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

Working Paper

Testing for the appropriate level of clustering in linear regression models

Queen's Economics Department Working Paper, No. 1428

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

Suggested Citation: MacKinnon, James G.; Nielsen, Morten Ørregaard; Webb, Matthew (2020) : Testing for the appropriate level of clustering in linear regression models, Queen's Economics Department Working Paper, No. 1428, Queen's University, Department of Economics, Kingston (Ontario)

This Version is available at:
https://hdl.handle.net/10419/230581

Queen's Economics Department Working Paper No. 1428

Testing for the Appropriate Level of Clustering in Linear Regression Models

James G. MacKinnon
Queen's University

Morten Ørregaard Nielsen
Queen's University
and CREATES

Matthew D. Webb
Carleton University

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

10-2020

Testing for the appropriate level of clustering in linear regression models*

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

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

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

October 30, 2020

Abstract

The overwhelming majority of empirical research that uses cluster-robust inference assumes that the clustering structure is known, even though there are often several possible ways in which a dataset could be clustered. We propose two tests for the correct level of clustering in regression models. One test focuses on inference about a single coefficient, and the other on inference about two or more coefficients. We provide both asymptotic and wild bootstrap implementations. The proposed tests work for a null hypothesis of either no clustering or "fine" clustering against alternatives of "coarser" clustering. We also propose a sequential testing procedure to determine the appropriate level of clustering. Simulations suggest that the bootstrap tests perform very well under the null hypothesis and can have excellent power. An empirical example suggests that using the tests leads to sensible inferences.

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

JEL Codes: C12, C15, C21, C23.

*We are grateful to participants at the 2018 Canadian Economics Association conference, the 2018 Canadian Econometric Study Group conference, the 2020 Econometric Society World Congress, the Université de Montréal, the University of Exeter, UCLA, and Michigan State University for comments. MacKinnon and Webb thank the Social Sciences and Humanities Research Council of Canada (SSHRC grant 435-2016-0871) for financial support. Nielsen thanks the Canada Research Chairs program and the SSHRC (grant 435-2017-0131) for financial support. Computer code for performing the testing procedures proposed here may be found at http://qed.econ.queensu.ca/pub/faculty/mackinnon/cluster-test/.

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

1 Introduction

The presence of within-cluster correlation can have very serious consequences for statistical inference. Theoretical work on cluster-robust inference almost always assumes that the structure of the clusters is known, even though the form of the correlation within clusters is arbitrary. Unless it is obvious that clustering must be at a certain level, however, this can leave empirical researchers in a difficult situation. They must generally rely on rules of thumb, their own intuition, or referees' suggestions to decide how the observations should be clustered. To make this process easier, we propose a test for any given level of clustering (including no clustering as a special case) against an alternative within which it is nested. When two or more levels of clustering are possible, we propose a sequence of such tests.

There has been a great deal of research on cluster-robust inference in the past two decades. Cameron and Miller (2015) is a very thorough survey that covers much of the literature up to a few years ago. More recent surveys include MacKinnon (2019), MacKinnon and Webb (2020), and Conley, Gonçalves, and Hansen (2018), which considers a broader class of methods for dependent data. Areas that have received particular attention include: asymptotic theory for cluster-robust inference (Djogbenou, MacKinnon, and Nielsen 2019; Hansen and Lee 2019); bootstrap methods with clustered data (Cameron, Gelbach, and Miller 2008; Djogbenou, MacKinnon, and Nielsen 2019; Roodman, MacKinnon, Nielsen, and Webb 2019); and inference with unbalanced clusters (Imbens and Kolesár 2016; Carter, Schnepel, and Steigerwald 2017; MacKinnon and Webb 2017; Djogbenou, MacKinnon, and Nielsen 2019).

Almost all of this literature assumes that the way in which observations are allocated to clusters is known to the econometrician. This is quite a strong assumption. Imagine that a dataset has many observations taken from individuals in different geographical locations. In order to utilize a cluster-robust variance estimator (CRVE), the researcher needs to specify at what level the clustering occurs. For example, there could possibly be clustering at the zip-code, city, county, state, or country level. Even in this relatively simple setting, there are many possible ways in which a researcher could 'cluster' the standard errors.

A few rules of thumb have emerged to cover some common cases. For instance, in the case of nested clusters, such as cities within states, Cameron and Miller (2015) advocates clustering at the larger, more aggregate level. In the case of randomized experiments, Athey and Imbens (2017) recommends clustering at the level of randomization. Similarly, in the case of experiments where treatment is assigned to groups in pairs, with one group treated and one not treated, de Chaisemartin and Ramirez-Cuellar (2020) recommends clustering at the pair level rather than the group level. While these rules of thumb are sometimes helpful, they may or may not be correct in any particular case.

Getting the level of clustering correct is extremely important. Simulation results in several papers have shown that ignoring clustering in a single dimension can result in rejection frequencies for tests at the 5% level that are actually well over 50% (Bertrand, Duflo, and Mullainathan 2004; Cameron, Gelbach, and Miller 2008) and confidence intervals that are too narrow by a factor of five or more (MacKinnon 2019). Similarly, simulations in MacKinnon and Webb (2020) show that clustering at too fine a level can result in gross size distortions. On the other hand, clustering at too coarse a level (say state-level clustering when there is actually city-level clustering) can lead to the problems associated with having few treated clusters, which can be very severe (MacKinnon and Webb 2017, 2018), and can also reduce power (MacKinnon and Webb 2020). Additionally, Abadie, Athey, Imbens, and Wooldridge (2017) suggest that clustering can be too conservative in situations where there is neither a clustered sample design nor cluster-specific treatment assignment.

In Section 3, we propose two tests for the cluster structure of the error variance matrix in a linear regression model. They test the null hypothesis of a fine level of clustering (or of no clustering at all) against an alternative hypothesis with a coarser level of clustering. The tests are based on the difference between two functions of the scores for the parameter(s) of interest. These functions are essentially the filling in the sandwich for two different cluster-robust variance estimators, one associated with the null level of clustering and one associated with the alternative level. Since the functions estimate the variance of the scores under two different clustering assumptions, we refer to the tests as score-variance tests. A procedure for sequential testing, described in Section 3.3, allows for determination of the appropriate level of clustering without inflating the family-wise error rate when there are several possible levels of clustering.

Ibragimov and Müller (2016) also proposes a test for the appropriate level of clustering. Their test, which involves simulation and is based in part on the procedure of Ibragimov and Müller (2010), requires that the model of interest be estimable on a (coarse) cluster-by-cluster basis. Unfortunately, this requirement is often not satisfied in difference-in-differences or randomized experiments where only certain groups are treated, or more generally when the regressor of interest is invariant within clusters. In contrast, our tests can be performed in any setting where it is possible to compute a CRVE for the whole sample for each level of clustering. The IM test, as we will refer to it, is described in detail in Appendix B.

The model of interest is discussed in Section 2. Our score-variance tests are described in Section 3, including the bootstrap implementation and the sequential testing procedure. Section 4 provides asymptotic theory for the two test statistics, the bootstrap tests, and the sequential testing procedure. The size and power of the proposed tests are analyzed by Monte Carlo simulations in Section 5. An empirical example that uses the STAR dataset (Finn

and Achilles 1990; Mosteller 1995) is discussed in Section 6. Finally, Section 7 concludes and offers some guidance for empirical researchers. All mathematical proofs are given in Appendix A, and additional simulation results are presented in Appendix C.

2 The Regression Model with Clustering

We focus on the linear regression model

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

where $\boldsymbol{y}$ and $\boldsymbol{u}$ are $N \times 1$ vectors of observations and error terms (or disturbances), and $\boldsymbol{X}$ is an $N \times K$ matrix of covariates. The coefficients on the regressors are in the $K \times 1$ parameter vector $\boldsymbol{\beta}$.

Suppose that the data are divided into G clusters, indexed by g, where the g^{th} cluster has N_g observations, so that $N = \sum_{g=1}^{G} N_g$. Thus, there are G vectors $\boldsymbol{y}_g$ and $\boldsymbol{u}_g$ of size N_g, along with G matrices $\boldsymbol{X}_g$, each with N_g rows and K columns. Using this notation, the OLS estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{\beta}_0 + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u} = \boldsymbol{\beta}_0 + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \sum_{g=1}^{G} \boldsymbol{X}_g^\top \boldsymbol{u}_g, \tag{2}$$

where $\boldsymbol{\beta}_0$ denotes the true value of $\boldsymbol{\beta}$. Now define the $K \times 1$ score vectors $\boldsymbol{s}_g = \boldsymbol{X}_g^\top \boldsymbol{u}_g$. We assume that these score vectors satisfy $\mathrm{E}(\boldsymbol{s}_g) = \boldsymbol{0}$ for all g and

$$\mathrm{E}(\boldsymbol{s}_g \boldsymbol{s}_{g'}^\top) = \mathbb{I}(g = g') \boldsymbol{V}_g, \quad g, g' = 1, \ldots, G, \tag{3}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function and $\boldsymbol{V}_g$ is a $K \times K$ variance matrix. If, in addition, we assumed that the regressors were exogenous, we could write

$$\mathrm{E}(\boldsymbol{u}_g | \boldsymbol{X}) = \boldsymbol{0} \quad \text{and} \quad \mathrm{E}(\boldsymbol{u}_g \boldsymbol{u}_{g'}^\top | \boldsymbol{X}) = \mathbb{I}(g = g') \boldsymbol{\Omega}_g, \quad g, g' = 1, \ldots, G, \tag{4}$$

where $\boldsymbol{\Omega}_g$ is an $N_g \times N_g$ variance matrix that forms a diagonal block of $\boldsymbol{\Omega} = \mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top | \boldsymbol{X})$. However, since the tests we will propose in Section 3 do not require the regressors to be exogenous, we maintain only the weaker assumption in (3).

It is clear from (2) that the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ depends on the properties of the score vectors. An estimator of the variance matrix of $\hat{\boldsymbol{\beta}}$ is given by the sandwich formula

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \hat{\boldsymbol{V}} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}, \tag{5}$$

where $\hat{\boldsymbol{V}}$ is an estimator of the variance matrix of the sum of scores, $\boldsymbol{V} = \mathrm{E}(\boldsymbol{X}^\top \boldsymbol{u}\boldsymbol{u}^\top \boldsymbol{X})$. The condition (3) implies that $\mathrm{E}(\boldsymbol{s}_g \boldsymbol{s}_{g'}^\top) = \boldsymbol{0}$ whenever $g \neq g'$. In this case $\boldsymbol{V} = \sum_{g=1}^{G} \boldsymbol{V}_g$, so that the usual estimator for $\boldsymbol{V}$ under condition (3) is

$$\hat{\boldsymbol{V}}_{\mathrm{c}} = m_{\mathrm{c}} \sum_{g=1}^{G} \boldsymbol{X}_g^\top \hat{\boldsymbol{u}}_g \hat{\boldsymbol{u}}_g^\top \boldsymbol{X}_g = m_{\mathrm{c}} \sum_{g=1}^{G} \hat{\boldsymbol{s}}_g \hat{\boldsymbol{s}}_g^\top, \tag{6}$$

where $\hat{\boldsymbol{u}}_g$ contains the residuals for cluster g and $\hat{\boldsymbol{s}}_g = \boldsymbol{X}_g^\top \hat{\boldsymbol{u}}_g$ is the $K \times 1$ vector of empirical scores for cluster g. The scalar factor m_{c} is a finite-sample correction, the most commonly employed factor being $m_{\mathrm{c}} = G/(G-1) \times (N-1)/(N-K)$, which is designed to account for degrees of freedom. Using $\hat{\boldsymbol{V}} = \hat{\boldsymbol{V}}_{\mathrm{c}}$ in (5) yields the most widely-used CRVE for $\hat{\boldsymbol{\beta}}$. Asymptotic inference on regression coefficients using the resulting CRVE is studied by Djogbenou, MacKinnon, and Nielsen (2019) and Hansen and Lee (2019).

Remark 1. In the special case in which each cluster has $N_g = 1$ observation, we can use

$$\hat{\boldsymbol{V}}_{\mathrm{het}} = \sum_{i=1}^{N} \hat{u}_i^2 \boldsymbol{X}_i^\top \boldsymbol{X}_i = \boldsymbol{X}^\top \mathrm{diag}(\hat{u}_1^2, \ldots, \hat{u}_N^2) \boldsymbol{X}, \tag{7}$$

where $\boldsymbol{X}_i$ is the i^{th} row of the $\boldsymbol{X}$ matrix and $\hat{u}_i$ is the i^{th} residual. The variance matrix obtained by setting $\hat{\boldsymbol{V}} = \hat{\boldsymbol{V}}_{\mathrm{het}}$ in (5) is the famous heteroskedasticity-consistent variance matrix estimator (HCCME) of Eicker (1963) and White (1980). Of course, the matrix $\hat{\boldsymbol{V}}_{\mathrm{het}}$ can be modified in various ways to improve its finite-sample properties (MacKinnon 2013), the simplest of which is to multiply it by $m_{\mathrm{het}} = N/(N-K)$. □

Remark 2. Abadie, Athey, Imbens, and Wooldridge (2017) have argued that whether or not to cluster, and at what level, depends fundamentally on assumptions about the sampling procedure. Our tests are based on what they call the "model-based" approach, as opposed to the "design-based" approach that they develop in detail. The model-based approach makes sense if we think of the sample as a random outcome from some sort of meta-population (or DGP), and the coefficients of interest as features of that meta-population; see MacKinnon, Nielsen, and Webb (2020) for additional details.

In contrast, the design-based approach makes sense when the investigator is concerned with the characteristics of a finite sample from a meta-population, and the observed sample constitutes a substantial proportion of that finite sample. In this case, Abadie et al. (2017) show that, depending on how treatment was assigned, it may be appropriate to use heteroskedasticity-robust standard errors rather than cluster-robust ones even when the latter are substantially larger than the former. Our tests are not designed for such a setting. □

3 The Testing Procedure

The fundamental idea of our testing procedure is to compare two estimates of the variance of the coefficient(s) that we want to estimate. We test the null hypothesis that a CRVE based on a "fine" clustering structure is valid against the alternative that the CRVE needs to be based on a "coarser" clustering structure. Since it is only the filling in the sandwich (5) that differs across different clustering structures, we are actually comparing two estimates of the variance matrix of the sum of scores. Our procedure is somewhat like the specification test of Hausman (1978). The "fine" CRVE is efficient when there actually is fine clustering, but it is invalid when there is coarse clustering. In contrast, the "coarse" CRVE is inefficient when there actually is fine clustering, but it is valid in both cases.

In practical applications, the number of coefficients in (1), and hence the size of the CRVE matrices, is often large, so comparing these matrices directly can be impractical. Furthermore, it is usually only a small subset of the coefficients that is actually of interest. Thus, suppose the regressors are partitioned as $\boldsymbol{X} = [\boldsymbol{X}_1,\ \boldsymbol{X}_2]$, where $\boldsymbol{X}_1$ denotes the $N \times k$ matrix of the regressors of interest and $\boldsymbol{X}_2$ denotes the $N \times (K-k)$ matrix of other regressors. Similarly, partition $\boldsymbol{\beta}^\top = [\boldsymbol{\beta}_1^\top,\ \boldsymbol{\beta}_2^\top]$, where the coefficients corresponding to the regressors of interest are in the $k \times 1$ parameter vector $\boldsymbol{\beta}_1$. In practice, many coefficients correspond to fixed effects and other conditioning variables that are not of primary interest, and these are collected in $\boldsymbol{\beta}_2$. If the coefficient vector of interest is actually a linear combination of the elements of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, we can redefine $\boldsymbol{X}$ as a nonsingular affine transformation of the original $\boldsymbol{X}$ matrix, so that $\boldsymbol{\beta}_1$ has the desired interpretation.

We regress each column of $\boldsymbol{X}_1$ on $\boldsymbol{X}_2$ and define $\boldsymbol{Z}$ as the matrix of residuals from those k regressions. The model (1) can then be rewritten as

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\delta} + \boldsymbol{u}, \qquad \boldsymbol{Z} = \boldsymbol{M}_{\boldsymbol{X}_2}\boldsymbol{X}_1, \tag{8}$$

where $\boldsymbol{M}_{\boldsymbol{X}_2} = \mathbf{I}_N - \boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top$ is the orthogonal projection matrix that projects off (or partials out) $\boldsymbol{X}_2$. The regressor matrices $\boldsymbol{Z}$ and $\boldsymbol{X}_2$ are orthogonal, and the models (1) and (8) have exactly the same explanatory power and the same errors, $\boldsymbol{u}$. The coefficient vector $\boldsymbol{\beta}_1$ in (8) is identical to the one defined in the previous paragraph, but the coefficient vector $\boldsymbol{\delta}$ is different from $\boldsymbol{\beta}_2$. Using the orthogonality between $\boldsymbol{Z}$ and $\boldsymbol{X}_2$, the OLS estimate of $\boldsymbol{\beta}_1$ is, c.f. (2),

$$\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{y} = \boldsymbol{\beta}_{1,0} + (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\sum_{g=1}^{G}\boldsymbol{Z}_g^\top\boldsymbol{u}_g. \tag{9}$$

To derive an estimate of the variance matrix of $\hat{\boldsymbol{\beta}}_1$, we make use of the relations

$$\begin{aligned} &\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{A}\big(1 + o_P(1)\big) \quad \text{with} \\ &\boldsymbol{Q} = [\mathbf{I}_k,\ -\boldsymbol{X}_1^\top \boldsymbol{X}_2 (\boldsymbol{X}_2^\top \boldsymbol{X}_2)^{-1}]^\top \text{ and } \boldsymbol{A} = [\mathbf{I}_k,\ -\boldsymbol{\Xi}_{12}\boldsymbol{\Xi}_{22}^{-1}]^\top = \operatorname*{plim}_{N\to\infty} \boldsymbol{Q}, \end{aligned} \tag{10}$$

where $\boldsymbol{A}$ is well defined and finite under Assumption 3 (to be stated below). Here $\boldsymbol{\Xi}_{12}$ and $\boldsymbol{\Xi}_{22}$ denote submatrices of the matrix $\boldsymbol{\Xi}$ to which $N^{-1}\boldsymbol{X}^\top\boldsymbol{X}$ tends as $N \to \infty$. Then, similarly to (5), we obtain the sandwich formula

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_1) = (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\hat{\boldsymbol{\Sigma}}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}, \tag{11}$$

where $\hat{\boldsymbol{\Sigma}}$ is an estimate of

$$\boldsymbol{\Sigma} = \boldsymbol{A}^\top \boldsymbol{V} \boldsymbol{A}, \tag{12}$$

which depends on the clustering structure in the same way as $\boldsymbol{V}$.

