximate moments of $\tilde{t}_a$.

Approximate first moment of $\tilde{t}_a$. By Lemma A.3 with $k = 3$, the expectation of each of the terms in (B.33)–(B.37) is $O(G^{-3/2})$. By the law of iterated expectations, the expectation of the first three terms in (B.32) is zero. This leaves only one term, and we find

$$\tilde{\mathrm{E}}(\tilde{t}_a) = -\frac{1}{2}G^{1/2}\mathrm{E}((\bar{Z}_2 - \bar{\mu}_2)\bar{Z}_6) = -\frac{1}{2}G^{-1/2}\gamma_{6,6} \tag{B.48}$$

from Lemma A.4.

Approximate second moment of $\tilde{t}_a$. The square of $\tilde{t}$ is, using (B.47),

$$\begin{aligned}
\tilde{t}_a^2 = {} & G\bar{Z}_6^2 + G\big((\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top \bar{\boldsymbol{Z}}_1\big)^2 + 4G(\bar{Z}_{11} - \bar{\mu}_{11})^2\bar{Z}_6^2 + G(\bar{Z}_2 - \bar{\mu}_2)^2\bar{Z}_6^2 \\
& - 2G(\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top \bar{\boldsymbol{Z}}_1\bar{Z}_6 + 2G(\bar{Z}_{11} - \bar{\mu}_{11})\bar{Z}_6^2 - G(\bar{Z}_2 - \bar{\mu}_2)\bar{Z}_6^2 \\
& + 2G\bar{Z}_6(\bar{Z}_2 - \bar{\mu}_2)(\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top \bar{\boldsymbol{Z}}_1 + 2G\bar{Z}_6(\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top(\bar{\boldsymbol{Z}}_9 - \bar{\boldsymbol{\mu}}_9)\bar{\boldsymbol{Z}}_1 \\
& - 4G\bar{Z}_6(\bar{Z}_{11} - \bar{\mu}_{11})(\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top \bar{\boldsymbol{Z}}_1 - G(\bar{\boldsymbol{Z}}_{10} - \bar{\boldsymbol{\mu}}_{10})^\top(\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)\bar{Z}_6^2 \\
& - 2G(\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top(\bar{\boldsymbol{Z}}_{12} - \bar{\boldsymbol{\mu}}_{12})\bar{Z}_6^2 + 2G(\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top(\bar{\boldsymbol{Z}}_8 - \bar{\boldsymbol{\mu}}_8)\bar{Z}_6^2 \\
& + 2G\bar{\boldsymbol{Z}}_1^\top \bar{\boldsymbol{Z}}_5\bar{Z}_6^2 - G\bar{\boldsymbol{Z}}_1^\top \bar{\boldsymbol{\mu}}_4\bar{\boldsymbol{Z}}_1\bar{Z}_6^2 - 4G(\bar{Z}_2 - \bar{\mu}_2)(\bar{Z}_{11} - \bar{\mu}_{11})\bar{Z}_6^2 + O_P(G^{-3/2}).
\end{aligned}$$

Applying Lemma A.4, the approximate second moment is

$$\begin{aligned}\tilde{\mathrm{E}}(\tilde{t}_a^2) = 1 &- G^{-1}\operatorname{Tr}\{\bar{\boldsymbol{\xi}}_{3,3}\boldsymbol{\xi}_{1,1}\} + 4G^{-1}(\zeta_{11,11} - \bar{\zeta}_{11,11}) + 2G^{-1}\gamma_{6,6}^2 \\ &+ 2G^{-1}(\gamma_{6,11} - \bar{\gamma}_{6,11}) + 2G^{-1}\boldsymbol{\xi}_{6,1}(\boldsymbol{\zeta}_{2,3} - \bar{\boldsymbol{\zeta}}_{2,3}) + 2G^{-1}\boldsymbol{\xi}_{6,1}(\boldsymbol{\zeta}_{9,3} - \bar{\boldsymbol{\zeta}}_{9,3}) \\ &- 4G^{-1}\boldsymbol{\xi}_{6,1}(\boldsymbol{\zeta}_{11,3} - \bar{\boldsymbol{\zeta}}_{11,3}) - G^{-1}(\zeta_{3,10} - \bar{\zeta}_{3,10}) - 2G^{-1}(\zeta_{3,12} - \bar{\zeta}_{3,12}) + 2G^{-1}(\zeta_{3,8} - \bar{\zeta}_{3,8}) \\ &+ 2G^{-1}(\operatorname{Tr}\{\bar{\boldsymbol{\gamma}}_{1,3}\} + 2\boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,2}) - 2G^{-1}\boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,6} - 4G^{-1}(\zeta_{11,2} - \bar{\zeta}_{11,2}) + O(G^{-2}),\end{aligned}$$

where we also used that $\xi_{6,6} = \boldsymbol{a}^\top\boldsymbol{\xi}_{1,1}\boldsymbol{a} = 1$. Using Lemma A.5 and $\boldsymbol{\zeta}_{2,3} = \boldsymbol{\xi}_{3,2}$, this simplifies to

$$\tilde{\mathrm{E}}(\tilde{t}_a^2) = 1 - G^{-1}\operatorname{Tr}\{\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,1}\} + 2G^{-1}\gamma_{6,6}^2 + 2G^{-1}\operatorname{Tr}\{\boldsymbol{\gamma}_{1,3}\} - 2G^{-1}\boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,6} + 4G^{-1}\boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,2} + O(G^{-2}).$$

Approximate third moment of $\tilde{t}_a$. Using (B.47) we find

$$\begin{aligned}\tilde{t}^3 = &G^{3/2}\bar{Z}_6^3 - 3G^{3/2}\bar{Z}_6^2(\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top\bar{\boldsymbol{Z}}_1 + 3G^{3/2}\bar{Z}_6^3(\bar{Z}_{11} - \bar{\mu}_{11}) \\ &- \frac{3}{2}G^{3/2}\bar{Z}_6^3(\bar{Z}_2 - \bar{\mu}_2) + 3G^{3/2}\bar{Z}_6^2(\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top(\bar{\boldsymbol{Z}}_9 - \bar{\boldsymbol{\mu}}_9)\bar{\boldsymbol{Z}}_1 \\ &- 9G^{3/2}\bar{Z}_6^2(\bar{Z}_{11} - \bar{\mu}_{11})(\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top\bar{\boldsymbol{Z}}_1 + \frac{9}{2}G^{3/2}\bar{Z}_6^2(\bar{Z}_2 - \bar{\mu}_2)(\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top\bar{\boldsymbol{Z}}_1 \\ &- \frac{3}{2}G^{3/2}\bar{Z}_6^3(\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top(\bar{\boldsymbol{Z}}_{10} - \bar{\boldsymbol{\mu}}_{10}) - 3G^{3/2}\bar{Z}_6^3(\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top(\bar{\boldsymbol{Z}}_{12} - \bar{\boldsymbol{\mu}}_{12}) \\ &+ 3G^{3/2}\bar{Z}_6^3(\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top(\bar{\boldsymbol{Z}}_8 - \bar{\boldsymbol{\mu}}_8) + 3G^{3/2}\bar{Z}_6^3\bar{\boldsymbol{Z}}_1^\top\bar{\boldsymbol{Z}}_5 \\ &- \frac{3}{2}G^{3/2}\bar{Z}_6^3\bar{\boldsymbol{Z}}_1^\top\bar{\boldsymbol{\mu}}_4\bar{\boldsymbol{Z}}_1 + \frac{15}{2}G^{3/2}\bar{Z}_6^3(\bar{Z}_{11} - \bar{\mu}_{11})^2 + \frac{15}{8}G^{3/2}\bar{Z}_6^3(\bar{Z}_2 - \bar{\mu}_2)^2 \\ &- \frac{15}{2}G^{3/2}\bar{Z}_6^3(\bar{Z}_2 - \bar{\mu}_2)(\bar{Z}_{11} - \bar{\mu}_{11}) + 3G^{3/2}\bar{Z}_6\big((\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top\bar{\boldsymbol{Z}}_1\big)^2 + O_P(G^{-3/2}),\end{aligned}$$