Our testing procedure is based on comparing two CRVE's for $\hat{\boldsymbol{\beta}}_1$, corresponding to a "fine" and a "coarse" clustering structure. To make the procedure operational, we formulate the hypotheses in terms of the parameters of the model. To this end, we first define some notation. There are G coarse clusters indexed by $g = 1, \ldots, G$. Within coarse cluster g, there are M_g fine clusters indexed by $h = 1, \ldots, M_g$. In total there are $G_\mathrm{f} = \sum_{g=1}^{G} M_g$ fine clusters. Fine cluster h in coarse cluster g contains N_{gh} observations indexed by $i = 1, \ldots, N_{gh}$. Coarse cluster g therefore contains $N_g = \sum_{h=1}^{M_g} N_{gh}$ observations, and the entire sample contains $N = \sum_{g=1}^{G} N_g = \sum_{g=1}^{G}\sum_{h=1}^{M_g} N_{gh}$ observations. Further, we let $\boldsymbol{X}_{ghi}$ and u_{ghi} denote the regressors and disturbance for observation i within fine cluster h in coarse cluster g. We then define the corresponding score as $\boldsymbol{s}_{ghi} = \boldsymbol{X}_{ghi}^\top u_{ghi}$, the score for fine cluster h in coarse cluster g as $\boldsymbol{s}_{gh} = \sum_{i=1}^{N_{gh}} \boldsymbol{s}_{ghi}$, and the score for coarse cluster g as $\boldsymbol{s}_g = \sum_{h=1}^{M_g} \boldsymbol{s}_{gh}$.

Under the coarse clustering structure, the $\boldsymbol{s}_g$ satisfy (3), so that in particular they are uncorrelated across g. Under the fine clustering structure, the $\boldsymbol{s}_{gh}$ are themselves uncorrelated across h, for each g. That is, for all $g = 1, \ldots, G$,

$$\mathrm{E}(\boldsymbol{s}_{gh}\boldsymbol{s}_{gh'}^\top) = \mathbb{I}(h = h')\boldsymbol{V}_{gh}, \quad h, h' = 1, \ldots, M_g, \tag{13}$$

where each of the $\boldsymbol{V}_{gh}$ is a $K \times K$ matrix. Equations (3) and (13) embody the assumption that the coarse and fine clustering structures are nested.

Now let $\boldsymbol{\Sigma}_\mathrm{c}$ and $\boldsymbol{\Sigma}_\mathrm{f}$ denote the matrix in (12) under the coarse and fine clustering structures, respectively. From (3), (12), and (13), these matrices are

$$\boldsymbol{\Sigma}_\mathrm{c} = \sum_{g=1}^{G} \boldsymbol{\Sigma}_g \quad \text{and} \quad \boldsymbol{\Sigma}_\mathrm{f} = \sum_{g=1}^{G}\sum_{h=1}^{M_g} \boldsymbol{\Sigma}_{gh}, \tag{14}$$

where, as in (12),

$$\boldsymbol{\Sigma}_g = \boldsymbol{A}^\top \boldsymbol{V}_g \boldsymbol{A} \quad \text{and} \quad \boldsymbol{\Sigma}_{gh} = \boldsymbol{A}^\top \boldsymbol{V}_{gh} \boldsymbol{A}. \tag{15}$$

We consider the null and alternative hypotheses

$$\mathrm{H}_0 \colon \lim_{N\to\infty} \boldsymbol{\Sigma}_\mathrm{f} \boldsymbol{\Sigma}_\mathrm{c}^{-1} = \mathbf{I} \quad \text{and} \quad \mathrm{H}_1 \colon \lim_{N\to\infty} \boldsymbol{\Sigma}_\mathrm{f} \boldsymbol{\Sigma}_\mathrm{c}^{-1} \neq \mathbf{I}. \tag{16}$$

The hypotheses are expressed in this way, rather than in terms of the difference between the limits of normalized versions of $\boldsymbol{\Sigma}_\mathrm{f}$ and $\boldsymbol{\Sigma}_\mathrm{c}$, because the appropriate normalizing factors will, in general, be unknown; see Djogbenou, MacKinnon, and Nielsen (2019).

Remark 3. In (16), we are not directly testing the fine clustering condition in (13). Instead, we are testing an important implication of the clustering structure. Specifically, we test whether $\boldsymbol{\Sigma}_\mathrm{c} = \boldsymbol{\Sigma}_\mathrm{f}$, which implies that a valid CRVE for $\hat{\boldsymbol{\beta}}_1$ is given by (11) with $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}_\mathrm{f}$. □

Remark 4. An important null hypothesis is that the HCCME (7) considered in Remark 1 is in fact valid, and no CRVE is needed. We note that this hypothesis is a special case of (16) in which each fine cluster has one observation, i.e. $N_{gh} = 1$ for all g and h. □

3.1 Test Statistics

Our score-variance test statistics are based on comparing estimates $\hat{\boldsymbol{\Sigma}}_\mathrm{f}$ and $\hat{\boldsymbol{\Sigma}}_\mathrm{c}$ obtained under fine and coarse clustering, respectively. There are many ways in which one could compare these $k \times k$ matrices. We focus on two quantities of particular interest, which define two test statistics. The first is obtained for $k = 1$, so that interest is focused on a particular coefficient that we are trying to make inferences about. This leads to a test statistic that has the form of a t-statistic. The second is obtained for $k > 1$, in which case our test statistic is a quadratic form involving all the unique elements of $\hat{\boldsymbol{\Sigma}}_\mathrm{f}$ and $\hat{\boldsymbol{\Sigma}}_\mathrm{c}$, as in White's (1980) "direct test" for heteroskedasticity. The first test is of course a special case of the second, but we treat it separately because it is particularly simple to compute and may often be of primary interest.

In order to derive the test statistics, we write $\hat{\boldsymbol{\Sigma}}_\mathrm{c}$ and $\hat{\boldsymbol{\Sigma}}_\mathrm{f}$ using common notation. Let

$$\boldsymbol{\zeta}_{ghi} = \boldsymbol{A}^\top \boldsymbol{s}_{ghi}, \quad \text{satisfying} \quad \boldsymbol{Z}_{ghi}^\top u_{ghi} = \boldsymbol{Q}^\top \boldsymbol{s}_{ghi} = \boldsymbol{\zeta}_{ghi}\Big(1 + o_P(1)\Big), \tag{17}$$

denote the conditional score for observation i within fine cluster h in coarse cluster g; see (10). Let $\hat{\boldsymbol{\zeta}}_{ghi} = \boldsymbol{Z}_{ghi}^\top \hat{u}_{ghi}$ denote the corresponding empirical conditional score. In what follows, we will generally omit the qualification "conditional" since this is implied by the notation. Similarly, let $\boldsymbol{\zeta}_{gh} = \sum_{i=1}^{N_{gh}} \boldsymbol{\zeta}_{ghi}$ and $\hat{\boldsymbol{\zeta}}_{gh} = \sum_{i=1}^{N_{gh}} \hat{\boldsymbol{\zeta}}_{ghi}$ denote the score and empirical score, respectively, for fine cluster h in coarse cluster g, and define the score and empirical

score for coarse cluster g as $\boldsymbol{\zeta}_g = \boldsymbol{Z}_g^\top \boldsymbol{u}_g = \sum_{h=1}^{M_g} \boldsymbol{\zeta}_{gh}$ and $\hat{\boldsymbol{\zeta}}_g = \sum_{h=1}^{M_g} \hat{\boldsymbol{\zeta}}_{gh}$, respectively. Under coarse clustering, the estimated $\boldsymbol{\Sigma}_\mathrm{c}$ matrix corresponding to (6) is

$$\hat{\boldsymbol{\Sigma}}_\mathrm{c} = m_\mathrm{c} \sum_{g=1}^{G} \boldsymbol{Z}_g^\top \hat{\boldsymbol{u}}_g \hat{\boldsymbol{u}}_g^\top \boldsymbol{Z}_g = m_\mathrm{c} \sum_{g=1}^{G} \hat{\boldsymbol{\zeta}}_g \hat{\boldsymbol{\zeta}}_g^\top = m_\mathrm{c} \sum_{g=1}^{G} \left(\sum_{h=1}^{M_g} \hat{\boldsymbol{\zeta}}_{gh} \right) \left(\sum_{h=1}^{M_g} \hat{\boldsymbol{\zeta}}_{gh} \right)^\top. \tag{18}$$

Similarly, we can write, c.f. (14) and (15),

$$\hat{\boldsymbol{\Sigma}}_\mathrm{f} = m_\mathrm{f} \sum_{g=1}^{G} \sum_{h=1}^{M_g} \left(\sum_{i=1}^{N_{gh}} \hat{\boldsymbol{\zeta}}_{ghi} \right) \left(\sum_{i=1}^{N_{gh}} \hat{\boldsymbol{\zeta}}_{ghi} \right)^\top = m_\mathrm{f} \sum_{g=1}^{G} \sum_{h=1}^{M_g} \hat{\boldsymbol{\zeta}}_{gh} \hat{\boldsymbol{\zeta}}_{gh}^\top, \tag{19}$$

where $m_\mathrm{f} = G_\mathrm{f}/(G_\mathrm{f} - 1) \times (N-1)/(N-K)$.

When interest focuses on just one coefficient, so that $k = 1$, the matrix $\boldsymbol{Z}$ becomes the vector $\boldsymbol{z}$, and the empirical scores are scalars. Specifically, $\hat{\zeta}_{ghi} = z_{ghi} \hat{u}_{ghi}$ and $\hat{\zeta}_{gh} = \sum_{i=1}^{N_{gh}} \hat{\zeta}_{ghi}$ denote the empirical scores for observation i and fine cluster h, respectively. Then the matrices (18) and (19) reduce to the scalars

$$\hat{\sigma}_\mathrm{c}^2 = m_\mathrm{c} \sum_{g=1}^{G} \left(\sum_{h=1}^{M_g} \hat{\zeta}_{gh} \right)^2 \quad \text{and} \quad \hat{\sigma}_\mathrm{f}^2 = m_\mathrm{f} \sum_{g=1}^{G} \sum_{h=1}^{M_g} \hat{\zeta}_{gh}^2. \tag{20}$$

The quantities given in (18), (19), and (20) are all defined in essentially the same way. They simply amount to different choices of empirical scores. When $N_{gh} = 1$, then $\hat{\sigma}_\mathrm{f}^2$ simplifies to

$$\hat{\sigma}_\mathrm{het}^2 = \sum_{g=1}^{G} \sum_{h=1}^{M_g} \sum_{i=1}^{N_{gh}} \hat{\zeta}_{ghi}^2, \tag{21}$$

which is just the sum of the squared empirical scores over all the observations.

Our first test is based on the difference between the two scalars in (20), namely,

$$\hat{\theta} = \hat{\sigma}_\mathrm{c}^2 - \hat{\sigma}_\mathrm{f}^2. \tag{22}$$

The empirical scores $\hat{\zeta}_{gh}$ that appear in (20), and implicitly in (22), depend on the vector $\boldsymbol{z}$, which is the residual vector from regressing $\boldsymbol{x}_1$ on $\boldsymbol{X}_2$. Different choices for $\boldsymbol{x}_1$ will yield different empirical scores, and hence different test statistics.

Our second test is based on the difference between the $k \times k$ matrices $\hat{\boldsymbol{\Sigma}}_\mathrm{c}$ and $\hat{\boldsymbol{\Sigma}}_\mathrm{f}$. For this test, we consider the vector of contrasts,

$$\hat{\boldsymbol{\theta}} = \mathrm{vech}(\hat{\boldsymbol{\Sigma}}_\mathrm{c} - \hat{\boldsymbol{\Sigma}}_\mathrm{f}), \tag{23}$$

where the operator $\mathrm{vech}(\cdot)$ returns a vector, of dimension $k(k+1)/2$ in this case, with all

the supra-diagonal elements of the symmetric $k \times k$ matrix argument removed.

In order to obtain test statistics with asymptotic distributions that are free of nuisance parameters, we need to derive the asymptotic means and variances of $\hat{\theta}$ and $\hat{\boldsymbol{\theta}}$, so that we can studentize the statistics in (22) and (23). To this end, suppose that we observe the (scalar) scores in (17), which for fine cluster h in coarse cluster g are denoted $\zeta_{gh} = \sum_{i=1}^{N_{gh}} \zeta_{ghi}$. Then the analog of $\hat{\theta}$ is the contrast

$$\theta = \sum_{g=1}^{G} \sum_{h_1=1}^{M_g} \sum_{h_2 \neq h_1}^{M_g} \zeta_{gh_1} \zeta_{gh_2}. \tag{24}$$

This is simply the sum of all the cross-products of scores that are in the same coarse cluster but different fine clusters. Under the null hypothesis, clearly, θ should have mean zero.

Under the null hypothesis, the variance of θ in (24) is

$$\mathrm{E}(\theta^2) = \sum_{g=1}^{G} \sum_{h_1=1}^{M_g} \sum_{\ell_1=1}^{M_g} \sum_{h_2 \neq h_1}^{M_g} \sum_{\ell_2 \neq \ell_1}^{M_g} \mathrm{E}(\zeta_{gh_1} \zeta_{g\ell_1} \zeta_{gh_2} \zeta_{g\ell_2}). \tag{25}$$

The expectation of any product of scores can only be nonzero, under the null, when their indices are the same in pairs. This implies that either $h_1 = \ell_1 \neq h_2 = \ell_2$ or $h_1 = \ell_2 \neq h_2 = \ell_1$. These cases are symmetric, and hence (25) is

$$2 \sum_{g=1}^{G} \sum_{h_1=1}^{M_g} \sum_{h_2 \neq h_1}^{M_g} \sigma_{gh_1}^2 \sigma_{gh_2}^2, \tag{26}$$

where $\sigma_{gh}^2 = \mathrm{Var}(\zeta_{gh})$ is used to denote $\boldsymbol{\Sigma}_{gh}$ in the scalar case; see (13), (15), and (17).

The sample analog of (26) is $2 \sum_{g=1}^{G} \sum_{h_1=1}^{M_g} \sum_{h_2 \neq h_1}^{M_g} \hat{\zeta}_{gh_1}^2 \hat{\zeta}_{gh_2}^2$; see (15) and (17). This suggests the variance estimator

$$\widehat{\mathrm{Var}}(\hat{\theta}) = 2 \sum_{g=1}^{G} \left(\sum_{h=1}^{M_g} \hat{\zeta}_{gh}^2 \right)^2 - 2 \sum_{g=1}^{G} \sum_{h=1}^{M_g} \hat{\zeta}_{gh}^4. \tag{27}$$

This equation avoids the triple summation in (26) by squaring the sums of squared empirical scores, which then requires that the second term be subtracted.[1] Combining (22) and (27) yields the studentized test statistic

$$\tau_\sigma = \frac{\hat{\theta}}{\sqrt{\widehat{\mathrm{Var}}(\hat{\theta})}}. \tag{28}$$

[1] In deriving (27), we have ignored the factors m_c and m_f, which are asymptotically irrelevant. Retaining them would have led to a much more computationally burdensome expression.

In Section 4, we show that τ_σ is asymptotically distributed as $\mathrm{N}(0,1)$.