but Lemma A.3 shows that the expectation of any term that contains a product of five or more $\bar{\boldsymbol{Z}}_j - \bar{\boldsymbol{\mu}}_j$ is at most $O(G^{-3/2})$. That leaves only the first four terms on the right-hand side. Of these, the law of iterated expectations shows that the second and third terms are $O(G^{-3/2})$, and thus

$$\tilde{\mathrm{E}}(\tilde{t}_a^3) = G^{-1/2}\gamma_{6,6} - \frac{9}{2}G^{-1/2}\gamma_{6,6} + O(G^{-3/2}) = -\frac{7}{2}\gamma_{6,6} + O(G^{-3/2}).$$

Approximate fourth moment of $\tilde{t}_a$. Using (B.47) we find

$$\begin{aligned}\tilde{t}_a^4 = &G^2\bar{Z}_6^4 - 4G^2\bar{Z}_6^3(\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top\bar{\boldsymbol{Z}}_1 + 4G^2\bar{Z}_6^4(\bar{Z}_{11} - \bar{\mu}_{11}) - 2G^2\bar{Z}_6^4(\bar{Z}_2 - \bar{\mu}_2) \\ &+ 8G^2\bar{Z}_6^3(\bar{Z}_2 - \bar{\mu}_2)(\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top\bar{\boldsymbol{Z}}_1 + 4G^2\bar{Z}_6^3(\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top(\bar{\boldsymbol{Z}}_9 - \bar{\boldsymbol{\mu}}_9)\bar{\boldsymbol{Z}}_1 \\ &- 16G^2\bar{Z}_6^3(\bar{Z}_{11} - \bar{\mu}_{11})(\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top\bar{\boldsymbol{Z}}_1 - 2G^2\bar{Z}_6^4(\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top(\bar{\boldsymbol{Z}}_{10} - \bar{\boldsymbol{\mu}}_{10}) \\ &- 4G^2\bar{Z}_6^4(\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top(\bar{\boldsymbol{Z}}_{12} - \bar{\boldsymbol{\mu}}_{12}) + 4G^2\bar{Z}_6^4(\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top(\bar{\boldsymbol{Z}}_8 - \bar{\boldsymbol{\mu}}_8) \\ &+ 4G^2\bar{Z}_6^4\bar{\boldsymbol{Z}}_1^\top\bar{\boldsymbol{Z}}_5 - 2G^2\bar{Z}_6^4\bar{\boldsymbol{Z}}_1^\top\bar{\boldsymbol{\mu}}_4\bar{\boldsymbol{Z}}_1 + 12G^2\bar{Z}_6^4(\bar{Z}_{11} - \bar{\mu}_{11})^2 + 3G^2\bar{Z}_6^4(\bar{Z}_2 - \bar{\mu}_2)^2 \\ &- 12G^2\bar{Z}_6^4(\bar{Z}_2 - \bar{\mu}_2)(\bar{Z}_{11} - \bar{\mu}_{11}) + 6G^2\bar{Z}_6^2\big((\bar{\boldsymbol{Z}}_3 - \bar{\boldsymbol{\mu}}_3)^\top\bar{\boldsymbol{Z}}_1\big)^2 + O_P(G^{-3/2}).\end{aligned}$$

Applying Lemma A.4, the approximate fourth moment is

$$\begin{aligned}\tilde{\mathrm{E}}(\tilde{t}_a^4) = 3 &- 2G^{-1}\xi_{2,2} - 12G^{-1}\boldsymbol{\xi}_{6,1}(\boldsymbol{\xi}_{3,2} - \bar{\boldsymbol{\xi}}_{3,2}) + 12G^{-1}(\gamma_{6,11} - \bar{\gamma}_{6,11}) + 12G^{-1}(\xi_{11,2} - \bar{\xi}_{11,2}) \\ &+ 28G^{-1}\gamma_{6,6}^2 + 24G^{-1}\boldsymbol{\xi}_{6,1}(\boldsymbol{\zeta}_{2,3} - \bar{\boldsymbol{\zeta}}_{2,3}) + 12G^{-1}\boldsymbol{\xi}_{6,1}(\boldsymbol{\zeta}_{9,3} - \bar{\boldsymbol{\zeta}}_{9,3}) - 48G^{-1}\boldsymbol{\xi}_{6,1}(\boldsymbol{\zeta}_{11,3} - \bar{\boldsymbol{\zeta}}_{11,3}) \\ &- 6G^{-1}(\zeta_{3,10} - \bar{\zeta}_{3,10}) - 12G^{-1}(\zeta_{3,12} - \bar{\zeta}_{3,12}) + 12G^{-1}(\zeta_{3,8} - \bar{\zeta}_{3,8}) \\ &+ 12G^{-1}\operatorname{Tr}\{\bar{\boldsymbol{\gamma}}_{1,3}\} + 48\boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,2} - 6G^{-1}\operatorname{Tr}\{\bar{\boldsymbol{\xi}}_{3,3}\boldsymbol{\xi}_{1,1}\} - 12G^{-1}\boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,6} \\ &+ 36G^{-1}(\zeta_{11,11} - \bar{\zeta}_{11,11}) - 36G^{-1}(\zeta_{11,2} - \bar{\zeta}_{11,2}) - 12G^{-1}\boldsymbol{\xi}_{6,1}\bar{\boldsymbol{\xi}}_{3,3}\boldsymbol{\xi}_{1,6} + O(G^{-2}).\end{aligned}$$

Using Lemma A.5 and $\boldsymbol{\zeta}_{2,3} = \boldsymbol{\xi}_{3,2}$, this simplifies to

$$\begin{aligned}\tilde{\mathrm{E}}(\tilde{t}^4) = 3 &- 2G^{-1}\xi_{2,2} + 28G^{-1}\gamma_{6,6}^2 - 24G^{-1}\boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,6} - 6G^{-1}\operatorname{Tr}\{\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,1}\} \\ &+ 12G^{-1}\operatorname{Tr}\{\boldsymbol{\gamma}_{1,3}\} + 48G^{-1}\boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,2} + O(G^{-2}).\end{aligned}$$

Approximate cumulants. Inserting the approximate moments into (B.23)–(B.26) and using (B.38)–(B.46), we obtain the approximate cumulants

$$\begin{aligned}d^{1/2}\Pi_1(t_a) &= -\frac{1}{2}G^{-1/2}\gamma_{6,6} = -\frac{1}{2}G^{-1/2}\boldsymbol{a}^\top\boldsymbol{\gamma}_{1,1}\boldsymbol{a}, \\ d\Pi_2(t_a) &= 1 - G^{-1}\operatorname{Tr}\{\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,1}\} + 2G^{-1}\gamma_{6,6}^2 + 2G^{-1}\operatorname{Tr}\{\boldsymbol{\gamma}_{1,3}\} - 2G^{-1}\boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,6} \\ &\quad + 4G^{-1}\boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,2} - \frac{1}{4}G^{-1}\gamma_{6,6}^2 + O(G^{-2}) \\ &= 1 + G^{-1}\Big(\frac{7}{4}(\boldsymbol{a}^\top\boldsymbol{\gamma}_{1,1}\boldsymbol{a})^2 - \operatorname{Tr}\{\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,1}\} + 2\operatorname{Tr}\{\boldsymbol{\gamma}_{1,3}\} - 2\boldsymbol{a}^\top\boldsymbol{\xi}_{1,1}\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,1}\boldsymbol{a} + 4\boldsymbol{a}^\top\boldsymbol{\xi}_{1,1}\boldsymbol{\xi}_{3,2}\Big) \\ &\quad + O(G^{-2}), \\ d^{3/2}\Pi_3(t_a) &= -2G^{-1/2}\gamma_{6,6} + O(G^{-3/2}) = -2G^{-1/2}\boldsymbol{a}^\top\boldsymbol{\gamma}_{1,1}\boldsymbol{a} + O(G^{-3/2}),\end{aligned}$$

and

$$\begin{aligned}d^2\Pi_4(t_a) &= 3 - 2G^{-1}\xi_{2,2} + 28G^{-1}\gamma_{6,6}^2 - 24G^{-1}\boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,6} - 6G^{-1}\operatorname{Tr}\{\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,1}\} \\ &\quad + 12G^{-1}\operatorname{Tr}\{\boldsymbol{\gamma}_{1,3}\} + 48G^{-1}\boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,2} - 4G^{-1}\frac{7}{4}\gamma_{6,6}^2 - 3 - 6\Big(-G^{-1}\operatorname{Tr}\{\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,1}\} \\ &\quad + 2G^{-1}\gamma_{6,6}^2 + 2G^{-1}\operatorname{Tr}\{\boldsymbol{\gamma}_{1,3}\} - 2G^{-1}\boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,6} + 4G^{-1}\boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,2}\Big) + 12\frac{1}{4}G^{-1}\gamma_{6,6}^2 + O(G^{-2}) \\ &= -2G^{-1}\xi_{2,2} + 28G^{-1}\gamma_{6,6}^2 - 24G^{-1}\boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,6} - 6G^{-1}\operatorname{Tr}\{\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,1}\} + 12G^{-1}\operatorname{Tr}\{\boldsymbol{\gamma}_{1,3}\} \\ &\quad + 48G^{-1}\boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,2} - 7G^{-1}\gamma_{6,6}^2 + 6G^{-1}\operatorname{Tr}\{\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,1}\} - 12G^{-1}\gamma_{6,6}^2 - 12G^{-1}\operatorname{Tr}\{\boldsymbol{\gamma}_{1,3}\} \\ &\quad + 12G^{-1}\boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,6} - 24G^{-1}\boldsymbol{\xi}_{6,1}\boldsymbol{\xi}_{3,2} + 3G^{-1}\gamma_{6,6}^2 + O(G^{-2}) \\ &= G^{-1}\Big(12(\boldsymbol{a}^\top\boldsymbol{\gamma}_{1,1}\boldsymbol{a})^2 - 2\xi_{2,2} - 12\boldsymbol{a}^\top\boldsymbol{\xi}_{1,1}\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,1}\boldsymbol{a} + 24\boldsymbol{a}^\top\boldsymbol{\xi}_{1,1}\boldsymbol{\xi}_{3,2}\Big) + O(G^{-2}).\end{aligned}$$

We finally conclude that

$$\begin{aligned}\kappa_{11} &= -\frac{1}{2}\boldsymbol{a}^\top\boldsymbol{\gamma}_{1,1}\boldsymbol{a}, \\ \kappa_{22} &= \frac{7}{4}(\boldsymbol{a}^\top\boldsymbol{\gamma}_{1,1}\boldsymbol{a})^2 - \operatorname{Tr}\{\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,1}\} + 2\operatorname{Tr}\{\boldsymbol{\gamma}_{1,3}\} - 2\boldsymbol{a}^\top\boldsymbol{\xi}_{1,1}\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,1}\boldsymbol{a} + 4\boldsymbol{a}^\top\boldsymbol{\xi}_{1,1}\boldsymbol{\xi}_{3,2}, \\ \kappa_{31} &= -2\boldsymbol{a}^\top\boldsymbol{\gamma}_{1,1}\boldsymbol{a}, \\ \kappa_{42} &= 12(\boldsymbol{a}^\top\boldsymbol{\gamma}_{1,1}\boldsymbol{a})^2 - 2\xi_{2,2} - 12\boldsymbol{a}^\top\boldsymbol{\xi}_{1,1}\boldsymbol{\xi}_{3,3}\boldsymbol{\xi}_{1,1}\boldsymbol{a} + 24\boldsymbol{a}^\top\boldsymbol{\xi}_{1,1}\boldsymbol{\xi}_{3,2}.\end{aligned}$$

In view of the moment conditions in Lemma A.4, we note that $\kappa_{11}, \kappa_{22}, \kappa_{31}$ exist under the conditions of the one-term expansion ($m = 1$) of Theorem 5.1, while κ_{42} exists under the conditions of the two-term expansion ($m = 2$). Thus, we obtain the results of Theorem 5.1 from (B.28) and (B.29).

B.4.3 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 $\boldsymbol{Z}_{jg}$ by $\boldsymbol{Z}_{jg}^*$ given in (32).