Remark 5. The statistic defined in (28) yields either a one-sided or a two-sided test. Right-tail tests may often be of primary interest, because we expect the diagonal elements of $\boldsymbol{\Sigma}_\mathrm{c}$ to exceed the corresponding elements of $\boldsymbol{\Sigma}_\mathrm{f}$ when there is positive correlation within clusters under the alternative. However, since this is not necessarily the case, two-sided tests based on τ_σ^2 may also be of interest. The asymptotic theory in Section 4 handles both cases. □

Remark 6. Consider again the special case in which the null is heteroskedasticity with no clustering. When the elements of $\boldsymbol{z}$ display little intra-cluster correlation, the contrast $\hat{\theta}$, and hence the absolute value of τ_σ, will tend to be small, even if the residuals display a great deal of intra-cluster correlation. This is what we should expect, because in that case the so-called Moulton (1986) factor, i.e., the ratio of clustered to non-clustered standard errors, will be relatively small. Of course, the opposite will be true when the elements of $\boldsymbol{z}$ display a lot of intra-cluster correlation. □

Remark 7. It might seem that we could directly test for intra-cluster correlation by basing a test on the residuals rather than the scores. This would involve replacing $\hat{\zeta}_{gh}$ in (20) by $\sum_{i=1}^{N_{gh}} \hat{u}_{ghi}$. Unfortunately, such a test will fail whenever there are cluster fixed effects. The problem is that the residuals must sum to zero over every cluster that has a fixed effect. This implies that either $\hat{\sigma}_\mathrm{c}^2 = 0$, when there are fixed effects at the coarse level, or both $\hat{\sigma}_\mathrm{c}^2$ and $\hat{\sigma}_\mathrm{f}^2$ equal 0, when there are fixed effects at the fine level. Since models that potentially involve clustered disturbances very often include cluster-level fixed effects, it does not seem interesting to investigate this sort of test. □

When $k > 1$, so that $\hat{\boldsymbol{\theta}}$ is a vector, the variance estimator analogous to (27) is

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}) = 2\sum_{g=1}^{G} \boldsymbol{H}_k \left(\sum_{h=1}^{M_g} \hat{\boldsymbol{\zeta}}_{gh}\hat{\boldsymbol{\zeta}}_{gh}^\top \otimes \sum_{h=1}^{M_g} \hat{\boldsymbol{\zeta}}_{gh}\hat{\boldsymbol{\zeta}}_{gh}^\top \right) \boldsymbol{H}_k^\top - 2\sum_{g=1}^{G}\sum_{h=1}^{M_g} \boldsymbol{H}_k \left(\hat{\boldsymbol{\zeta}}_{gh}\hat{\boldsymbol{\zeta}}_{gh}^\top \otimes \hat{\boldsymbol{\zeta}}_{gh}\hat{\boldsymbol{\zeta}}_{gh}^\top \right) \boldsymbol{H}_k^\top. \quad (29)$$

Here $\boldsymbol{H}_k$ is the so-called elimination matrix satisfying $\mathrm{vech}(\boldsymbol{S}) = \boldsymbol{H}_k \mathrm{vec}(\boldsymbol{S})$ for any $k \times k$ symmetric matrix $\boldsymbol{S}$ (Harville 1997, p. 354), and $\otimes$ denotes the Kronecker product. A studentized (Wald) statistic is then given by

$$\tau_\Sigma = \hat{\boldsymbol{\theta}}^\top \widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})^{-1} \hat{\boldsymbol{\theta}}. \quad (30)$$

In Section 4, we show that τ_Σ is asymptotically distributed as $\chi^2(k(k+1)/2)$.

In this section, we have proposed two score-variance tests of (16). They both involve comparing different variance estimates of the empirical scores, namely, the two scalars in (20) for the τ_σ test and the matrices in (18) and (19) for the τ_Σ test. The former is a special

case of the latter, where all regressors except one have been partialed out. This special case is interesting, because the τ_σ test can be directional and because many equations simplify neatly in the scalar case.

As we show in Section 5, the finite-sample properties of our asymptotic tests are often good but could sometimes be better, especially when the number of clusters under the alternative is quite small. In such cases, we therefore recommend the use of bootstrap tests based on the statistics (28) and (30), which often perform much better in finite samples; see Section 5. These bootstrap implementations are described next.

3.2 Bootstrap Implementation

The simplest way to implement a bootstrap test based on any of our test statistics is to compute a bootstrap P value, say $\hat{P}^*$, and reject the null hypothesis when it is less than the level of the test. The bootstrap methods that we propose are based on either the ordinary wild bootstrap (Wu 1986; Liu 1988) or the wild cluster bootstrap (Cameron, Gelbach, and Miller 2008). These bootstrap methods are normally used to test hypotheses about $\boldsymbol{\beta}$, and we are not aware of any previous work in which they have been used to test hypotheses about the variances of parameter estimates. The asymptotic validity of the bootstrap tests that we now describe is established in Section 4.2.

The key idea of the wild bootstrap is to obtain the bootstrap disturbances by multiplying the residuals by realizations of an auxiliary random variable with mean 0 and variance 1. In contrast to many applications of the wild bootstrap, the residuals in this case are unrestricted, meaning that they do not impose a null hypothesis on $\boldsymbol{\beta}$. This is because we are not testing any restrictions on $\boldsymbol{\beta}$ when testing the level of clustering. In the special case of testing the null of heteroskedasticity, as in Remark 4, we use the ordinary wild bootstrap. When the null involves clustering, we use the wild cluster bootstrap. Because the test statistics depend only on residuals, the value of $\boldsymbol{\beta}$ in the bootstrap DGP does not matter, and so we set it to zero.

The b^{th} wild (cluster) bootstrap sample is thus generated by $\boldsymbol{y}^{*b} = \boldsymbol{u}^{*b}$, where the vector of bootstrap disturbances $\boldsymbol{u}^{*b}$ has typical element given by either $u^{*b}_{ghi} = v^{*b}_{ghi}\hat{u}_{ghi}$ for the wild bootstrap or $u^{*b}_{ghi} = v^{*b}_{gh}\hat{u}_{ghi}$ for the wild cluster bootstrap. The auxiliary random variables v^{*b}_{ghi} and v^{*b}_{gh} are assumed to follow the Rademacher distribution, which takes the values $+1$ and -1 with equal probabilities. Notice that there is one such random variable per observation for the wild bootstrap and one per cluster for the wild cluster bootstrap. Other distributions can also be used; see Davidson and Flachaire (2008), Webb (2014), and Djogbenou, MacKinnon, and Nielsen (2019).

The algorithm for a wild (cluster) bootstrap-based implementation of our tests is as

follows. It applies to both τ_σ and τ_Σ. For simplicity, the algorithm below simply refers to one test statistic, τ. However, it is easy to perform two or more tests at the same time, using just one set of bootstrap samples for all of them. For example, if there are three possible regressors of interest, we might perform four tests, one with $k = 3$ based on τ_Σ and three with $k = 1$ based on different versions of τ_σ.

Algorithm 1 (Bootstrap test implementation)**.** Let $B >> 1$ denote the number of bootstrap replications, and let τ denote the chosen test statistic.

1. Estimate the model (1), or equivalently the model (8), by OLS regression to obtain the residuals $\hat{\boldsymbol{u}}$.
2. Compute the empirical score vector $\hat{\boldsymbol{\zeta}}$ and use it to compute τ.
3. For $b = 1, \ldots, B$,
   (a) generate the vector of bootstrap dependent variables $\boldsymbol{y}^{*b} = \boldsymbol{u}^{*b}$ from the residual vector $\hat{\boldsymbol{u}}$ using the wild cluster bootstrap corresponding to the null hypothesis, or the ordinary wild bootstrap if the null does not involve clustering.
   (b) Regress $\boldsymbol{y}^{*b}$ on $\boldsymbol{X}$ to obtain the bootstrap residuals $\hat{\boldsymbol{u}}^{*b}$, and use these, together with $\boldsymbol{z}$ or $\boldsymbol{Z}$, to compute τ^{*b}, the bootstrap analog of τ.
4. Compute the bootstrap P value $\hat{P}^* = B^{-1} \sum_{b=1}^{B} \mathbb{I}(\tau^{*b} > \tau)$.

As usual, if α is the level of the test, then B should be chosen so that $(1-\alpha)B$ is an integer (Racine and MacKinnon 2007). Numbers like 999 and 9,999 are commonly used because they satisfy this condition for conventional values of α. The power of the test increases in B, but it does so very slowly once B exceeds a few hundred (Davidson and MacKinnon 2000).[2]

Remark 8. When τ is defined as τ_σ, Algorithm 1 yields a one-sided right-tail test. When τ is defined as $|\tau_\sigma|$, τ_σ^2, or τ_Σ, it yields a two-sided test; see Remark 5. □

Remark 9. We could use the ordinary wild bootstrap instead of the wild cluster bootstrap in Algorithm 1, even when the null hypothesis involves clustering. The same intuition as in Djogbenou et al. (2019) applies, whereby the ordinary wild bootstrap would lead to asymptotically valid tests because the statistics are asymptotically pivotal. There may be cases, like the ones considered in MacKinnon and Webb (2018) and/or ones in which the number of fine clusters is small, in which the wild bootstrap would perform better than the wild cluster bootstrap. However, we believe that such cases are likely to be rare. □

[2] If desired, bootstrap critical values can be calculated as quantiles of the τ^{*b}. For example, when $B = 999$ and the τ^{*b} are sorted from smallest to largest, the 0.05 critical value for a one-sided upper-tail test is number $(1 - 0.05)(B + 1) = 950$ in the sorted list.

3.3 Choosing the Level of Clustering by Sequential Testing

In many applications, there are several possible levels of clustering. In such situations, we suggest a sequential testing procedure. The statistical principle upon which we base our testing procedure is the intersection-union (IU) principle (e.g., Berger and Sinclair 1984), whereby a hypothesis is rejected if and only if the hypothesis itself, along with any hypotheses nested within it, are all rejected. The IU principle leads naturally to a bottom-up testing strategy for the level of clustering, whereby a model is chosen if and only if the model itself is not rejected, but all models nested within it are rejected. Berger and Sinclair (1984) shows that the IU principle does not imply an inflation of the family-wise rejection rate in the context of multiple testing; that is, there is no accumulation of size due to testing multiple hypotheses. We prove a similar result for our sequential procedure below.

Suppose the potential levels of clustering are sequentially nested, and denote their corresponding $\mathbf{\Sigma}$ matrices by $\mathbf{\Sigma}_0, \mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_p$; see (5) and (14). Here we assume that $\mathbf{\Sigma}_0$ corresponds to no clustering, c.f. Remarks 1 and 4, and that, in addition, there are p potential levels of clustering of the data. All these levels of clustering are assumed to be nested from fine to increasingly more coarse clustering.

In this situation, following the IU statistical principle mentioned above, we reject clustering at level m if and only if levels $0, \ldots, m$ are all rejected. That is, the natural testing strategy here is to test clustering at level m against the coarser level $m+1$ sequentially, for $m = 0, 1, \ldots, p-1$, and choose the level of clustering in the first non-rejected test. Algorithmically, we perform the following sequential testing procedure.

Algorithm 2 (Nested sequential testing procedure)**.** Let $m = 0$. Then:

1. Test $\mathrm{H}_0 \colon \lim_{N\to\infty} \mathbf{\Sigma}_m \mathbf{\Sigma}_{m+1}^{-1} = \mathbf{I}$ against $\mathrm{H}_1 \colon \lim_{N\to\infty} \mathbf{\Sigma}_m \mathbf{\Sigma}_{m+1}^{-1} \neq \mathbf{I}$.
2. If the test in step 1 does not reject, choose $\hat{m} = m$ and stop.
3. If $m = p - 1$ and the test in step 1 rejects, choose $\hat{m} = p$ and stop.
4. If $m \leq p - 2$ and the test in step 1 rejects, increment m by 1 and go to step 1.

We can equivalently state the sequential testing problem in Algorithm 2 as a type of estimation problem. Specifically,

$$\hat{m} = \min\{m \in \{0, 1, \ldots, p\} \text{ such that } \mathrm{H}_0 \colon \operatorname*{plim}_{N\to\infty} \mathbf{\Sigma}_m \mathbf{\Sigma}_{m+1}^{-1} = \mathbf{I} \text{ is not rejected}\}. \tag{31}$$

Of course, the $\hat{m}$ resulting from Algorithm 2 and from (31) will be identical.

Because each individual test will reject a false null hypothesis with probability converging to one, this procedure will (at least asymptotically) never choose a level of clustering

that is too fine. In other words, $\hat{m}$ defined in either Algorithm 2 or (31) will be (nearly) consistent. Precise asymptotic properties of the proposed sequential procedure are established in Section 4.3, and finite-sample performance is investigated using Monte Carlo simulation methods in Section 5.3.

3.4 Inference about Regression Coefficients

The ultimate purpose of using score-variance tests is to make more reliable inferences about the coefficient(s) of interest, that is, $\boldsymbol{\beta}_1$ in (8). This may or may not involve some sort of formal pre-testing or model averaging procedure. We intend to provide a detailed treatment of this important issue in future work. In this subsection, we briefly discuss the key issues.

For simplicity, suppose we are attempting to construct a confidence interval for β_1, the (scalar) coefficient of interest, when there are just two levels of clustering, fine and coarse. Without a testing procedure, an investigator must choose between fine and coarse clustering on the basis of prior beliefs about which level is appropriate, perhaps informed by the observed standard errors associated with the two levels of clustering. With the testing procedures of this paper, an investigator can instead choose the level of clustering based on the outcome of a score-variance test. This involves picking a level α for the test and deciding whether to use a one-sided or a two-sided test. We form the interval based on coarse clustering when the score-variance test rejects, and we form the interval based on fine clustering when it does not reject.

Of course, this procedure can never work as well as the infeasible procedure of simply choosing the correct level of clustering. Since it involves pre-testing, it inevitably suffers from some of the classic problems associated with pre-testing (e.g., Leeb and Pötscher 2005). When there is actually fine clustering, the pre-test will sometimes make a Type I error and reject, leading to an interval that is too long. When there is actually coarse clustering, the pre-test will sometimes make a Type II error and fail to reject, leading to an interval that is too short. We report the results of some preliminary simulation experiments that compare confidence intervals based on several alternative procedures in Appendix C.5.

4 Asymptotic Theory

In Section 4.1, we derive the asymptotic distributions of the two score-variance test statistics under the null hypothesis and show that they are divergent under the alternative. Then we prove the validity of the bootstrap implementation (Section 4.2) and prove asymptotic results for the sequential testing procedure (Section 4.3). We first state and discuss the

assumptions needed for our proofs, which may be found in Appendix A.

Assumption 1. The sequence $\boldsymbol{s}_{gh} = \sum_{i=1}^{N_{gh}} \boldsymbol{X}_{ghi}^\top u_{ghi}$ is independent across both g and h. □

Assumption 2. For all g, h, it holds that $\mathrm{E}(\boldsymbol{s}_{gh}) = \boldsymbol{0}$ and $\mathrm{Var}(\boldsymbol{s}_{gh}) = \boldsymbol{V}_{gh}$. Furthermore, $\sup_{g,h,i} \mathrm{E}\|\boldsymbol{s}_{ghi}\|^{2\lambda} < \infty$ for some $\lambda > 1$. □

Assumption 3. The regressor matrix $\boldsymbol{X}$ satisfies $\sup_{g,h,i} \mathrm{E}\|\boldsymbol{X}_{ghi}\|^2 < \infty$ and $N^{-1}\boldsymbol{X}^\top\boldsymbol{X} \xrightarrow{P} \boldsymbol{\Xi}$, where $\boldsymbol{\Xi}$ is finite and positive definite. □

Assumption 4. Let $\omega_{\min}(\cdot)$ and $\omega_{\max}(\cdot)$ denote the minimum resp. maximum eigenvalue of the argument. Then $\inf_{g,h} N_{gh}^{-1}\omega_{\min}(\boldsymbol{\Sigma}_{gh}) > 0$ and $\sup_{g,h} \omega_{\max}\Big(\boldsymbol{\Sigma}_{gh}(\sum_{h=1}^{M_g} \boldsymbol{\Sigma}_{gh})^{-1}\Big) < 1$. □

Assumption 5. For λ defined in Assumption 2, the cluster sizes satisfy

$$\frac{\sup_g N_g^2 \sup_{g,h} N_{gh}^2}{\sum_{g=1}^G \omega_{\min}\Big(\sum_{h=1}^{M_g} \boldsymbol{\Sigma}_{gh}\Big)^2} \longrightarrow 0 \quad \text{and} \quad \frac{N^{1/\lambda} \sup_g N_g \sup_{g,h} N_{gh}^{3-1/\lambda}}{\sum_{g=1}^G \omega_{\min}\Big(\sum_{h=1}^{M_g} \boldsymbol{\Sigma}_{gh}\Big)^2} \longrightarrow 0. \qquad \square$$

Assumption 1 is the assumption of (at most) "fine" clustering, which implies that the null hypothesis in (16) is satisfied, even without taking the limit. In fact, it is slightly weaker than that, because we do not make the stronger assumption that all observations in any fine cluster are independent of those in a different fine cluster; we only assume that the cluster sums are independent across fine clusters. The moment conditions in Assumption 2 and the multicollinearity condition in Assumption 3 are standard in linear regression models.

Next, the conditions in Assumption 4 rule out degenerate cases. The minimum eigenvalue condition rules out perfect negative correlation between scores within fine clusters. The maximum eigenvalue condition ensures that the variance of a single fine cluster cannot dominate the sum of the variances within a coarse cluster. It is basically satisfied if $M_g > 1$ for all g. The latter holds by construction of the test statistics, because any coarse cluster with $M_g = 1$ will not contribute to $\hat{\boldsymbol{\theta}}$, and hence not to the test statistic.

The conditions in Assumption 5 restrict the amount of heterogeneity of cluster sizes that is allowed under both the null and the alternative. Neither the fine cluster sizes nor the coarse cluster sizes are required to be bounded under these conditions, which allow the cluster sizes to diverge with the sample size. The first condition is used in the proofs to replace residuals with disturbances and for convergence of the variance. The second condition trades off moments and cluster size heterogeneity to rule out the possibility that one cluster dominates the test statistic in the limit in such a way that the central limit theorem does not apply; technically, it is used to verify Lyapunov's condition for the central limit theorem. When $\lambda \to \infty$, the second condition is implied by the first.

Under Assumptions 1 and 4, the denominators of both conditions in Assumption 5 are bounded from below by $\sum_{g=1}^{G} N_g^2 \geq cN \inf_g N_g$, and a sufficient condition for Assumption 5 is

$$\sup_{g,h} N_{gh}^2 \left(\frac{\sup_g N_g}{\inf_g N_g}\right)\left(\frac{\sup_g N_g}{N}\right) \longrightarrow 0 \quad \text{and} \quad \left(\frac{\sup_g N_g}{\inf_g N_g}\right)^{\lambda}\left(\frac{\sup_{g,h} N_{gh}^{3\lambda-1}}{N^{\lambda-1}}\right) \longrightarrow 0. \tag{32}$$

If the cluster sizes are bounded under the alternative, i.e. $\sup_g N_g < \infty$, then (32) is easily satisfied. Note that $\sup_g N_g/N \to 0$, and hence $G \to \infty$, is implied by Assumption 5, and it is therefore not stated explicitly. Moreover, Assumption 5 allows the possibility that $\sup_{g,h} N_{gh} = \sup_g N_g$, in which case the conditions simplify accordingly.

Remark 10. Consider again the important special case in which the scores are independent, but heteroskedastic under the null (or more generally that cluster sizes are bounded under the null, i.e. $\sup_{g,h} N_{gh} < \infty$). We consider two examples. First, let $N_g = N^{\alpha}$ for $g = 1, \ldots, G$ and $G = N^{1-\alpha}$. We can interpret α small as many small clusters and α large as few large clusters (under the alternative). In this relatively homogeneous case, (32), and hence Assumption 5, is satisfied for any $\alpha < 1$. Second, let $N_g = N^{\alpha}$ for $g = 1, \ldots, G_1$ with G_1 fixed, and suppose N_g is bounded for $g = G_1, \ldots, G$. We interpret this as a small (fixed) number of large clusters and many small clusters under the alternative. In this very heterogeneous case, the two conditions of (32) are satisfied when $\alpha < 1/2$ and $\alpha < 1 - 1/\lambda$, respectively. □

The denominators of both terms in Assumption 5 show that these conditions trade off intra-cluster dependence and cluster-size heterogeneity. That is, the sufficient condition in (32) can be relaxed somewhat when there is more correlation within fine clusters. Intuitively, the greater the amount of intra-cluster correlation, the less information large clusters provide relative to small clusters, which allows the large ones to be relatively larger without dominating the limit; a similar tradeoff was found in Djogbenou et al. (2019). Technically, the reason in our case is that the lower bound for the denominators in Assumption 5 can be made larger, for a given fine clustering structure, by assuming more correlation within fine clusters. We illustrate this tradeoff in the following remark.

Remark 11. Assumption 4 could be strengthened to assume that $\inf_{g,h} N_{gh}^{-2} \omega_{\min}(\boldsymbol{\Sigma}_{gh}) > 0$. Then the denominators in Assumption 5 would be bounded from below by $N \inf_g N_g \inf_{g,h} N_{gh}^2$, so that a sufficient condition for Assumption 5 would be

$$\left(\frac{\sup_{g,h} N_{gh}}{\inf_{g,h} N_{gh}}\right)^2 \left(\frac{\sup_g N_g}{\inf_g N_g}\right)\left(\frac{\sup_g N_g}{N} + \frac{\sup_{g,h} N_{gh}}{N^{1-1/\lambda}}\right) \longrightarrow 0. \tag{33}$$

For example, if a random effects model is assumed under the null, i.e., $\boldsymbol{s}_{ghi} = \boldsymbol{\eta}_{gh} + \boldsymbol{\varepsilon}_{ghi}$, where $\boldsymbol{\eta}_{gh}$ and $\boldsymbol{\varepsilon}_{ghi}$ are independent across all subscripts with 2λ finite moments, then (33) would

be sufficient for Assumption 5. The same would be true for many factor-type models. □

When $\sup_{g,h} N_{gh} < \infty$, as in Remark 10, the conditions (32) and (33) are identical. Thus, the distinction between (32) and (33) is only relevant when $\sup_{g,h} N_{gh}$ is unbounded.

Remark 12. We note from (15) that another sufficient condition for Assumption 5 could be obtained by replacing $\boldsymbol{\Sigma}_{gh}$ with $\boldsymbol{V}_{gh}$, because $\omega_{\min}(\boldsymbol{V}_{gh}) \leq \omega_{\min}(\boldsymbol{\Sigma}_{gh})$. While this would be attractive in the sense that $\boldsymbol{V}_{gh}$ is defined directly in terms of the scores $\boldsymbol{s}_{gh}$, it would result in a much stronger assumption. Suppose, for example, that $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are (asymptotically) orthogonal, such that $\boldsymbol{\Sigma}$ equals the diagonal block of $\boldsymbol{V}$ corresponding to $\boldsymbol{X}_1^\top \boldsymbol{u}$. Suppose also that $\boldsymbol{X}_1$ and $\boldsymbol{u}$ are both finely clustered, but $\boldsymbol{X}_2$ is independent. Then $\boldsymbol{\Sigma}_{gh}$ satisfies the condition in Remark 11, while $\boldsymbol{V}_{gh}$ only satisfies the corresponding condition in Assumption 4, and hence using $\boldsymbol{V}_{gh}$ in Assumption 5 would lead to a stronger condition. □

Remark 13. Consider the following prototypical setup for clusters that are relatively homogeneous, but possibly unbounded, in size. Suppose the coarse clusters have size $N_g = N^\alpha$ for $g = 1, \ldots, G = N^{1-\alpha}$, and for each g, the fine clusters have size $N_{gh} = N_g^\gamma$ for $h = 1, \ldots, M_g = N_g^{1-\gamma}$. That is, when α (γ) is large, there are few but large coarse (fine) clusters. Conversely, when α (γ) is small, there are many small coarse (fine) clusters. In this case, (32) is satisfied if $\alpha(2\gamma + 1) < 1$ and $\alpha\gamma < (\lambda - 1)/(3\lambda - 1)$. On the other hand, (33) is much weaker and is satisfied if $\alpha < 1$ and $\alpha\gamma < 1 - 1/\lambda$. □

Remark 14. The sufficient condition in (32) applies because the observations could possibly be generated independently, regardless of the null and alternative hypotheses. On the other hand, as explained in Remark 11, the weaker sufficient condition in (33) is obtained when there is more substantial clustering in the data-generating process. This is relevant for the sequential testing procedure. At any given step in that procedure, the previous step concluded that clustering is at least at the level of the null hypothesis of the current step, which implies that the condition in (33) and Remark 11 applies. □

4.1 Theory for Asymptotic Tests

Theorem 1. *Let Assumptions 1–5 be satisfied. Then, as $N \to \infty$, it holds that*

$$\mathrm{Var}(\boldsymbol{\theta})^{-1/2}\hat{\boldsymbol{\theta}} \xrightarrow{d} \mathrm{N}(0, \mathbf{I}), \qquad \mathrm{Var}(\boldsymbol{\theta})^{-1}\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}) \xrightarrow{P} \mathbf{I}, \quad \textit{and}$$

$$\frac{\hat{\theta}}{\sqrt{\mathrm{Var}(\theta)}} \xrightarrow{d} \mathrm{N}(0, 1), \qquad \frac{\widehat{\mathrm{Var}}(\hat{\theta})}{\mathrm{Var}(\theta)} \xrightarrow{P} 1.$$

Remark 15. Observe that the statement of the asymptotic distributions in Theorem 1 only concerns quantities that are self-normalized. For example, in the scalar case, these are either $\hat{\theta}$ divided by its true standard error or the estimated variance of $\hat{\theta}$ divided by the true variance. This is because the appropriate rates of convergence are not known in general; see the discussion below (16). □

The asymptotic distributions of the test statistics follow immediately from Theorem 1:

Corollary 1. *Let Assumptions 1–5 be satisfied. Then, as $N \to \infty$, it holds that*

$$\tau_\Sigma \xrightarrow{d} \chi^2\big(k(k+1)/2\big) \quad \textit{and} \quad \tau_\sigma \xrightarrow{d} \mathrm{N}(0,1).$$

We next consider the asymptotic behavior of the test statistics under the alternative. Because Assumption 1 implies that H_0 is true, that assumption is not made. Instead, we impose the following conditions:

Assumption 6. The sequence $\boldsymbol{s}_g = \boldsymbol{X}_g^\top \boldsymbol{u}_g = \sum_{h=1}^{N_g} \boldsymbol{s}_{gh}$ is independent across g. □

Assumption 7. The cluster sizes satisfy

$$\frac{\sup_g N_g^{3/2} N^{1/2}}{\sum_{g=1}^G \omega_{\min}(\boldsymbol{\Sigma}_g)} \longrightarrow 0.$$ □

Assumption 6 is the assumption of (at most) coarse clustering. This assumption is very general, and departures from the null could be very small and inconsequential. In order for our tests to be able to detect departures from the null hypothesis, with probability converging to one in the limit, we need to impose sufficient correlation within the coarse clusters. That is, we need $\boldsymbol{\Sigma}_g = \sum_{h_1=1}^{M_g} \sum_{h_2=1}^{M_g} \mathrm{E}(\boldsymbol{\zeta}_{gh_1} \boldsymbol{\zeta}_{gh_2}^\top)$ to be sufficiently large, in aggregate. This condition is embodied in Assumption 7.

Remark 16. As in Remark 11, there is a tradeoff between cluster size heterogeneity and correlation, in this case correlation within coarse clusters. Specifically, under Assumption 4, the denominator in Assumption 7 is bounded from below by $\sum_{g=1}^G N_g = N$, and hence a sufficient condition for Assumption 7 is

$$\frac{\sup_g N_g^3}{N} \longrightarrow 0. \tag{34}$$

Suppose instead that Assumption 4 were strengthened to assume that $\inf_g N_g^{-2} \omega_{\min}(\boldsymbol{\Sigma}_g) > 0$ (as in Remark 11, this could be due to a random effects model or a factor-type model). That is, more correlation is assumed within the coarse clusters, so that there is a stronger departure

from the null hypothesis. In this case, the denominator in Assumption 7 is bounded from below by $\sum_{g=1}^{G} N_g^2 \geq \inf_g N_g N$. Therefore, a sufficient condition for Assumption 7 is

$$\frac{\sup_g N_g^3}{\inf_g N_g^2 N} \longrightarrow 0. \tag{35}$$

With relatively homogeneous coarse clusters as in Remark 13, i.e. coarse clusters where $\sup_g N_g$ and $\inf_g N_g$ are of the same order of magnitude, the condition (35) reduces to $\sup_g N_g/N \to 0$, which is clearly minimal and implied by Assumption 5. □

Theorem 2. *Let Assumptions 2–7 be satisfied, and suppose H_0 in (16) is not true. Then, as $N \to \infty$, it holds that*

$$\tau_\Sigma \xrightarrow{P} +\infty \quad \text{and} \quad |\tau_\sigma| \xrightarrow{P} +\infty.$$

It follows immediately from Theorem 2 that tests based on either of our statistics reject with probability converging to one under the alternative. That is, they are consistent tests.

4.2 Theory for Bootstrap Tests

We now demonstrate the asymptotic validity of the bootstrap implementation of our tests. To this end, let τ denote either of our statistics, and let the cumulative distribution function of τ under H_0 be denoted $P_0(\tau \leq x)$. The corresponding bootstrap statistic is denoted τ^*. As usual, let P^* denote the bootstrap probability measure, conditional on a given sample, and let E^* denote the corresponding expectation conditional on a given sample.

Theorem 3. *Let Assumptions 2–6 be satisfied with $\lambda \geq 2$, and assume that $\mathrm{E}^*|v^*|^{2\lambda} < \infty$. Then, as $N \to \infty$, it holds for any $\epsilon > 0$ that*

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

First, note that the bootstrap theory requires a slight strengthening of the moment condition since at least four moments are now required. Second, Theorem 3 shows that the bootstrap P values in Algorithm 1 are asymptotically valid under Assumption 1 and H_0. Third, note that neither the null hypothesis nor Assumption 1 is imposed in Theorem 3. Thus Theorems 1–3 together show immediately that the bootstrap tests are consistent. We summarize these results in the following corollary.

Corollary 2. *Let Assumptions 2–5 be satisfied with $\lambda \geq 2$, and assume that $\mathrm{E}^*|v^*|^{2\lambda} < \infty$. As $N \to \infty$, it holds that:*

(i) If Assumption 1 is satisfied and H_0 is true, then $\hat{P}^* \xrightarrow{d} \mathrm{U}(0,1)$, where $\mathrm{U}(0,1)$ is a uniform random variable on $[0,1]$.

(ii) If Assumptions 6 and 7 are satisfied and H_0 is not true, then $\hat{P}^* \xrightarrow{P} 0$.

4.3 Theory for Sequential Testing Procedure

We next provide a theoretical justification for the sequential testing procedure in Algorithm 2.

Theorem 4. *Suppose Assumption 1 is satisfied when the "fine" clustering level in (16) is $m = m_0 \in \{0, 1, \ldots, p\}$ (and hence also for $m > m_0$), and suppose H_0 in (16) is not true for cluster levels $m < m_0$. Suppose also that Assumptions 2–5 and 7 are satisfied, and let α denote the nominal level of the tests. As $N \to \infty$, it holds that*

(i) if $m_0 \leq p-1$ then $P(\hat{m} \leq m_0 - 1) \to 0$, $P(\hat{m} = m_0) \to 1-\alpha$, and $P(\hat{m} \geq m_0 + 1) \to \alpha$,

(ii) if $m_0 = p$ then $P(\hat{m} \leq m_0 - 1) \to 0$ and $P(\hat{m} = m_0) \to 1$.

The results in Theorem 4 show that $\hat{m}$ defined in Algorithm 2 or (31) is nearly consistent, in the sense that it is asymptotically correct with probability converging to $1 - \alpha$ when $m_0 \leq p - 1$ and with probability converging to 1 when $m_0 = p$. It is worth emphasizing that the sequential procedure will never "under-estimate" the cluster level, at least asymptotically, in the sense that $\hat{m} < m_0$ with probability converging to 0.

5 Simulation Experiments

Most of the papers cited in the second paragraph of Section 1 employ simulation experiments to study the finite-sample properties of various methods for cluster-robust inference. To our knowledge, all of these papers use some sort of random effects, or single-factor, model to generate the data. The key feature of these models is that all of the intra-cluster correlation for every cluster g arises from a single random variable, say ξ_g, that affects every observation within that cluster *equally*. This yields disturbances that are equi-correlated within each cluster. Although this type of model is convenient to work with and can readily generate any desired level of intra-cluster correlation, it cannot be used when a regression model has cluster fixed effects. The fixed effects completely explain the ξ_g so that, no matter how highly correlated within each cluster the disturbances may be, the residuals are uncorrelated.

Investigators very often wish to test whether it is valid to use standard errors that are robust to heteroskedasticity but not to intra-cluster correlation in regression models that have cluster fixed effects. It is always valid to use heteroskedasticity-robust (HR) standard errors for such models when the intra-cluster correlation of either the disturbances or the regressors arises solely from a random effects model. Therefore, the null hypothesis of our tests is satisfied, and they will have no (asymptotic) power. Of course, this is the desired

outcome both in the statistical sense, because the null is satisfied, and in the practical sense, because cluster-robust (CR) standard errors are not needed.

In practice, HR and CR standard errors often differ greatly in models with cluster fixed effects; see, for example, Bertrand et al. (2004), MacKinnon (2019), and Section 6. Therefore, whatever processes are generating intra-cluster correlation in real-world data must be more complicated than simple random effects models. Since we wish to investigate models with cluster fixed effects, we need to employ a data-generating process (DGP) for which cluster fixed effects do not remove all of the intra-cluster correlation. To this end, the DGPs for both the regressors and the disturbances in our experiments employ J correlated random factors per cluster instead of just one. The idea is that the observations within each cluster all depend on one of these J unobserved factors. In practice, we set $J = 10$, but the results are qualitatively the same for other reasonable values of J.

Consider first the generation of the disturbances, $\boldsymbol{u}_g$, for $g = 1, \ldots, G$. These are generated with one level of clustering, so there is no need to distinguish fine and coarse clusters. Specifically, the disturbance vector for cluster g is generated by the factor model

$$\boldsymbol{u}_g = \boldsymbol{W}_\xi \boldsymbol{\xi}_g + w_\epsilon \boldsymbol{\epsilon}_g, \quad \boldsymbol{\epsilon}_g \sim \mathrm{N}(\boldsymbol{0}, \mathbf{I}_g), \quad g = 1, \ldots, G, \tag{36}$$

where $\boldsymbol{\epsilon}_g$ is an idiosyncratic noise term and $\boldsymbol{\xi}_g = [\xi_{g1}, \ldots, \xi_{gJ}]^\top$ is a J-vector of unobserved random factors. The $N_g \times J$ loading matrix $\boldsymbol{W}_\xi$ has $(i,j)^{\text{th}}$ entry $w_\xi \mathbb{I}\Big(j = \lfloor (i-1)J/N_g \rfloor + 1\Big)$, where $\lfloor \cdot \rfloor$ denotes the integer part of the argument. We normalize $w_\xi^2 + w_\epsilon^2 = 1$. When $J = 1$, $\boldsymbol{W}_\xi$ is a vector of ones (prior to normalization), and clearly (36) is the random effects model. In practice, we choose N_g so that it is a multiple of J, and hence precisely N_g/J observations in each cluster depend on each of the J unobserved factors. The most important parameter here is w_ξ, which is the weight on the unobserved factor. When $w_\xi = 0$, there is no intra-cluster correlation.

If the ξ_{gj} in (36) were independent, the only within-cluster correlation would arise from columns in $\boldsymbol{W}_\xi$ with multiple non-zero entries. We generate additional correlation within cluster g by letting the factors ξ_{gj} be correlated across j. Specifically, we let ξ_{gj} follow a stationary first-order autoregression with unconditional variance 1, given as

$$\xi_{g1} \sim \mathrm{N}(0,1), \quad \xi_{gj} = \rho \xi_{g,j-1} + e_{gj}, \quad e_{gj} \sim \mathrm{N}(0, 1-\rho^2), \quad j = 2, \ldots, J, \tag{37}$$

where $\rho^2 < 1$. Although the average intra-cluster correlation increases with ρ, the value of ρ should not be too large, because the ξ_{gj} for each g become more similar as ρ increases. When ρ is very close to 1, the DGP consisting of (36) and (37) becomes hard to distinguish from a random effects model, so that the fixed effects remove most of the intra-cluster correlation.