B.5 Proof of Theorem 5.2

First, we find that

$$\boldsymbol{a}^\top \ddot{\boldsymbol{\gamma}}_{1,1}\boldsymbol{a} = \frac{1}{G}\sum_{g=1}^{G} \mathrm{E}^*(\boldsymbol{a}^\top \boldsymbol{Z}_{1g}^*)^3 = \ddot{\nu}_a^{-3/2}\frac{1}{G}\sum_{g=1}^{G} \mathrm{E}^*(\boldsymbol{a}^\top \ddot{\boldsymbol{\Xi}}^{-1}\boldsymbol{X}_g^\top \boldsymbol{u}_g^*)^3. \tag{B.49}$$

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

$$\boldsymbol{a}^\top \ddot{\boldsymbol{\gamma}}_{1,1}\boldsymbol{a} = \mathrm{E}^*(v^{*3})\ddot{\nu}_a^{-3/2}\frac{1}{G}\sum_{g=1}^{G} (\boldsymbol{a}^\top \ddot{\boldsymbol{\Xi}}^{-1}\boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g)^3 = \mathrm{E}^*(v^{*3})(\boldsymbol{a}^\top \boldsymbol{\gamma}_{1,1}\boldsymbol{a} + B_1 + B_2 + B_3 + B_4 + B_5),$$

where

$$\begin{aligned}
B_1 &= \nu_a^{-3/2}\frac{1}{G}\sum_{g=1}^{G}\Big((\boldsymbol{a}^\top \boldsymbol{\Xi}^{-1}\boldsymbol{X}_g^\top \boldsymbol{u}_g)^3 - \mathrm{E}(\boldsymbol{a}^\top \boldsymbol{\Xi}^{-1}\boldsymbol{X}_g^\top \boldsymbol{u}_g)^3\Big),\\
B_2 &= \nu_a^{-3/2}\frac{1}{G}\sum_{g=1}^{G}\Big((\boldsymbol{a}^\top \boldsymbol{\Xi}^{-1}\boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g)^3 - (\boldsymbol{a}^\top \boldsymbol{\Xi}^{-1}\boldsymbol{X}_g^\top \boldsymbol{u}_g)^3\Big),\\
B_3 &= \nu_a^{-3/2}\frac{1}{G}\sum_{g=1}^{G}\Big((\boldsymbol{a}^\top \ddot{\boldsymbol{\Xi}}^{-1}\boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g)^3 - (\boldsymbol{a}^\top \boldsymbol{\Xi}^{-1}\boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g)^3\Big),\\
B_4 &= (\ddot{\nu}_a^{-3/2} - \nu_a^{-3/2})\frac{1}{G}\sum_{g=1}^{G}(\boldsymbol{a}^\top \ddot{\boldsymbol{\Xi}}^{-1}\boldsymbol{X}_g^\top \boldsymbol{u}_g)^3, \text{ and}\\
B_5 &= (\ddot{\nu}_a^{-3/2} - \nu_a^{-3/2})\frac{1}{G}\sum_{g=1}^{G}\Big((\boldsymbol{a}^\top \ddot{\boldsymbol{\Xi}}^{-1}\boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g)^3 - (\boldsymbol{a}^\top \ddot{\boldsymbol{\Xi}}^{-1}\boldsymbol{X}_g^\top \boldsymbol{u}_g)^3\Big),
\end{aligned}$$

and we analyze each term B_i, for $i = 1, \ldots, 4$, in turn.

First note that, by (28), (30), and (B.40), we have $\boldsymbol{a}^\top \boldsymbol{\Xi}^{-1}\boldsymbol{X}_g^\top \boldsymbol{u}_g = \boldsymbol{a}^\top \boldsymbol{\Xi}^{-1}\boldsymbol{W}_{1g} = \boldsymbol{a}^\top \boldsymbol{Z}_{1g} = Z_{6g}$, such that $B_1 = G^{-1}\sum_{g=1}^{G}(Z_{6g}^3 - \mathrm{E}(Z_{6g}^3))$. Because $Z_{6g}^3 - \mathrm{E}(Z_{6g}^3)$ is an independent, mean-zero sequence with finite second moments by Assumption 5, it follows from Lemma A.1 that $B_1 = O_P(G^{-1/2})$.

To analyze B_2, we use the decomposition $\boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g = \boldsymbol{X}_g^\top \boldsymbol{u}_g - \boldsymbol{X}_g^\top \boldsymbol{X}_g(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ and find

$$B_2 = 3B_{21} - 3B_{22} - B_{23},$$

where, see (30) and (B.40),

$$\begin{aligned}
B_{21} &= \nu_a^{-3/2}\frac{1}{G}\sum_{g=1}^{G}\boldsymbol{a}^\top \boldsymbol{\Xi}^{-1}\boldsymbol{X}_g^\top \boldsymbol{u}_g(\boldsymbol{a}^\top \boldsymbol{\Xi}^{-1}\boldsymbol{X}_g^\top \boldsymbol{X}_g(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_0))^2 = \nu_a^{-1}\frac{1}{G}\sum_{g=1}^{G} Z_{6g}(\boldsymbol{Z}_{3g}^\top(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_0))^2,\\
B_{22} &= \nu_a^{-3/2}\frac{1}{G}\sum_{g=1}^{G}(\boldsymbol{a}^\top \boldsymbol{\Xi}^{-1}\boldsymbol{X}_g^\top \boldsymbol{u}_g)^2\boldsymbol{a}^\top \boldsymbol{\Xi}^{-1}\boldsymbol{X}_g^\top \boldsymbol{X}_g(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \nu_a^{-1/2}\frac{1}{G}\sum_{g=1}^{G} Z_{6g}^2\boldsymbol{Z}_{3g}^\top(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_0),\\
B_{23} &= \nu_a^{-3/2}\frac{1}{G}\sum_{g=1}^{G}(\boldsymbol{a}^\top \boldsymbol{\Xi}^{-1}\boldsymbol{X}_g^\top \boldsymbol{X}_g(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_0))^3 = \nu_a^{-3/2}\frac{1}{G}\sum_{g=1}^{G}(\boldsymbol{Z}_{3g}^\top(\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_0))^3.
\end{aligned}$$

It follows directly from (26), (B.30), and Assumption 5 that $B_{2j} = O_P(G^{-1/2})$ for $j = 1, 2, 3$.