Note that (36) and (37) make no reference to fine and coarse clusters. As stated, they could be used to generate either finely or coarsely clustered data. When there is coarse clustering, we order the observations so that, when all clusters are the same size, every fine cluster contains an equal number of observations that depend on each of the factors. For example, when each coarse cluster contains 200 observations, each fine cluster contains 50 observations, and there are $J = 10$ factors, every coarse (or fine) cluster will have 20 (or 5) observations that depend on each of the factors. This ensures that, when there is coarse clustering, there will be the same level of correlation both within fine clusters, and across fine clusters within each coarse cluster.

For example, in an analysis of house prices where the fine and coarse clusters corresponded to small and large geographic areas, the factors in (36) and (37) might correspond to streets of different types. In an analysis of wages where the clusters corresponded to plants or firms, the factors might correspond to occupations. However, there is no need to interpret the DGP in this way. It is simply a way to generate scores that display intra-cluster correlation even in the presence of fixed effects.

We described the generation of the disturbances above. Except in the experiments of Appendix C.4, the regressors $\boldsymbol{X}_1$ are generated in the same way, but they are always coarsely clustered. This ensures that, if the disturbances are either independent ($w_\xi = 0$), finely clustered, or coarsely clustered, the scores will also be independent, finely clustered, or coarsely clustered, respectively. When the DGP in (36) and (37) is applied at the coarse level with $w_\xi > 0$ and $\rho > 0$, there is at least some correlation between every pair of observations within each coarse cluster, and consequently also between every pair of observations within each fine cluster. If instead (36) and (37) are applied at the fine level, then there is correlation between every pair of observations within each fine cluster, but there is no correlation (or dependence) across fine clusters. Importantly, cluster fixed effects at either the coarse or fine levels do not eliminate these correlations, although they will usually reduce their magnitude.

In all experiments, the regressors in $\boldsymbol{X}_1$ are generated independently. This implies that there is no correlation among the coefficients. It might seem that the extent of any such correlation would be important for the properties of the τ_Σ tests. However, that is not the case. We find numerically that the τ_Σ statistic is invariant to any transformation of $\boldsymbol{X}_1$ that does not change the subspace spanned by its columns. Thus there is no loss of generality in generating the regressors in $\boldsymbol{X}_1$ independently.

5.1 Performance under the Null Hypothesis

Our first set of experiments is designed to investigate the rejection frequencies of asymptotic and bootstrap score-variance tests under the null hypothesis. The model is

$$y_{ghi} = \sum_{\ell=1}^{k} \beta_\ell X_{\ell ghi} + \boldsymbol{X}_{2,gh}\boldsymbol{\delta} + u_{ghi}, \tag{38}$$

where the regressors $X_{\ell ghi}$ are generated independently across ℓ by (36) and (37) at the coarse level with $w_\xi = 0.7$, $J = 10$, and $\rho = 0.5$. The additional regressors in $\boldsymbol{X}_{2,gh}$ are either a constant term or a set of cluster fixed effects at the fine level. The number of coarse clusters, which in this section we denote by G_c, is allowed to vary. In the first set of experiments, there are always four fine clusters in each coarse cluster, so that $G_\mathrm{f} = 4G_\mathrm{c}$.

Figure 1 shows rejection frequencies at the 0.05 level for asymptotic tests for $k = 1, \ldots, 5$, which implies that the number of degrees of freedom for the tests is 1, 3, 6, 10, or 15. Panels (a) through (c) show rejection frequencies when the model includes fine-level fixed effects. Tests of no clustering versus fine clustering, in Panel (a), always over-reject less severely than the other tests. Tests of no clustering versus coarse clustering, in Panel (b), appear to over-reject considerably more severely, but this is mainly because there are not as many coarse clusters. When $G_\mathrm{c} = 16$, for example, rejection frequencies for tests of no clustering versus coarse clustering are very similar to those for tests of no clustering versus fine clustering when $G_\mathrm{f} = 16$. The modest differences arise because the coarse clusters are four times as large as the fine clusters and because the fixed effects are at the fine level.

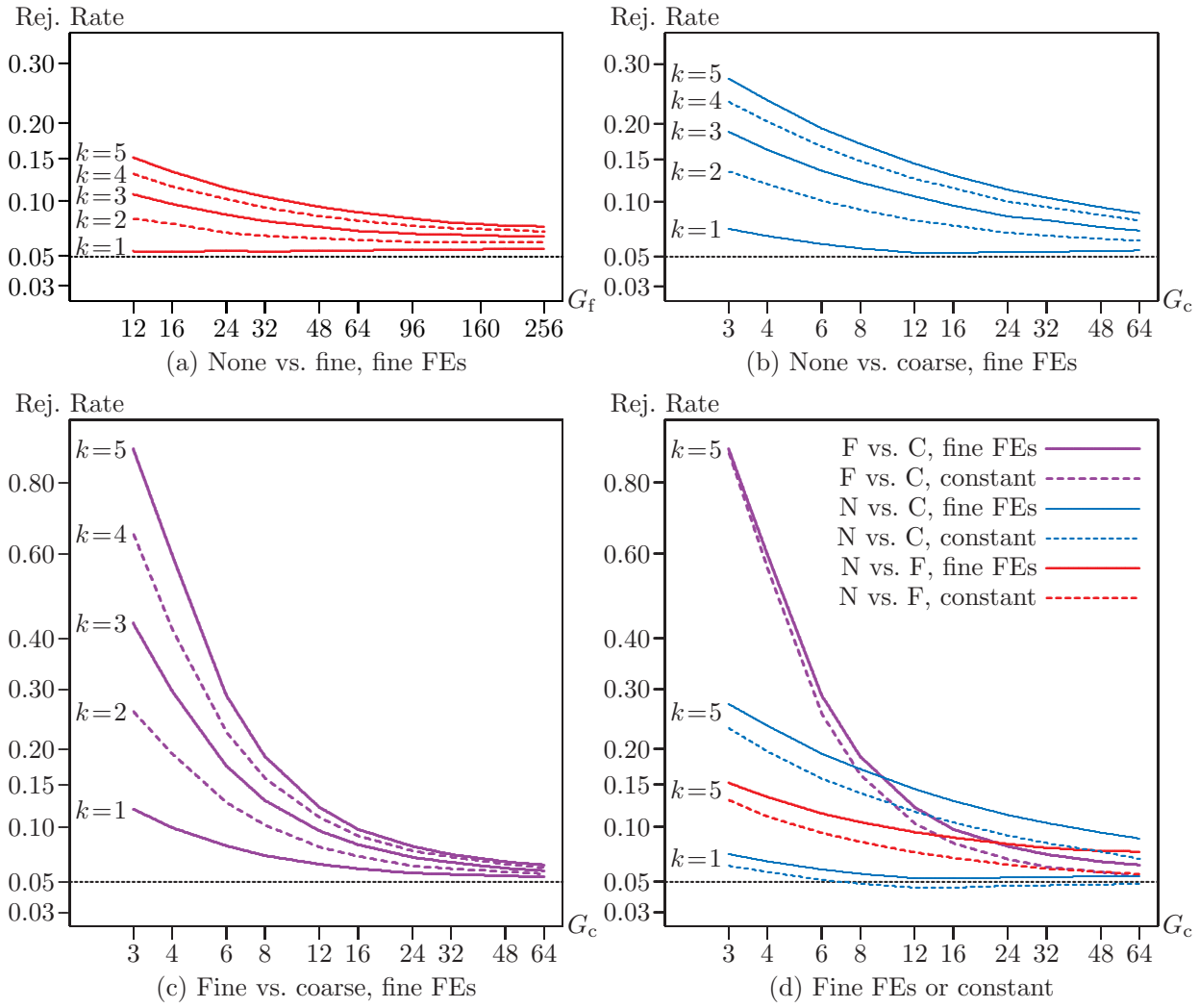
For small and moderate values of G_c, tests of fine against coarse clustering, in Panel (c), over-reject more severely than do the other tests. When k is large and G_c is small, the over-rejection can be quite extreme. However, as can most clearly be seen in Panel (d), the performance of these tests improves more rapidly with the number of clusters than does the performance of the other tests.

The most striking feature of Figure 1 is that over-rejection increases sharply with k. This should not have been a surprise in view of the fact that, like the information matrix test (White 1982), the τ_Σ test has degrees of freedom that are $O(k^2)$. Davidson and MacKinnon (1992) found a similar tendency for the rejection rate of the information matrix test (in particular, the popular NR^2 form of it) to increase rapidly with k.

In Panel (d), for just a few cases, we compare the rejection frequencies for models that have fine-level fixed effects with ones for corresponding models that just have a constant term. For all values of G_c, the rejection frequencies seem to be slightly smaller for the latter models than for the former ones.

The bootstrap versions of the tests perform very much better than the asymptotic ones.

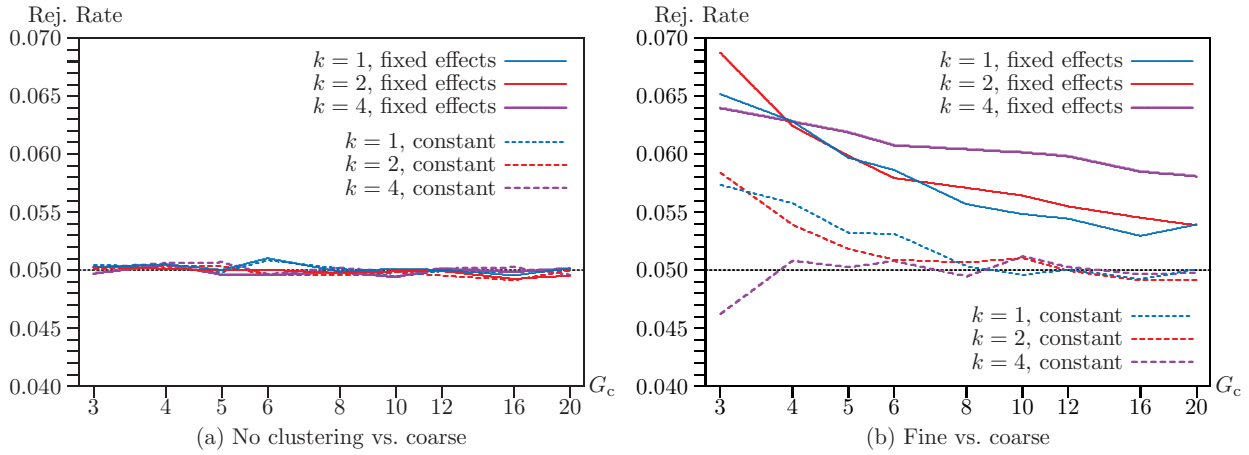
Figure 1: Rejection frequencies for asymptotic τ_Σ tests at 0.05 level



Notes: The data are generated by (38) with independent standard normal disturbances and $1 \leq k \leq 5$. G_c denotes the number of coarse clusters. There are $G_f = 4G_c$ fine clusters. Each fine cluster contains 100 observations, so that $N = 400G_c$. There are 400,000 replications.

Figure 2 is based on experiments similar to those that underlie Figure 1, but with 200,000 replications instead of 400,000 because they are much more expensive. Panel (a) shows rejection frequencies at the 0.05 level for the wild bootstrap test of no clustering against coarse clustering described in Section 3.2. These appear to be equal to 0.05 plus random noise. For clarity, there are only three values of k, namely, 1, 2, and 4. Panel (b) shows rejection frequencies at the 0.05 level for the wild cluster bootstrap test of fine against coarse clustering. With fixed effects, the bootstrap tests perform quite well (compare Panel (c) of Figure 1), but they always over-reject slightly. With just a constant term, on the other hand, the bootstrap tests perform very well even for G_c as small as 3. Note that this corresponds to

Figure 2: Rejection frequencies for bootstrap τ_Σ tests at 0.05 level



(a) No clustering vs. coarse

(b) Fine vs. coarse

Notes: The data are generated by (38) with independent standard normal disturbances. G_c denotes the number of coarse clusters. There are $G_f = 4G_c$ fine clusters. Each fine cluster contains 100 observations, so that $N = 400G_c$. Panel (a) uses the ordinary wild bootstrap, and Panel (b) uses the wild cluster bootstrap. There are 200,000 replications and 399 bootstraps.

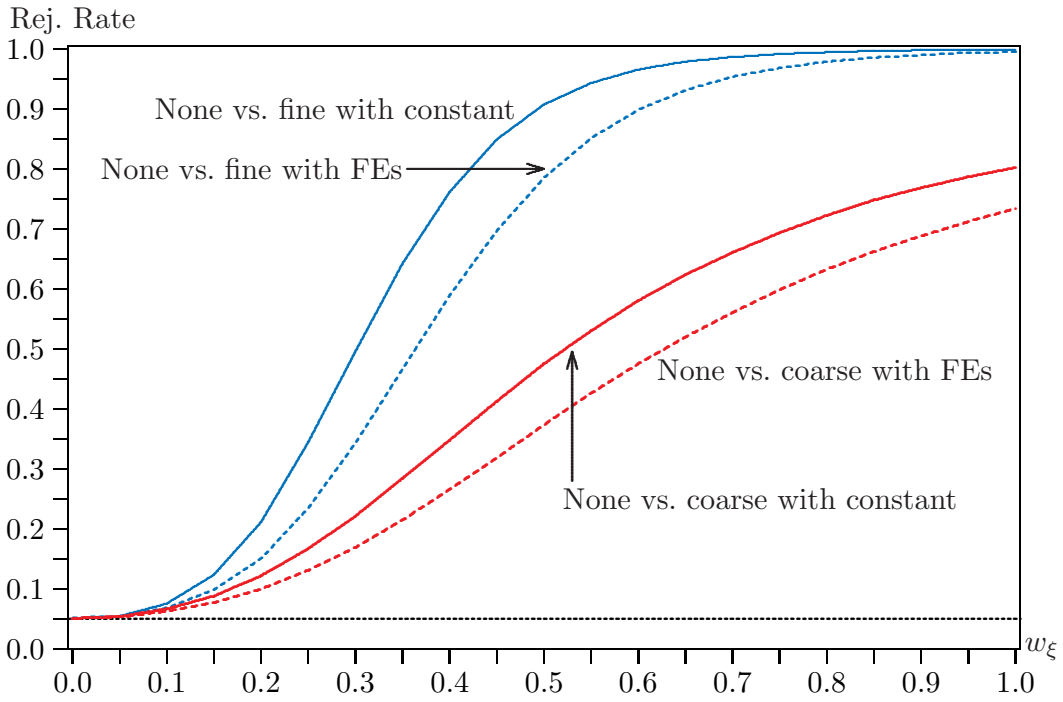
$G_f = 12$, so that, in this case, each wild cluster bootstrap sample is based on 12 values of v^{*b}.

The bootstrap tests can be computationally demanding when the sample size is large, particularly for larger values of k. This is especially true for tests where the null hypothesis is no clustering, because the calculations in (18), (19), and (29) involve score vectors of which the size is the number of clusters under the null hypothesis. This number is N for tests of no clustering but only G_f for tests of fine clustering.

In some additional experiments that are not reported, we find that asymptotic tests in models with coarse-level fixed effects work a bit better than the same tests in models with fine-level fixed effects, but not as well as in models with just a constant term. The same pattern is observed for bootstrap tests of fine versus coarse clustering, especially for larger values of k. It is not surprising that performance should be a bit better when the number of fixed effects is reduced by a factor of four.

In all the experiments reported here, the disturbances in the DGP are independent. For the tests of fine against coarse clustering, this is unnecessarily restrictive. In some additional unreported experiments, we allow there to be intra-cluster correlation at the fine level. Interestingly, we find that asymptotic tests of fine against coarse clustering perform a bit better as the amount of fine-level intra-cluster correlation increases, which is in line with Remark 11. On the other hand, bootstrap tests perform a bit worse. The differences are always very small, however.

Figure 3: Power of two-sided bootstrap τ_σ tests when there is fine clustering



Notes: The data are generated by (38) with two regressors. There are 8 coarse clusters, 32 fine clusters, and 3200 observations. The disturbances have fine clustering, with $\rho = 0.5$ and weight w_ξ between 0.0 and 1.0. There are 400,000 replications and 399 bootstraps.

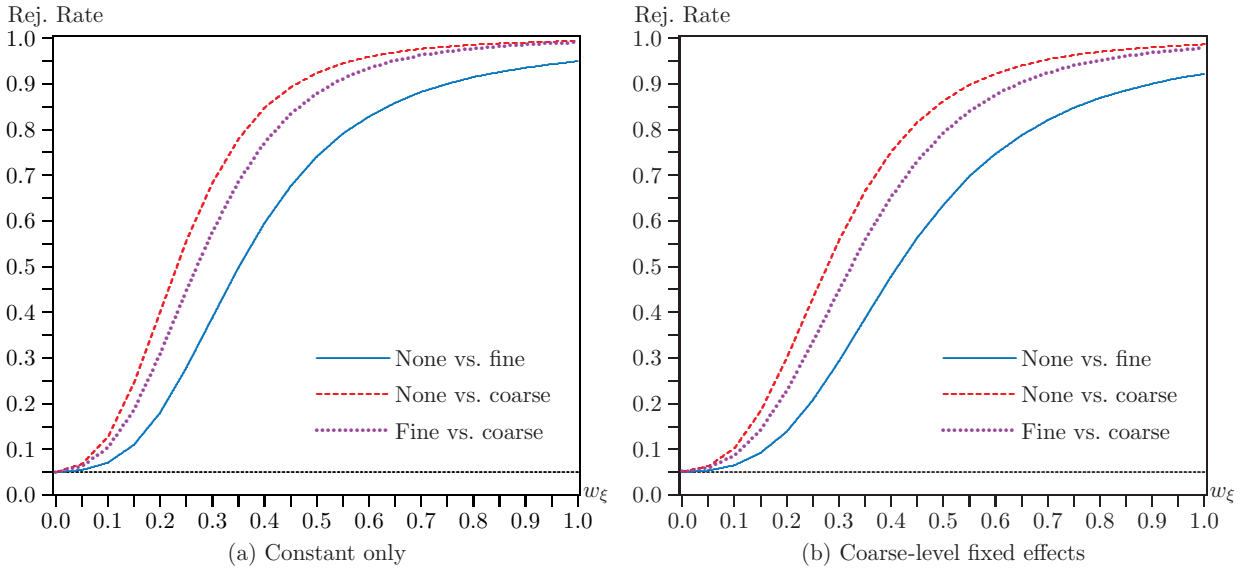
5.2 Performance under the Alternative Hypothesis

In the next set of experiments, we turn our attention to power, initially focusing on the special case of the τ_σ test for a single coefficient. The model actually contains two regressors, which are independent and have the same distribution. Including an additional regressor seems to have very little effect on the properties of the tests.

Figure 3 shows the power of four bootstrap τ_σ tests (no clustering against fine and no clustering against coarse, with and without fine-level fixed effects) at the 0.05 level as functions of w_ξ when there is fine clustering, with $G_\mathrm{c} = 8$, $G_\mathrm{f} = 32$, and $N = 3200$. Since the null hypothesis is true for tests of fine against coarse clustering, we do not show results for them. The null hypothesis is false for the tests of no clustering, and it is evident that they have power which increases monotonically with w_ξ. As expected, the tests have more power for the models with just a constant term than for the models with fine-level fixed effects.

Since clustering is actually at the fine level, it should be no surprise that the tests against fine clustering have more power than the tests against coarse clustering. The numerator of the test statistic, $\hat{\theta}$, equals either $\hat{\sigma}^2_\mathrm{c}$ or $\hat{\sigma}^2_\mathrm{f}$, both given in (20), minus $\hat{\sigma}^2_\mathrm{het}$ given in (21). In

Figure 4: Power of two-sided bootstrap τ_σ tests when there is coarse clustering



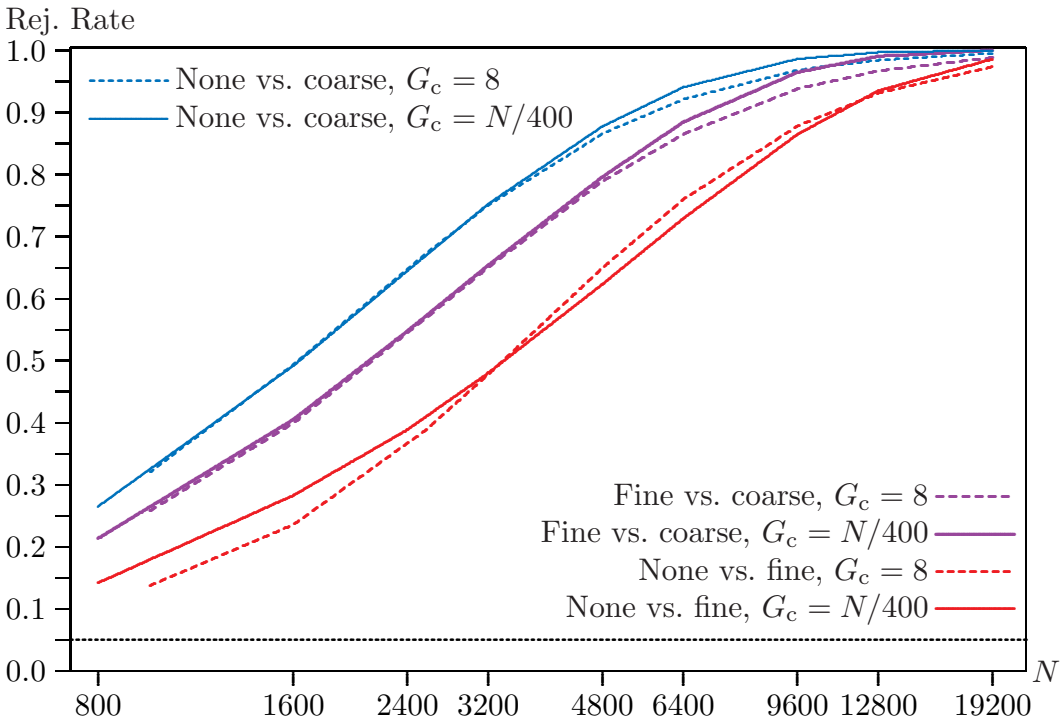
Notes: The data are generated by (38) with two regressors. There are 8 coarse clusters, 32 fine clusters, and 3200 observations. The disturbances have coarse clustering, with $\rho = 0.5$ and weight w_ξ between 0.0 and 1.0. There are 400,000 replications and 399 bootstraps.

the case of the test for fine clustering, all of the ζ_{ghi} that are implicitly being tested have non-zero means, but in the case of the test for coarse clustering only the ones that correspond to the same fine clusters do so. In other words, the test for coarse clustering incorporates many additional terms that are zero in expectation, when compared to the test for fine clustering, and consequently it has less power.

Figure 4 is similar to Figure 3, but there are two panels and the DGP now has coarse clustering. This means that, except when $w_\xi = 0$, the null hypothesis is false for all the tests. In Panel (a), there is only a constant term, and in Panel (b) there are coarse-level fixed effects. As before, any given test for a model with fixed effects has less power than the corresponding test for a model without fixed effects. In both panels, the test of no clustering against coarse clustering has the most power, followed by the test of fine against coarse clustering, followed in turn by the test of no clustering against fine clustering.

This ranking of the tests is quite different from the ranking in Figure 3, as it should be. Intuitively, the ranking follows from the fact that different numbers of terms with non-zero means contribute to the power of the tests. The test of no clustering against coarse clustering incorporates the largest number of terms that are non-zero in expectation under the alternative but not under the null. The test of no clustering against fine clustering omits many of these terms, because they do not belong to the same fine clusters. In contrast,

Figure 5: Power of two-sided bootstrap τ_σ tests with fixed effects and coarse clustering



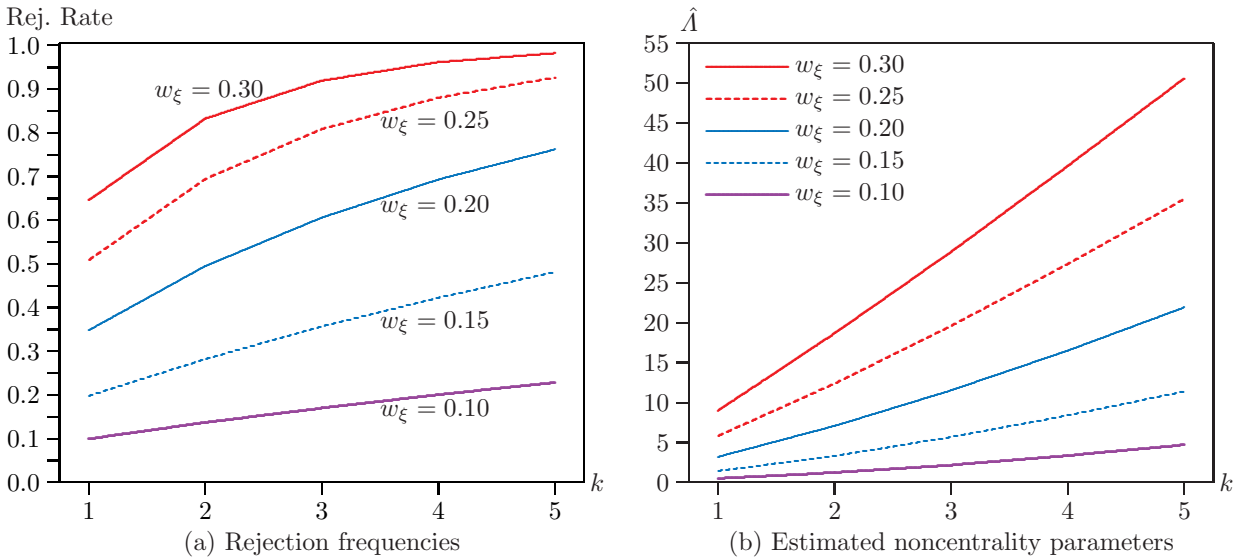
Notes: The data are generated by (38) with two regressors. There are either 8 coarse clusters with varying numbers of observations per cluster, or there are between 2 and 32 coarse clusters with 400 observations per cluster. The disturbances have coarse clustering, with $\rho = 0.5$ and $w_\xi = 0.4$. There are 400,000 replications and 399 bootstraps.

because the test of fine against coarse clustering has a different null, the terms that are non-zero in expectation and belong to the same fine clusters no longer contribute to power.

It is also of interest to see what happens to power as the sample size increases. There are many different ways in which this can happen. One of them is for the number of clusters to increase, with cluster sizes held constant. Another, which however does not satisfy the assumptions of Theorem 2, is for the number of observations per cluster to increase, with the number of clusters held constant. Figure 5 shows the power of the three bootstrap τ_σ tests as a function of N for both ways of increasing the sample size. All models include fixed effects at the fine level; power would be somewhat higher if they only included a constant term. Interestingly, power seems to increase with N at roughly the same rate for both ways of increasing the sample size.

In Figures 3–5, we report power only for bootstrap tests. The power functions for the asymptotic tests look very similar, and the figures would be too crowded if they were added. In these experiments, the power of the asymptotic tests is always higher than that of the bootstrap tests, but of course this apparently greater power is spurious and due to the fact

Figure 6: Power of asymptotic τ_Σ tests (fine vs. coarse) as a function of k



Notes: The data are generated by (38). Disturbances are clustered at the coarse level, with parameters $\rho = 0.5$ and w_ξ between 0.1 and 0.3. There are 40 coarse clusters, 320 fine clusters, 32,000 observations, and 400,000 replications. Tests are at the 0.05 level.

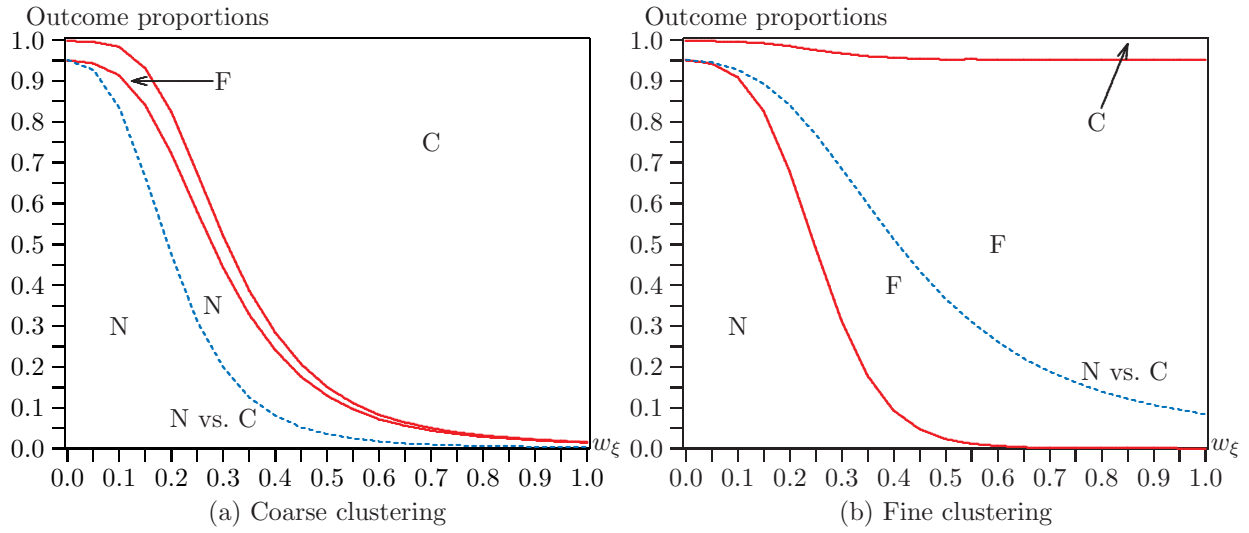
that the asymptotic tests over-reject noticeably under the null. We present additional results about the power of bootstrap and asymptotic tests in Appendix C.1, where we also compare the power of one-sided and two-sided tests.

In the remainder of this subsection, we focus on the τ_Σ tests for $k \geq 1$ coefficients. Since, by Corollary 1, these tests are asymptotically distributed as $\chi^2(d)$ under the null hypothesis, where $d = k(k+1)/2$, we expect their distribution under the alternative hypothesis to be approximately $\chi^2(d, \Lambda)$, where Λ is a noncentrality parameter that might be expected to increase with k if the scores were actually clustered for every coefficient.

In order to investigate this conjecture, we perform a set of experiments in which k varies from 1 to 5 and there is coarse clustering, the extent of which depends on the parameter w_ξ. In the hope that asymptotic approximations will be reasonably accurate, the model has 40 coarse clusters, 320 fine clusters, and 32,000 observations. We focus on tests of fine against coarse clustering. Panel (a) of Figure 6 shows rejection frequencies for asymptotic τ_Σ tests at the 0.05 level as a function of k for five values of w_ξ. It is evident that power increases with both w_ξ and k. Both relationships are inevitably nonlinear, because power is bounded above by 1 and below by (approximately) 0.05.

Panel (b) of Figure 6 is more revealing than Panel (a). The horizontal axis is the same, but the vertical axis shows $\hat{\Lambda}$, the estimated noncentrality parameter. For each value of k and w_ξ, $\hat{\Lambda}$ is computed as the mean of the 400,000 test statistics minus d. If the statistics

Figure 7: Outcomes for sequential two-sided bootstrap tests at 0.05 level



Notes: The data are generated by (38) with clustered disturbances at either the coarse level (left panel) or the fine level (right panel), for w_ξ between 0.0 and 1.0. There are 8 coarse clusters, 32 fine clusters, and 6400 observations. Bootstrap tests use $B = 999$, and there are 200,000 replications. The two solid red curves separate the three outcomes of the sequential procedure (N, F, and C). The dashed blue curve shows the outcome of a direct test of N against C.

were actually distributed as $\chi^2(d, \Lambda)$, then $\hat{\Lambda}$ would be an unbiased estimate of Λ. It can be seen from the figure that, to a very good approximation, $\hat{\Lambda}$ is linear in k. In fact, it is almost proportional to k.

Of course, the clustering of the regressors is important. In our experiments, every regressor is clustered in the same way. That is why $\hat{\Lambda}$ is almost proportional to k. If instead the regressors were such that, say, the scores for coefficient 1 were clustered but the scores for other coefficients were not, then we would expect the τ_σ test based on coefficient 1 to have more power than the τ_Σ test based on two or more coefficients.

5.3 The Sequential Testing Procedure

Our next set of experiments concerns the sequential testing procedure of Section 4.3, using bootstrap tests. These experiments are quite similar to the ones in Figures 3 and 4, except that there are 6400 observations instead of 3200. There are also 999 bootstrap samples instead of 399, in order to reduce the (already quite small) power loss caused by using a finite number (Davidson and MacKinnon 2000). Because both these changes to the experimental design increase computational cost, the number of replications is now 200,000. The model always contains fine-level fixed effects, and all of the tests are at the 0.05 level.

In Panel (a) of Figure 7, there is coarse clustering in the DGP, except when $w_\xi = 0$. In the latter case, as expected, the procedure chooses no clustering (N) almost exactly 95% of the time, fine clustering (F) almost exactly 4.75% of the time, and coarse clustering (C) almost exactly 0.25% of the time. These results illustrate why the sequential testing algorithm does not inflate the Type I error. In this case, the true null is rejected almost exactly α% of the time. Amongst the replications with false positives, the test concludes that fine clustering is appropriate about $(1 - \alpha)$% of the time and that coarse clustering is appropriate the remaining α% of the time.

As w_ξ increases, the procedure chooses N less and less often. Initially, it chooses both C and F more frequently. For F, the highest percentage is 10.16% when $w_\xi = 0.20$. At that point, the procedure is already choosing C 17.51% of the time. As w_ξ increases further, it chooses C more and more often, at the expense of both N and F. When $w_\xi = 1$, the largest possible value with our DGP, it chooses C 98.49% of the time, N 1.41%, and F just 0.01%.

Figure 7 does not show the outcomes of sequential tests of the model with just a constant term, because it would make the figure too hard to read. Since these tests are more powerful than the same tests in the model with fixed effects, the correct level of clustering is always chosen more often, for any value of w_ξ.

The sequential procedure inevitably has less power than testing no clustering directly against coarse clustering. The outcome of testing N directly against C at the 0.05 level is shown by the blue dashed curve in Panel (a). The gap between this curve and the one that separates the F and C regions shows the power loss from using the sequential procedure. This power loss arises for two reasons. First, the test of N against F has less power than the test of N against C; see Figure 4. Second, even when N is correctly rejected against F, the latter is sometimes not rejected against C. When the investigator finds coarse clustering more plausible than fine clustering, it may therefore make sense to test no clustering directly against the former rather than to employ the sequential procedure.