Next, we find that $B_3 = B_{31} + B_{32} + B_{33}$, where

$$B_{31} = \nu_a^{-3/2} \frac{1}{G} \sum_{g=1}^{G} \boldsymbol{a}^\top (\ddot{\boldsymbol{\Xi}}^{-1} - \boldsymbol{\Xi}^{-1}) \boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g (\boldsymbol{a}^\top \ddot{\boldsymbol{\Xi}}^{-1} \boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g)^2,$$

$$B_{32} = \nu_a^{-3/2} \frac{1}{G} \sum_{g=1}^{G} \boldsymbol{a}^\top (\ddot{\boldsymbol{\Xi}}^{-1} - \boldsymbol{\Xi}^{-1}) \boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g \boldsymbol{a}^\top \ddot{\boldsymbol{\Xi}}^{-1} \boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g \boldsymbol{a}^\top \boldsymbol{\Xi}^{-1} \boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g,$$

$$B_{33} = \nu_a^{-3/2} \frac{1}{G} \sum_{g=1}^{G} \boldsymbol{a}^\top (\ddot{\boldsymbol{\Xi}}^{-1} - \boldsymbol{\Xi}^{-1}) \boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g (\boldsymbol{a}^\top \boldsymbol{\Xi}^{-1} \boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g)^2.$$

It follows from Lemma A.1 and Assumption 5 that $\ddot{\boldsymbol{\Xi}}^{-1} - \boldsymbol{\Xi}^{-1} = O_P(G^{-1/2})$. It is then straightforward to show, using the same arguments as applied to B_2 combined with (B.30) and Assumption 5, that $B_{3j} = O_P(G^{-1/2})$ for $j = 1, 2, 3$.

For the analysis of B_4, we first note that $\boldsymbol{\Sigma}_g = \mathrm{E}(\boldsymbol{W}_{1g} \boldsymbol{W}_{1g}^\top)$ and write

$$\ddot{\nu}_a - \nu_a = \frac{1}{G} \sum_{g=1}^{G} \Big((\boldsymbol{a}^\top \ddot{\boldsymbol{\Xi}}^{-1} \boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g)^2 - \mathrm{E}(\boldsymbol{a}^\top \boldsymbol{\Xi}^{-1} \boldsymbol{X}_g^\top \boldsymbol{u}_g)^2 \Big) = B_{41} + B_{42} + B_{43},$$

where

$$B_{41} = \frac{1}{G} \sum_{g=1}^{G} \Big((\boldsymbol{a}^\top \boldsymbol{\Xi}^{-1} \boldsymbol{W}_{1g})^2 - \mathrm{E}(\boldsymbol{a}^\top \boldsymbol{\Xi}^{-1} \boldsymbol{W}_{1g})^2 \Big),$$

$$B_{42} = \frac{1}{G} \sum_{g=1}^{G} \Big((\boldsymbol{a}^\top \boldsymbol{\Xi}^{-1} \boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g)^2 - (\boldsymbol{a}^\top \boldsymbol{\Xi}^{-1} \boldsymbol{W}_{1g})^2 \Big),$$

$$B_{43} = \frac{1}{G} \sum_{g=1}^{G} \Big((\boldsymbol{a}^\top \ddot{\boldsymbol{\Xi}}^{-1} \boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g)^2 - (\boldsymbol{a}^\top \boldsymbol{\Xi}^{-1} \boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g)^2 \Big).$$

We can write $B_{41} = \nu_a G^{-1} \sum_{g=1}^{G} (Z_{6g}^2 - \mathrm{E}(Z_{6g}^2))$, so that $B_{41} = O_P(G^{-1/2})$ by the same argument as applied to B_1. Next, for B_{42} we use the decomposition $\boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g = \boldsymbol{X}_g^\top \boldsymbol{u}_g - \boldsymbol{X}_g^\top \boldsymbol{X}_g (\ddot{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ and the same arguments as applied to B_2, which show that $B_{42} = O_P(G^{-1/2})$. Finally, we write

$$B_{43} = \frac{1}{G} \sum_{g=1}^{G} \boldsymbol{a}^\top (\ddot{\boldsymbol{\Xi}}^{-1} - \boldsymbol{\Xi}^{-1}) \boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g \boldsymbol{a}^\top (\ddot{\boldsymbol{\Xi}}^{-1} + \boldsymbol{\Xi}^{-1}) \boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g,$$

so that $B_{43} = O_P(G^{-1/2})$ by the same argument as applied to B_3. It follows that

$$\ddot{\nu}_a - \nu_a = O_P(G^{-1/2}). \tag{B.50}$$

Next, by Taylor-series expansion and using (B.50),

$$\ddot{\nu}_a^{-3/2} - \nu_a^{-3/2} = -\frac{3}{2} \nu_a^{-5/2} (\ddot{\nu}_a - \nu_a)(1 + O_P(G^{-1/2})), \tag{B.51}$$

which implies

$$B_4 = -\frac{3}{2} \nu_a^{-1} (\ddot{\nu}_a - \nu_a)(1 + O_P(G^{-1/2})) \frac{1}{G} \sum_{g=1}^{G} Z_{6g}^3.$$

The right-hand side is $O_P(G^{-1/2})$ by (B.30), (B.50), and Assumption 5. Finally, the result for B_5 follows by combining (B.51) with the same arguments as applied for B_2.

To prove the result for $\ddot{\xi}_{2,2}$, we proceed as in (B.49) and find

$$\ddot{\xi}_{2,2} = \mathrm{E}^*(v^{*4})\ddot{\nu}_a^{-2}\frac{1}{G}\sum_{g=1}^{G}\left(\boldsymbol{a}^\top\ddot{\boldsymbol{\Xi}}^{-1}\boldsymbol{X}_g^\top\ddot{\boldsymbol{u}}_g\right)^4 = \mathrm{E}^*(v^{*4})(\xi_{2,2} + C_1 + C_2 + C_3 + C_4 + C_5),$$

where C_i, for $i = 1, \ldots, 5$, are given by the same expressions as B_i, for $i = 1, \ldots, 5$, replacing the powers $-3/2$ and 3 in B_i by -2 and 4, respectively. Consequently, the proofs that $C_i = o_P(1)$, for $i = 1, \ldots, 5$, are nearly identical to those for the corresponding B_i given above, and are therefore omitted.

Appendix C: Additional Simulation Experiments

The linear regression model with clustered errors is very general. Thus, in principle, we could perform an infinite number of simulation experiments for it. Since that is infeasible, we have to make choices about which results to report. In Appendix C.1, we consider three key features of the DGP used in most of the simulation experiments in Section 4 of the paper and justify the choices that were made there. In Appendix C.2, we consider an extended version of the model that includes additional regressors.

C.1 Choice of Key Parameters

In this subsection, we present the results of three simulation experiments which focus on certain features of the data-generating processes used in the main experiments. Specifically, we look at the parameter ρ, the intra-cluster correlation for the error terms, the parameter γ, which determines how much cluster sizes vary, and alternative ways of generating the regressor, which primarily affect how much right skew it has.

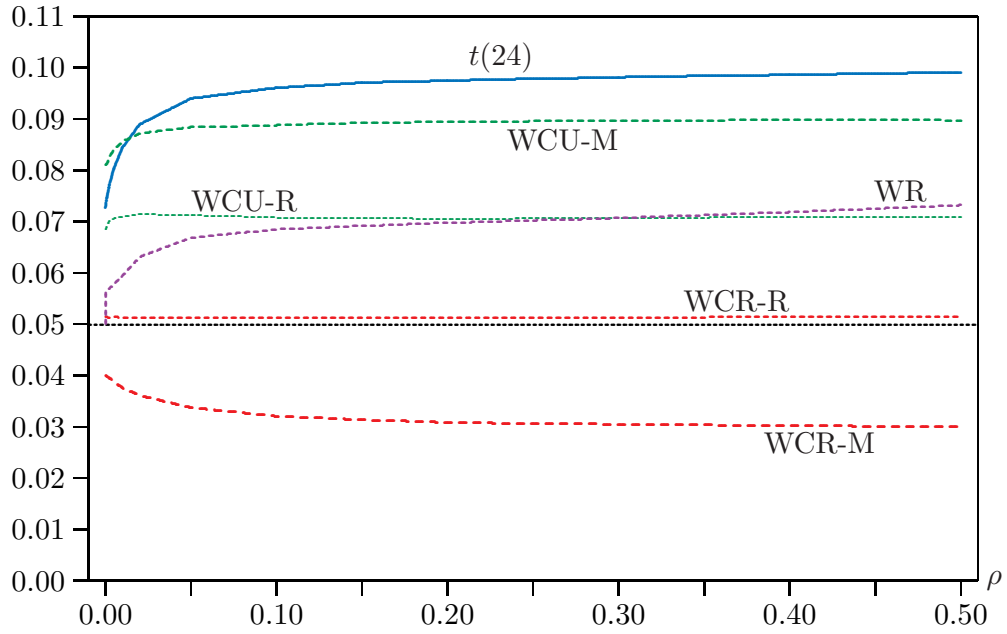
Figure C.1 shows rejection frequencies for the t-test, four variants of the wild cluster bootstrap, and the ordinary restricted wild bootstrap (WR) as a function of ρ, the within-cluster correlation coefficient. The other parameters are identical to the ones in Panel (b) of Figure 1, with $\rho_x = 0.7$. It might seem that ρ would be an important parameter. The value of ρ is indeed important in determining how efficiently the parameters of a regression model with clustered errors are estimated. As ρ increases, the amount of information contained in any given cluster diminishes. However, as Figure C.1 shows, rejection frequencies are not very sensitive to ρ for $\rho \geq 0.05$.

The t-test and the ordinary wild bootstrap are most sensitive to the value of ρ, but, even for them, rejection frequencies increase much more between 0 and 0.05 than between 0.05 and 0.50. The two WCB procedures that use the Rademacher distribution are almost totally insensitive to the value of ρ. The ones that use the Mammen distribution are a bit more sensitive to it, but not much so for $\rho \geq 0.05$. Based on the results in Figure C.1, ρ is set to 0.10 in all the simulation results reported in the paper. None of those results would have changed appreciably if we had used any value between 0.05 and 0.50.

As both the theory of Sections 2, 3 and 5 and the simulations of Section 4 show, variation in cluster sizes is very important. In most of our experiments, the cluster sizes are determined by (21), which depends on the parameter γ. In the paper, we report results for $\gamma = 0$ (equal-sized clusters) and $\gamma = 2$. Figure C.2 shows how rejection frequencies for the same six tests vary as a function of γ. Not surprisingly, they increase or decrease monotonically (allowing for simulation errors, which are small but noticeable even with 400,000 replications) as γ increases.

If we had used a value of γ larger than 2, we would evidently have obtained somewhat worse results for all methods. WCR-M would have underrejected slightly more severely, and all other

Figure C.1: Rejection frequencies at 0.05 level, $G = 25$, $N = 2500$, $\rho_x = 0.7$, $\gamma = 2$



methods would have overrejected more severely (although the effect is very modest for WCR-R). However, our view is that values greater than 2 are not very realistic. For $G = 25$, the ratio of the largest to the smallest cluster is 7.3 when $\gamma = 2$. It increases to 19.4 when $\gamma = 3$ and to 48.1 when $\gamma = 4$. Moreover, since Assumption 3 of Theorem 2.1 limits the extent of heterogeneity of cluster sizes, albeit not in a way that can be stated explicitly for any finite sample size, using large values of γ in the simulations would risk relying on simulation results for cases that are not actually covered by the theory.

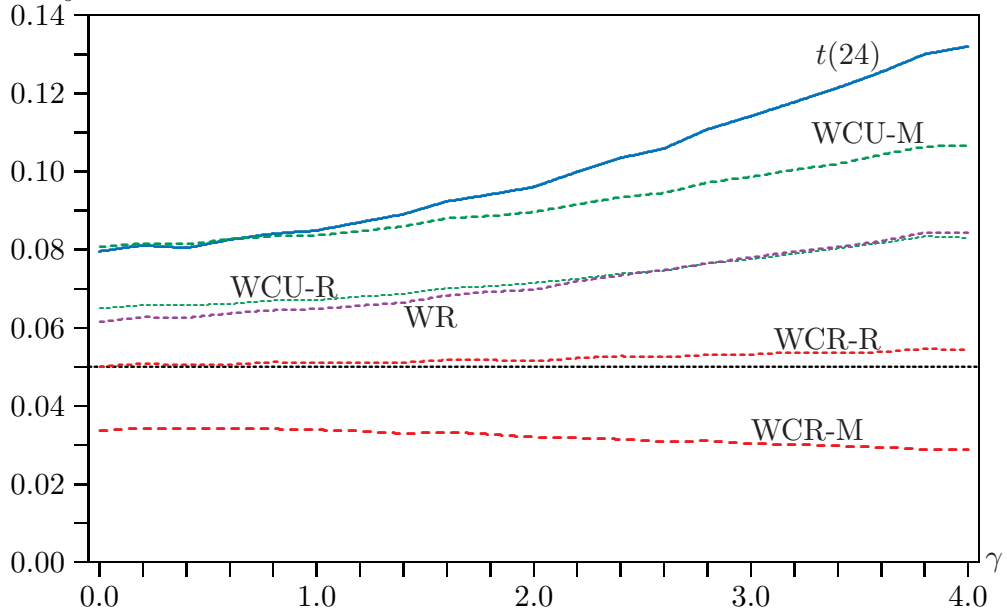
In all the experiments reported in the paper, the regressor is generated as a weighted sum of $\chi^2(8)$ random variates, recentered and rescaled to have mean 0 and variance 1, with the weights chosen so that the intra-cluster correlation is ρ_x. In generating the regressor in this way, our objective is to make it leptokurtic and right-skewed, but not excessively so.

In Figure C.3, we report rejection frequencies for the same six tests for eight ways of generating the regressor. These rejection frequencies are not very sensitive to the choice of distribution, except when skewness is relatively extreme: It is 1 for $\chi^2(8)$, 1.41 for $\chi^2(4)$, 2 for $\chi^2(2)$, 2.83 for $\chi^2(1)$, and 6.18 for the standard lognormal distribution. None of our results would have changed very much if we had used $\chi^2(4)$ or $\chi^2(16)$ instead of $\chi^2(8)$. Some of them would have changed noticeably if we had used the lognormal distribution, which is, in our view, much too extreme. If we had used a moderately extreme distribution, like $\chi^2(2)$, the WCR-R, WCR-M, and WR tests would have performed about the same as in the paper, but the t-test and the two unrestricted WCB tests would have overrejected more severely.