In Panel (b) of Figure 7, there is fine clustering in the DGP, except when $w_\xi = 0$. The sequential procedure again works very well. As w_ξ increases, it chooses no clustering a rapidly diminishing fraction of the time, which drops to essentially zero for $w_\xi \geq 0.6$. For larger values of w_ξ, it incorrectly chooses coarse clustering about 4.9% of the time.[3] Once again, the outcome of testing N directly against C is shown by the blue dashed line. This test works much less well than the sequential procedure, often failing to reject the false hypothesis that the disturbances are not clustered. Again, this is expected from the results

[3]This is just under the nominal level of the test for F against C and reflects the fact that the bootstrap test tends to under-reject very slightly when there is actually fine clustering. Recall that there are only 8 coarse and 32 fine clusters in these experiments, so it should not be a surprise that even the bootstrap test does not work quite perfectly.

in Figure 3; compare the two dashed lines in that figure.

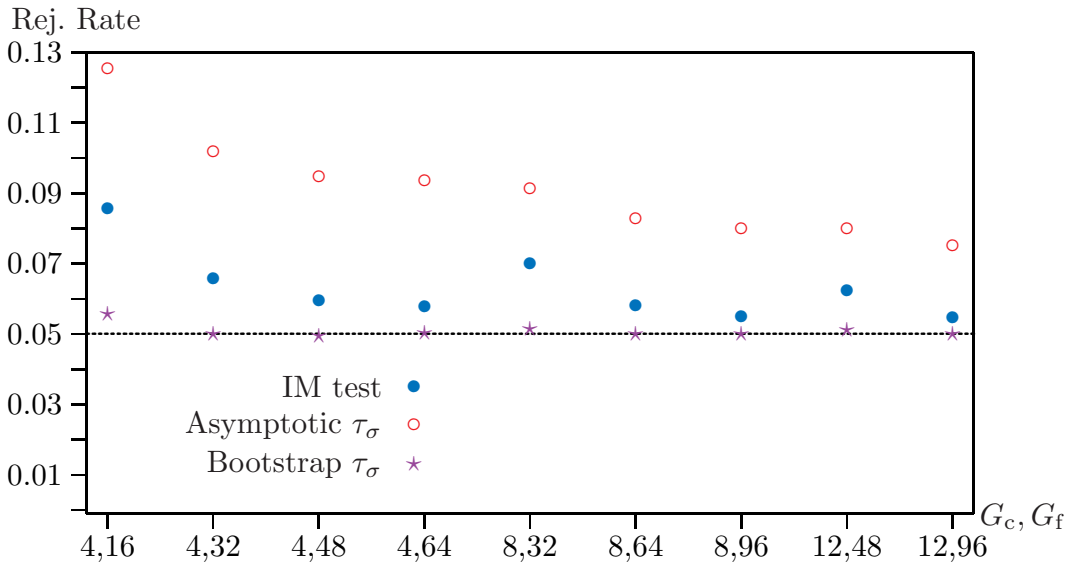
5.4 The Ibragimov-Müller Test

The only existing test for the level of clustering of which we are aware was proposed by Ibragimov and Müller (2016) and is described in detail in Appendix B. The IM test is based on estimating the model separately for every coarse cluster. Unfortunately, this is impossible to do whenever treatment is invariant within clusters. Even when it is possible to do it for some clusters, it may not be possible to do it for all of them. For example, in a difference-in-differences context, any cluster that is always treated or never treated will have to be omitted. Especially when many of the explanatory variables are dummies, there may also be perfect collinearity at the cluster level between the regressor of interest and some explanatory variables, or just among some of the latter. This is never a problem for our tests, because they do not require the model to be estimated on a cluster-by-cluster basis.

For our DGP, the IM test may be used to test either no clustering or fine clustering against coarse clustering. Since its performance seems likely to depend on the numbers of coarse and fine clusters, our simulations focus on varying these numbers. Because the IM test is one-sided, we compare it with a one-sided (upper tail) version of the τ_σ test rather than with the two-sided version studied so far. We do not consider sequential testing, because an IM test of no clustering against fine clustering assumes fine-level fixed effects, while an IM test of fine against coarse clustering assumes coarse-level fixed effects. Thus the sequence of IM tests is not nested, and consequently it is impossible to perform sequential IM tests for the same regression model. In contrast, our sequential tests always keep the model unchanged, including the level of the fixed effects, if any.

In the first set of experiments, the null hypothesis is true, $N = 9600$, the value of G_c is 4, 8, or 12, and the number of fine clusters per coarse cluster varies. Figure 8 shows rejection frequencies for three tests of fine against coarse clustering. As in Figure 2, the bootstrap τ_σ test works nearly perfectly. In the worst case, it rejects just 5.57% of the time. In contrast, the IM test over-rejects moderately, and the asymptotic τ_σ test over-rejects somewhat more seriously. These experiments differ in several respects from the ones in Figures 1 and 2. The sample size is three times as large, the fixed effects are at the coarse instead of the fine level, and the tests are one-sided instead of two-sided. For cases where G_c and G_f are the same as before, the asymptotic tests now reject somewhat more often, and the bootstrap tests reject less often. We do not report rejection frequencies for tests of no clustering against coarse clustering because the IM test, like the bootstrap τ_σ test, always seems to work very well; see Figures C.2 and C.3 in Appendix C for some examples.

Figure 8: Rejection frequencies for one-sided tests of fine against coarse clustering



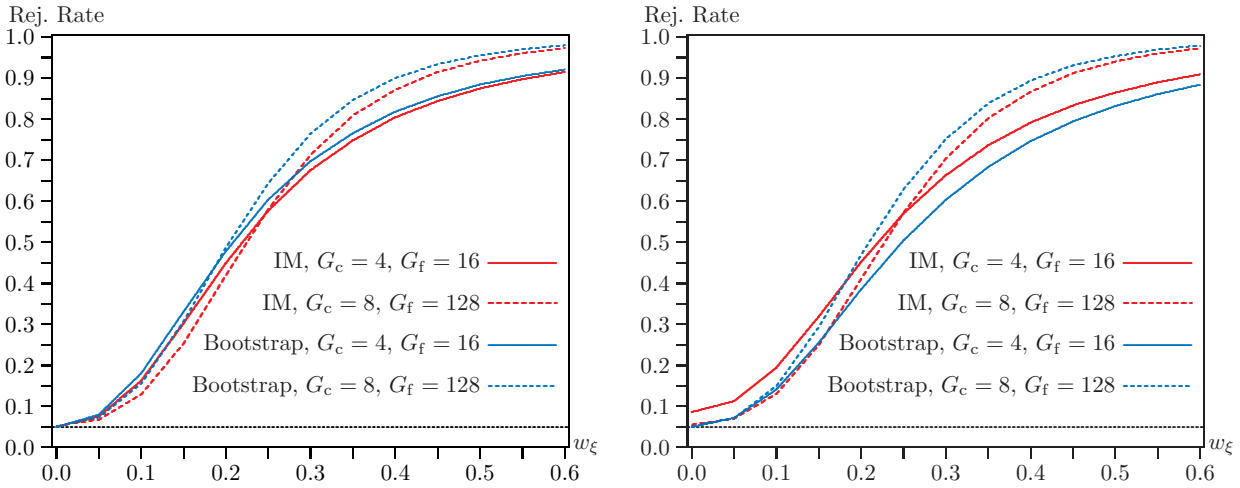
Notes: The data are generated by (38). There are 9600 observations and 4, 8, or 12 coarse clusters, together with various numbers of fine clusters per coarse cluster. There are coarse-level fixed effects, and the disturbances are uncorrelated. The only regressor has coarse clustering, with $\rho^x = 0.5$ and $w_\xi^x = 0.7$. There are 400,000 replications, 399 bootstraps, and 9,999 simulations for the IM tests. Tests are at the 0.05 level.

The next set of experiments concerns power as a function of the parameter w_ξ, which determines how much intra-cluster correlation the disturbances display. Panel (a) of Figure 9 shows four power functions for tests of no clustering against coarse clustering. There are two DGPs and two tests, namely, the IM test and the bootstrap τ_σ test. Not surprisingly, both tests have more power when there are more clusters, even though the sample size is unchanged. The IM test always has somewhat less power than the bootstrap score-variance test. Since both tests perform more or less perfectly under the null hypothesis, this lower power is clearly not spurious. For a large range of values of w_ξ, the difference in power when $G_c = 8$ is between 6 and 9 percentage points.

Panel (b) of Figure 9 shows power functions for tests of fine versus coarse clustering for the same experiments. When $G_c = 4$ and $G_f = 16$, the IM test over-rejects under the null, and it apparently has more power than the bootstrap score-variance test. Because of the over-rejection, this higher power is spurious. When $G_c = 8$ and $G_f = 128$, both tests perform extremely well under the null hypothesis, and the bootstrap score-variance test now has greater power than the IM test. The difference is between 3 and 6 percentage points for a large range of values of w_ξ.

On the basis of these, admittedly limited, experiments, we tentatively conclude that the IM test generally performs well under the null, but not as well as the bootstrap version of

Figure 9: Power of one-sided bootstrap τ_σ and IM tests at 0.05 level



(a) No clustering vs. coarse

(b) Fine clustering vs. coarse

Notes: The data are generated by (38) with two regressors. There are 6400 observations and either 4 coarse and 16 fine clusters or 8 coarse and 128 fine clusters. The disturbances have coarse clustering, with w_ξ varying on the horizontal axis, and there are coarse-level fixed effects. The regressor of interest has coarse clustering, with $\rho^x = 0.5$ and $w_\xi^x = 0.7$. There are 200,000 replications and 999 bootstraps.

our test. When both tests have the correct rejection frequency under the null, the bootstrap score-variance test seems to have somewhat more power. This is for cases where cluster-by-cluster estimation of the original model is possible for all clusters. When that is not possible, as in the STAR example of Section 6, our test and the IM test are actually testing different hypotheses, because the interpretation of the coefficients differs across clusters.

5.5 Additional Experiments

In additional simulation experiments that are discussed in Appendix C, we modify the design of these experiments in several ways. In particular, we consider DGPs for which there is heteroskedasticity in Appendix C.2 and ones for which the clusters vary in size in Appendix C.3. We also study a model in which the regressors are all dummy variables, taken from the STAR example of Section 6, in Appendix C.4. The principal conclusions of this section, and of Appendix C, can be summarized as follows.

Asymptotic score-variance tests perform well under the null hypothesis in many cases, but they can over-reject noticeably when either the total number of coarse clusters or the number of fine clusters per coarse cluster is small. This is particularly true for the experiments of Appendix C.4, where the combination of coarse-level fixed effects, dummy regressors, and few fine clusters per coarse cluster sometimes leads to severe over-rejection. The asymptotic tests

can also under-reject, sometimes severely, when cluster sizes vary a lot; see Appendix C.3.

Bootstrap score-variance tests generally work very well. The only case in which they do not work almost perfectly is for some of the experiments in Appendix C.4, where they nevertheless perform very much better than the asymptotic tests. When they can be calculated, IM tests of no clustering generally perform well, but IM tests of fine against coarse clustering can over-reject substantially in some cases. The power of all the tests is often very good. The sequential testing procedure seems to perform very well when the sample size and intra-cluster correlations are large enough for each of the sequential tests to have good power.

6 Empirical Example

We now illustrate the use of our score-variance tests, and of the IM test suggested by Ibragimov and Müller (2016), in a realistic empirical setting. Our example employs the widely-used data from the Tennessee Student Teacher Achievement Ratio (STAR) experiment (Finn and Achilles 1990; Mosteller 1995). We use these data to estimate a cross-sectional model similar to one in Krueger (1999). The STAR experiment randomly assigned students either to small-sized classes, regular-sized classes without a teacher's aide, or regular-sized classes with a teacher's aide. We are interested in the effect of being in a small class, or being in a class with an aide, on standardized test scores in reading.

We estimate the following cross-sectional regression model:

$$\text{read-one}_{sci} = \alpha + \beta_s \text{small-class}_{sc} + \beta_a \text{aide-class}_{sc} + \boldsymbol{x}_{sci}^\top \boldsymbol{\delta} + u_{sci}. \tag{39}$$

The outcome variable read-one_{sci} is the reading score in grade one of student i in classroom c in school s. We are interested in β_s and β_a, which are the coefficients for the small-class and aide-class dummies. Small-class equals 1 if a student attended a small class in grade one and equals 0 otherwise; aide-class is constructed in the same way for classes with or without teacher's aides. Additional control variables are collected in the vector of regressors $\boldsymbol{x}_{sci}$. These include dummy variables for whether the student was male, non-white, or received free lunches, as well as a dummy variable for whether the student's teacher was non-white. They also include the teacher's years of experience and the student's reading score in kindergarten. Finally, there are dummy variables for the student's quarter of birth, the student's year of birth, and the teacher's highest degree. There are thus 17 coefficients in total, not counting the constant term or the school fixed effects, if any.

OLS estimates for the model (39) are presented in the top half of Table 1. The model is estimated both without school fixed effects (left panel) and with school fixed effects (right

Table 1: STAR Example

Estimation		Without School FE HR	Without School FE CR-class	Without School FE CR-school	With School FE HR	With School FE CR-class	With School FE CR-school
small	$\hat{\beta}_s$	9.211	9.211	9.211	8.095	8.095	8.095
	s.e.	1.631	3.203	3.178	1.538	2.322	3.127
	t-stat.	5.649	2.876	2.899	5.263	3.486	2.589
aide	$\hat{\beta}_a$	6.245	6.245	6.245	4.170	4.170	4.170
	s.e.	1.661	3.260	2.790	1.569	2.109	2.422
	t-stat.	3.759	1.916	2.238	2.658	1.977	1.722

Cluster tests		Without School FE stat.	Without School FE asy. P	Without School FE boot P	With School FE stat.	With School FE asy. P	With School FE boot P	With School FE IM P
small	H_N vs H_R	28.388	0.000	0.000	12.757	0.000	0.000	—
	H_N vs H_S	16.409	0.000	0.000	18.308	0.000	0.000	0.251
	H_R vs H_S	−0.101	0.920	0.925	4.366	0.000	0.004	0.000
aide	H_N vs H_R	25.693	0.000	0.000	7.625	0.000	0.000	—
	H_N vs H_S	10.102	0.000	0.000	7.696	0.000	0.000	0.438
	H_R vs H_S	−1.765	0.080	0.084	1.871	0.061	0.344	0.000
both	H_N vs H_R	1075.469	0.000	0.000	180.448	0.000	0.000	—
	H_N vs H_S	322.367	0.000	0.000	385.950	0.000	0.000	—
	H_R vs H_S	5.215	0.157	0.171	28.673	0.000	0.011	—

Notes: There are 3,989 observations and either 330 (classroom) or 75 (school) clusters. Values of the τ_σ statistic (for "small" and "aide") or the τ_Σ statistic (for "both") are shown under "stat." All other numbers in the lower panel are P values. For the τ_σ tests, asymptotic P values are two-sided and based on the $N(0,1)$ distribution. For the τ_Σ tests, they are based on the $\chi^2(3)$ distribution. Bootstrap tests use $B = 99{,}999$. IM tests use $S = 9{,}999$.

panel). It is impossible to use classroom fixed effects, because treatment was assigned at the classroom level. Three sets of standard errors and t-statistics are reported for each variant of the model. One set is heteroskedasticity-robust (HR). The other two sets are cluster-robust (CR) at either the classroom level or the school level. Because treatment was assigned at the classroom level, it seems plausible that clustering at that level would be appropriate. However, since there are multiple classrooms per school, and students from the same school probably have many common characteristics and peer effects, it might also seem natural to cluster at the school level instead of the classroom level.

Unfortunately, the dataset does not contain a classroom indicator. One was created by using the information on the school ID, teacher's race, teacher's experience, teacher's highest degree, teacher's career ladder stage, and treatment status. It is possible that this procedure occasionally grouped two classes into one class, when two teachers in the same school had

exactly the same observable characteristics. However, since the largest observed class had only 29 students, this seems unlikely to have happened often. Moreover, it would not be a problem, because the true classes would always be nested within the larger, assumed class. What would be a problem is if classes were incorrectly partitioned, but this cannot happen.

Without school fixed effects, the estimate for the impact of being in a small class on test scores is $\hat{\beta}_s = 9.211$. Based on an HR standard error of 1.631, the t-statistic for the null hypothesis that $\beta_s = 0$ is 5.65. When we instead use CR standard errors clustered at the classroom level, the standard error for β_s increases to 3.203, and the t-statistic decreases to 2.88. Finally, when we use CR standard errors clustered at the school level, the standard error is 3.178, and the t-statistic is 2.90. The CR t-statistics provide quite strong evidence against the null, but not as strong as the evidence from the HR t-statistics. If one were to rely on the rule of thumb always to use the largest standard error, then one would want to cluster at the classroom level. However, if one instead relied on the rule of thumb to cluster at the coarsest possible cluster, then one would want to cluster at the school level.

When we estimate the model with school fixed effects, $\hat{\beta}_s = 8.095$, so it has not changed much. The HR t-statistic is now 5.26, the classroom-level CR t-statistic is 3.49, and the school-level CR t-statistic is 2.59. Thus, with school fixed effects, the "largest-standard-error" and "coarsest-clusters" approaches both suggest clustering at the school level.

The pattern is largely similar for the effect on test scores of being in a class with an aide. The estimate is $\hat{\beta}_a = 6.245$ without fixed effects and $\hat{\beta}_a = 4.170$ with fixed effects. The standard errors and t-statistics are shown in the bottom part of the top half of Table 1. The only surprising thing is that, without fixed effects, the CR standard error clustered at the school level (2.790) is smaller than the one clustered at the classroom level (3.260).

The lower panel of Table 1 shows the values of our test statistics, and the associated asymptotic and bootstrap P values, for the two coefficients of interest. It also shows results for the IM test for the model with school fixed effects, when that test can be calculated; see Appendix B. For each specification, we consider three hypotheses: $\mathrm{H_N}$ is no clustering with possible heteroskedasticity (HR), $\mathrm{H_R}$ is classroom-level clustering (CR-class), and $\mathrm{H_S}$ is school-level clustering (CR-school). These are nested as $\mathrm{H_N} \subseteq \mathrm{H_R} \subseteq \mathrm{H_S}$.