C.2 Additional Regressors

The model (20) used in all the simulation experiments in the paper has just one regressor and a constant term. This means that the restricted residuals are simply deviations from a sample mean. It is therefore natural to speculate that the excellent performance of WCR-R observed in

Figure C.2: Rejection frequencies at 0.05 level, $G = 25$, $N = 2500$, $\rho = 0.10$, $\rho_x = 0.7$



Section 4 of the paper, and in Figures C.1–C.3, may be a consequence of that feature of the model.[1] In this subsection, we therefore consider a model with additional regressors. As we shall see, the performance of all methods deteriorates, but WCR-R continues to perform relatively well.

The model we study here is

$$\boldsymbol{y}_g = \beta_1 + \beta_2 \boldsymbol{x}_g + \sum_{j=1}^{J} \delta_j \boldsymbol{z}_g^{(j)} + \boldsymbol{u}_g, \quad \mathrm{E}(\boldsymbol{u}_g \boldsymbol{u}_g^\top) = \boldsymbol{\Omega}_g, \quad g = 1, \ldots, G, \tag{C.1}$$

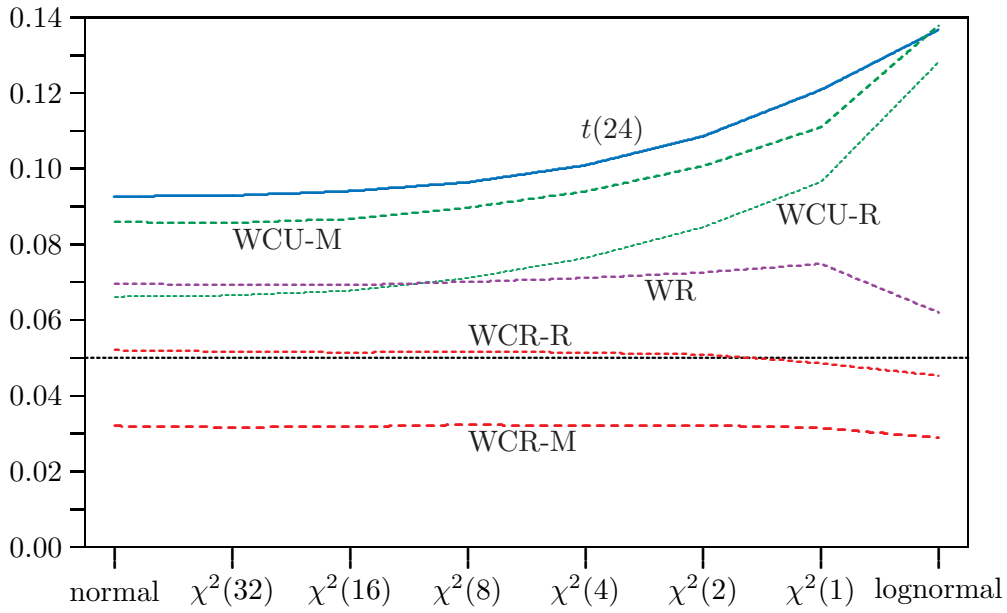
where both the $\boldsymbol{u}_g$ and the regressor of interest $\boldsymbol{x}_g$ are distributed as described in Section 4, with intra-cluster correlations 0.1 and ρ_x, respectively. Each of the J additional regressors $\boldsymbol{z}_g^{(j)}$ is normally distributed, uncorrelated with all the other regressors, and uncorrelated across clusters, with intra-cluster correlation ϕ. We performed experiments for three values of J (2, 4, and 8) and five values of ϕ (0.00, 0.25, 0.50, 0.75, and 1.00).

Figure C.4 shows rejection frequencies for cluster-robust t-tests at the 0.05 level for $G = 25$, $N = 2500$, $\rho = 0.10$, $\gamma = 2$, and 21 values of ρ_x between 0 and 1. This is the same case as Panel (b) of Figure 1 in the paper. In Panel (a) of Figure C.4, $J = 4$, and in Panel (b), $J = 8$. In both panels, the lowest curve (for $\phi = 0$) is almost identical to the corresponding curve in Panel (b) of Figure 1 in the paper. For example, the rejection frequency for $\rho_x = 1$ increases from 0.1045 with no extra regressors to 0.1065 with either 4 or 8 extra regressors. Thus, for this model, adding additional regressors that vary only at the individual level appears to have almost no effect.

In contrast, except when $\rho_x = 0$, adding additional regressors that vary partly or entirely at the cluster level increases rejection frequencies substantially. There is evidently an important interaction between the regressor of interest and the additional regressors when they both exhibit intra-cluster correlation. In the worst case, for $J = 8$ and $\rho_x = 1$, the rejection frequency rises from 0.1045 to 0.1879.

[1] However, if this were the case, one might expect WCR-M to work equally well, and it does not.

Figure C.3: Rejection frequencies at 0.05 level, $G = 25$, $N = 2500$, $\gamma = 2$, $\rho = 0.10$, $\rho_x = 0.7$



It would be impractical to present results for the bootstrap methods for all 10 cases shown in Figure C.4. We therefore focus on two cases, one moderate and one quite extreme. In the first of them, $J = 4$ and $\phi = 0.5$. This corresponds to the middle curve in Panel (a). In the second case, $J = 8$ and $\phi = 1.0$. This corresponds to the top curve in Panel (b). In the second case, we are using just 25 clusters to estimate nine coefficients on regressors that vary only at the cluster level.

Figure C.5, which is directly comparable to Panel (b) of Figure 1 in the paper, shows results for the t-test, all four WCB tests, and the ordinary wild bootstrap (WR) test for these two cases. In both panels, the ordering of the WCB tests is the same as in Figure 1 in the paper. They all reject more often than they did previously. This actually improves the performance of WCR-M, but not of the other tests.

Figure C.4: Rejection frequencies for t-tests at 0.05 level, $G = 25$, $N = 2500$, $\gamma = 2$, $\rho = 0.10$

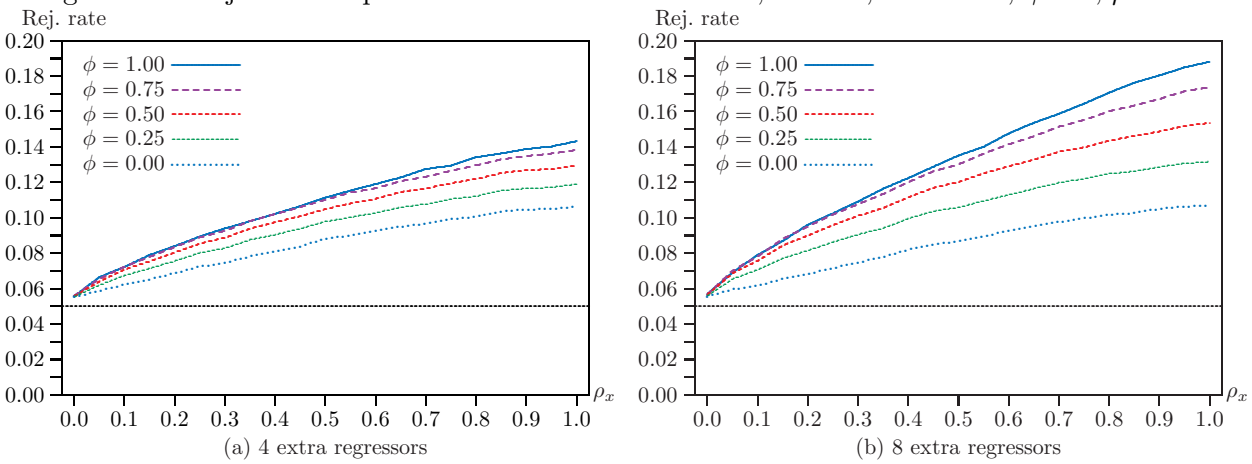
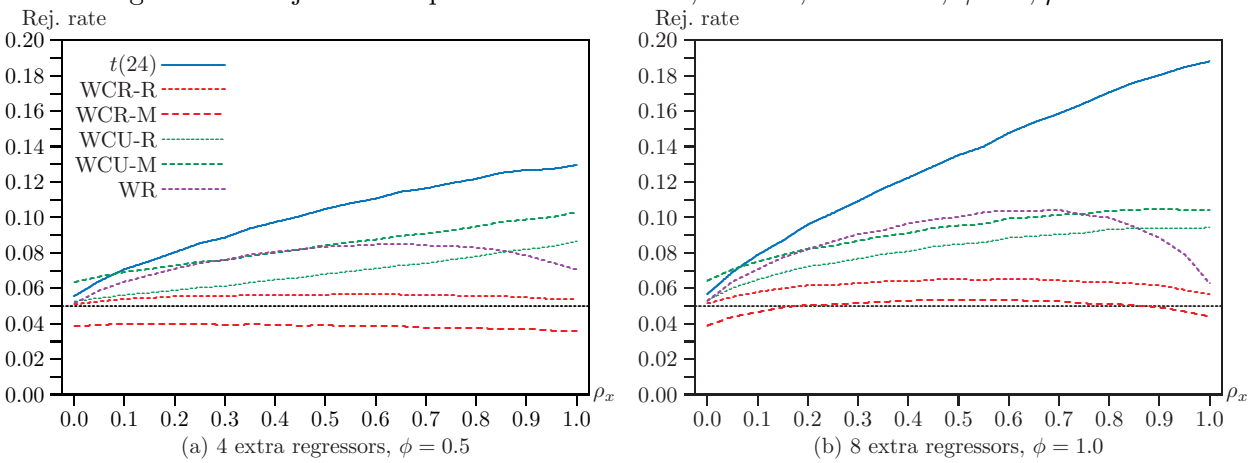


Figure C.5: Rejection frequencies at 0.05 level, $G = 25$, $N = 2500$, $\gamma = 2$, $\rho = 0.10$



(a) 4 extra regressors, $\phi = 0.5$ (b) 8 extra regressors, $\phi = 1.0$

In Panel (a) of Figure C.5, where $J = 4$ and $\phi = 0.5$, all tests reject more often than they did previously, but the differences are not large. WCR-R remains the best method, but it now overrejects (albeit very slightly) for all values of ρ_x. In the worst case, for $\rho_x = 0.60$, it rejects 5.67% of the time. WCR-M continues to underreject for all values of ρ_x, a bit more than before when ρ_x is small, but not quite as much when ρ_x is large. The two unrestricted WCB tests perform somewhat worse than they did before, especially when ρ_x is large. For example, WCU-M rejects 10.27% of the time when $\rho_x = 1$, versus 9.25% in Panel (b) of Figure 1 in the paper.

In Panel (b) of Figure C.5, where $J = 8$ and $\phi = 1.0$, all tests reject more often than in Panel (a). This is most noticeable for the t-test, the WR bootstrap test, and the two restricted WCB tests. Perhaps surprisingly, WCR-M is now the best test for most values of ρ_x. There seem to be two factors at work here. One is whatever causes WCR-M to underreject in all other cases, and the other is whatever is causing all the tests to reject more often in this case. The net effect is that WCR-M can either underreject or overreject, but it generally does so only modestly. Nevertheless, even though WCR-R is no longer the best test throughout Panel (b), it performs reasonably well. In the worst case, when $\rho_x = 0.60$, it rejects 6.52% of the time. For WCU-R and WCU-M, the rejection frequency curves, which are essentially straight lines in Panel (a), are clearly concave in Panel (b). For intermediate values of ρ_x, they overreject quite a bit more severely than before, but for extreme values their rejection frequencies do not change much.

One interesting result in both panels is that the ordinary wild bootstrap test (WR) overrejects noticeably more often than it did before. It is the worst bootstrap test for a range of intermediate values of ρ_x, albeit only slightly worse than WCU-M. The rejection frequency curve still has an inverted U shape, but the peak in the middle is much higher than it was in Figure 1 of the paper. These results suggest that it may become more important for the bootstrap DGP to match the actual structure of the clusters as the number of regressors that display substantial intra-cluster correlation increases.

C.3 Concluding Remarks

The results in Figures C.1–C.3 suggest that the choices of certain key parameters in the simulation experiments reported in the paper either have little effect on the results or have moderate but predictable effects that do not change the ordering of rejection frequencies for any of the tests.

The results in Figures C.4 and C.5 suggest that using a model with more than one regressor and a constant term would have had more substantial effects. In particular, all methods would have rejected more frequently, especially the t-test, the WR test, and the two unrestricted WCB tests. The effects on WCR-R and WCR-M would have been much smaller, with the latter no longer underrejecting in all cases.

References for Supplementary Material

Bhattacharya, R. N., and J. K. Ghosh. 1978. On the validity of the formal Edgeworth expansion. *Annals of Statistics* 6:434–451.

Bhattacharya, R. N., and R. R. Rao. 1976. *Normal Approximation and Asymptotic Expansions*. Philadelphia: SIAM.

Djogbenou, A. A., J. G. MacKinnon, and M. Ø. Nielsen. 2019. Asymptotic theory and wild bootstrap inference with clustered errors. *Journal of Econometrics* to appear.

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

Skovgaard, I. B. 1981. Transformation of an Edgeworth expansion by a sequence of smooth functions. *Scandinavian Journal of Statistics* 8:207–217.