For testing $\mathrm{H_N}$ against $\mathrm{H_R}$ (that is, HR against CR-class), our tests, both asymptotic and bootstrap, very strongly reject the null in all cases. IM tests cannot be computed for this hypothesis, because the procedure requires the model to be estimated classroom by classroom, and the two treatment variables are invariant at that level; see Appendix B.

For testing $\mathrm{H_N}$ against $\mathrm{H_S}$ (that is, HR against CR-school), all our tests also very strongly reject the null in all cases. This is not surprising in view of the results for testing $\mathrm{H_N}$ against $\mathrm{H_R}$. Since there is overwhelming evidence against $\mathrm{H_N}$ when tested against $\mathrm{H_R}$, and classrooms

are nested within schools, there is inevitably also strong evidence against H_N when tested against H_S. IM tests can be computed when testing against H_S, but only for the model with school fixed effects. For both coefficients, the IM tests suggest that H_N should not be rejected. This is inconsistent with the results of the score-variance tests and surprising in view of the standard errors reported in the top part of the table.

The results for testing H_R against H_S differ considerably depending on the model, the coefficient(s) of interest, and the testing procedure. For small-class, the τ_σ test fails to reject when there are no school fixed effects, but it rejects quite strongly when there are. This makes sense, because the CR-class and CR-school standard errors are almost the same without fixed effects but quite different with them. The IM test also rejects in the latter case. For aide-class, the τ_σ test rejects H_R at the 10% level without fixed effects but fails to reject with them. The IM test rejects strongly in this case.

Many of the differences between the score-variance tests and the IM test in Table 1 probably arise because calculating the latter for the model (39) is tricky. The problem is that estimating all the coefficients for every one of the 75 schools is infeasible. For 34 schools, it is impossible to estimate at least one of β_s and β_a (17 schools in the case of β_s and 21 schools in the case of β_a). This means that the IM tests have to be based on either 58 or 54 coarse clusters, instead of all 75. Additionally, the other regressors that are included vary across clusters, so that the coefficients β_s and β_a may have different interpretations for different clusters. The IM tests may effectively be testing different null hypotheses than the score-variance tests, which are always based on estimates for the entire sample.

The τ_Σ tests for both coefficients reject in all but one case. The only exception is the test of H_R against H_S in the model with just a constant. This is not surprising, because neither of the τ_σ tests is significant in that case.

For the model with just a constant, the asymptotic and bootstrap P values are quite similar. However, for the model with fixed effects, the latter are often considerably larger than the former for the tests of H_N against H_R and H_R against H_S. Although it cannot be seen in the table, this is true even when the bootstrap P values are 0.000, because the bootstrap critical values are much larger than the asymptotic ones. For example, the test statistic for H_N against H_R for β_a is 7.625. The asymptotic critical value is, of course, 1.96, but the bootstrap critical value is 3.48.

It may be surprising that a bootstrap distribution should differ so greatly from its asymptotic counterpart when there are 330 classroom clusters. With 75 school fixed effects and 17 regressors, several of which vary only at the classroom level, the asymptotic critical values are evidently not to be trusted. Interestingly, when we perform a Monte Carlo experiment that combines the actual data for all regressors with independent, normally distributed dis-

turbances (in Appendix C.4), the 0.05 critical value for the test of H_N against H_R for β_a is 3.77. This is very much larger than the asymptotic critical value and much closer to the bootstrap critical value of 3.48. This suggests that, for datasets where key regressors often vary at the fine cluster level, it can be very important to bootstrap when there are coarse-level fixed effects. When the number of fine clusters per coarse cluster is small, even bootstrap tests may tend to over-reject somewhat for tests of H_R against H_S; see Appendix C.4.

In summary, our score-variance tests suggest that clustering at either the classroom or school level is essential, because the null hypothesis of no clustering is always strongly rejected against both alternatives. Whether we should cluster at the school or classroom level is not so clear. With just a constant term, the sequential testing procedure, using either asymptotic or bootstrap tests, suggests that we should choose H_R. However, with fixed effects, we should apparently choose H_R if interest focuses on β_s, but we should choose H_S otherwise.

7 Conclusion

Most empirical research that uses cluster-robust inference assumes that the level of clustering is known. When this strong assumption fails, the consequences can be serious. Clustering at too fine a level can result in tests that over-reject severely and confidence intervals that under-cover dramatically. However, clustering at too coarse a level can lead to loss of power and to confidence intervals that vary greatly in length across samples and are, on average, excessively long.

We have proposed two direct tests for the level of clustering in a linear regression model, which we call score-variance tests. Both tests are based on the variances of the scores for two nested levels of clustering, because it is these variances that appear in the "filling" of the sandwich covariance matrices that correspond to the two levels. Under the null hypothesis that the finer level is appropriate, many of these variances are zero. The test statistics are functions of the empirical counterparts of those variances. Tests based on them can be used either to test the null of no clustering against an alternative of clustering at a certain level or to test the null of "fine" clustering against alternatives of "coarser" clustering. We have also proposed a sequential procedure which can be used to determine the correct level of clustering without inflating the family-wise error rate.

It is often assumed that including group fixed effects in a model makes it unnecessary to use a CRVE. However, this is demonstrably false in many cases (Bertrand et al. 2004; MacKinnon 2019). Whenever it is appropriate to include fixed effects at a certain level, it may well also be appropriate to cluster at that level. Investigators should either do that routinely or test whether they need to do so by using our score-variance tests.

The simplest of our two tests is based on the statistic τ_σ. It has the form of a t-statistic and tests whether the variance of a particular coefficient estimate is the same for two different levels of clustering. It will be attractive whenever interest focuses on a single coefficient, and it can be implemented as either a one-sided or a two-sided test. The second variant, based on the Wald-like statistic τ_Σ, tests whether the covariance matrix of a vector of coefficient estimates is the same for two different levels of clustering. It is necessarily two-sided.

Our tests can be implemented as either asymptotic tests or as wild bootstrap tests. In Section 4 and Appendix A, we derive the asymptotic distribution of our tests, prove that they are consistent tests, and also prove the validity of the wild bootstrap implementations. In the simulation experiments of Section 5, the asymptotic tests often work well for tests of a single coefficient, but they can be seriously over-sized for tests of several coefficients. The problem is most severe when testing a moderate number of fine clusters against a small number of coarse clusters. For the empirical example of Section 6, where several regressors, including the key ones, vary only at the fine-cluster level, the asymptotic tests seem to be quite over-sized when there are school fixed effects. In contrast, the wild bootstrap versions perform very well under the null hypothesis in almost all cases.

Unlike the IM test proposed in Ibragimov and Müller (2016) and described in Appendix B, our tests do not require cluster-by-cluster estimation, which is impossible in many cases; see e.g. Section 6. In most of the simulation experiments of Section 5.4, where cluster-by-cluster estimation is always feasible, our tests seem to be a little more powerful than the IM test.

Both our simulation results and the empirical example suggest that the tests can have excellent power. In many cases, with both actual and simulated data, the value of the test statistic is so far beyond any reasonable critical value that we can reject the null hypothesis with something very close to certainty even without bothering to use the bootstrap. However, when our tests are used as pre-tests to choose the level of clustering, they inevitably make some Type I errors when the true clustering level is fine, and they inevitably make some Type II errors when the true clustering level is coarse but the sample size and the extent of coarse clustering are not large enough for rejection to occur all the time. The Type I errors lead to confidence intervals that over-cover, and the Type II errors lead to confidence intervals that under-cover; see Appendix C.5.

The score-variance tests we have proposed are intended to provide guidance for applied researchers. In our view, it should be routine to test any proposed level of clustering, including no clustering, against a coarser alternative whenever such an alternative is plausible. This is especially important when investigators are considering the use of heteroskedasticity-robust standard errors or clustering at a very fine level, such as by individual or by family. When there are three or more plausible levels of clustering, including no clustering at all, it

will often be attractive to employ the sequential procedure proposed in Section 3.3. In practice, it may be safest to report inferences based on more than one level of clustering, along with the outcomes of score-variance tests, as we did in Section 6.

Appendix A: Proofs of Main Results

A.1 Proof of Theorem 1

We give the proof for τ_σ only, so that, in particular, the matrices $\boldsymbol{A}$ and $\boldsymbol{Q}$ become the vectors $\boldsymbol{a}$ and $\boldsymbol{q}$. The proof for τ_Σ is essentially the same but with more complicated notation. Also, because the factors m_c and m_f both converge to 1, we can ignore them in the proof.

Recall the contrast $\theta = \sum_{g=1}^{G} \sum_{h_1=1}^{M_g} \sum_{h_2 \neq h_1}^{M_g} \zeta_{gh_1}\zeta_{gh_2}$ defined in (24). To prove the first result of the theorem, we show that

$$\frac{\hat{\theta} - \theta}{\sqrt{\mathrm{Var}(\theta)}} \xrightarrow{P} 0 \quad \text{and} \tag{A.1}$$

$$\frac{\theta}{\sqrt{\mathrm{Var}(\theta)}} \xrightarrow{d} \mathrm{N}(0,1). \tag{A.2}$$

Under Assumptions 1 and 2, it holds that $\sigma_g^2 = \sum_{h=1}^{M_g} \sigma_{gh}^2$. From (26) and Lemma A.4 we then find that

$$\mathrm{Var}(\theta) = 2 \sum_{g=1}^{G} \sum_{h_1=1}^{M_g} \sum_{h_2 \neq h_1}^{M_g} \sigma_{gh_1}^2 \sigma_{gh_2}^2 \geq c \sum_{g=1}^{G} \sigma_g^4. \tag{A.3}$$

It follows from Lemma A.2(i) and (A.3) that the left-hand side of (A.1) is

$$O_P\left(\frac{\sup_{g,h} N_{gh} \sup_g N_g}{(\sum_{g=1}^{G} \sigma_g^4)^{1/2}} \right) = o_P(1)$$

by the first condition of Assumption 5. This proves (A.1).

To prove (A.2), we write

$$\theta = \sum_{g=1}^{G} \sum_{h=1}^{M_g} w_{gh} \quad \text{with} \quad w_{gh} = 2\zeta_{gh} \sum_{j=1}^{h-1} \zeta_{gj}, \tag{A.4}$$

where we note that w_{gh} is a martingale difference sequence with respect to the filtration $\mathcal{F}_{gh} = \sigma(\{\zeta_{mn}\}_{m=1,\ldots,g-1, n=1,\ldots,M_m}, \{\zeta_{gn}\}_{n=1,\ldots,h})$, i.e. $\mathrm{E}(w_{gh}|\mathcal{F}_{g,h-1}) = 0$. Then (A.2) follows

from the martingale central limit theorem (e.g., Brown 1971, Theorem 2) if

$$\mathrm{Var}(\theta)^{-\lambda}\sum_{g=1}^{G}\sum_{h=1}^{M_g}\mathrm{E}|w_{gh}|^{2\lambda}\longrightarrow 0 \quad \text{for some } \lambda>1, \tag{A.5}$$

$$\mathrm{Var}(\theta)^{-1}\sum_{g=1}^{G}\sum_{h=1}^{M_g}\mathrm{E}(w_{gh}^2|\mathcal{F}_{g,h-1})\xrightarrow{P}1. \tag{A.6}$$

We first prove the Lyapunov condition in (A.5). We find $\mathrm{E}|\zeta_{gh}|^{2\lambda}\leq cN_{gh}^{2\lambda}$ by (17) and Lemma A.1. We also find that

$$\mathrm{E}\Big|\sum_{j=1}^{h-1}\zeta_{gj}\Big|^{2\lambda}\leq c\mathrm{E}\Big|\sum_{j=1}^{h-1}\zeta_{gj}^2\Big|^{\lambda}\leq c\Big|\sum_{j=1}^{h-1}(\mathrm{E}\zeta_{gj}^{2\lambda})^{1/\lambda}\Big|^{\lambda}\leq c\Big|\sum_{j=1}^{M_g}(N_{gj}^{2\lambda})^{1/\lambda}\Big|^{\lambda}\leq cN_g^{\lambda}\sup_{g,h}N_{gh}^{\lambda}, \tag{A.7}$$

where the first inequality is Marcinkiewicz-Zygmund, the second is Minkowski, and the third is due to Lemma A.1. Thus, we obtain the bound

$$\mathrm{E}|w_{gh}|^{2\lambda}\leq 2^{2\lambda}\mathrm{E}|\zeta_{gh}|^{2\lambda}\mathrm{E}\Big|\sum_{j=1}^{h-1}\zeta_{gj}\Big|^{2\lambda}\leq cN_{gh}^{2\lambda}N_g^{\lambda}\sup_{g,h}N_{gh}^{\lambda}, \tag{A.8}$$

and hence

$$\sum_{g=1}^{G}\sum_{h=1}^{M_g}\mathrm{E}|w_{gh}|^{2\lambda}\leq c\sup_{g,h}N_{gh}^{3\lambda-1}\sup_{g}N_g^{\lambda}N. \tag{A.9}$$

Combining (A.3) and (A.9), the Lyapunov condition in (A.5) is satisfied by the second condition of Assumption 5.

We next prove convergence of the conditional variance in (A.6). Because $\mathrm{Var}(\theta)$ equals $\sum_{g=1}^{G}\sum_{h=1}^{M_g}\mathrm{E}(w_{gh}^2)$, we decompose $\mathrm{E}(w_{gh}^2|\mathcal{F}_{g,h-1})-\mathrm{E}(w_{gh}^2)=q_{1,gh}+q_{2,gh}$, where $q_{1,gh}=\sigma_{gh}^2\sum_{j=1}^{h-1}(\zeta_{gj}^2-\sigma_{gj}^2)$ and $q_{2,gh}=\sigma_{gh}^2\sum_{j_1=1}^{h-1}\sum_{j_2\neq j_1}^{h-1}\zeta_{gj_1}\zeta_{gj_2}$. Then (A.6) follows if

$$\mathrm{Var}(\theta)^{-1}\sum_{g=1}^{G}\sum_{h=1}^{M_g}q_{m,gh}\xrightarrow{P}0 \quad \text{for } m=1,2. \tag{A.10}$$

For $m=1$, we reverse the summations and find that $\sum_{h=1}^{M_g}q_{1,gh}=\sum_{h=1}^{M_g}r_{1,gh}$, where $r_{1,gh}=(\zeta_{gh}^2-\sigma_{gh}^2)\sum_{j=h+1}^{M_g}\sigma_{gj}^2$ is mean zero and independent across both g and h. We prove convergence in L_λ-norm. We find $\mathrm{E}|r_{1,gh}|^{\lambda}\leq c\mathrm{E}|\zeta_{gh}|^{2\lambda}(\sum_{j=h+1}^{M_g}\sigma_{gj}^2)^{\lambda}\leq cN_{gh}^{2\lambda}(\sum_{j=1}^{M_g}\sigma_{gj}^2)^{\lambda}=cN_{gh}^{2\lambda}\sigma_g^{2\lambda}$ using Lemma A.1 and $\sum_{j=1}^{M_g}\sigma_{gj}^2=\sigma_g^2$. By the Marcinkiewicz-Zygmund and Minkowski inequalities we find that $\mathrm{E}\big|\sum_{g=1}^{G}\sum_{h=1}^{M_g}r_{1,gh}\big|^{\lambda}\leq c\Big(\sum_{g=1}^{G}\sum_{h=1}^{M_g}(\mathrm{E}|r_{1,gh}|^{\lambda})^{2/\lambda}\Big)^{\lambda/2}$, and hence

$$\mathrm{E}\Big|\sum_{g=1}^{G}\sum_{h=1}^{M_g}r_{1,gh}\Big|^{\lambda}\leq c\Big(\sum_{g=1}^{G}\sum_{h=1}^{M_g}N_{gh}^4\sigma_g^4\Big)^{\lambda/2}\leq\sup_{g,h}N_{gh}^{3\lambda/2}\sup_{g}N_g^{\lambda/2}\Big(\sum_{g=1}^{G}\sigma_g^4\Big)^{\lambda/2}.$$

Combining this with the bound (A.3), the result (A.10) for $m = 1$ follows if

$$\sup_{g,h} N_{gh}^{3\lambda/2} \sup_{g} N_g^{\lambda/2} \Big(\sum_{g=1}^{G} \sigma_g^4 \Big)^{-\lambda/2} \longrightarrow 0,$$

which is satisfied by the first condition of Assumption 5.

For $m = 2$, we use symmetry and reverse the summations to find $\sum_{h=1}^{M_g} q_{2,gh} = \sum_{h=1}^{M_g} r_{2,gh}$, where $r_{2,gh} = 2\zeta_{gh} \sum_{j_1=h+1}^{M_g} \sigma_{gj_1}^2 \sum_{j_2=1}^{h-1} \zeta_{gj_2} = w_{gh} \sum_{j=h+1}^{M_g} \sigma_{gj}^2$ is a martingale difference sequence with respect to $\mathcal{F}_{gh}$. We prove convergence in mean square. By (A.8) with $\lambda = 1$ the variance is $\mathrm{E}(r_{2,gh}^2) \leq \mathrm{E}(w_{gh}^2)\sigma_g^4 \leq cN_{gh}^2 N_g \sup_{g,h} N_{gh}\sigma_g^4$, and hence

$$\mathrm{E}\Big(\sum_{g=1}^{G} \sum_{h=1}^{M_g} r_{2,gh} \Big)^2 = \sum_{g=1}^{G} \sum_{h=1}^{M_g} \mathrm{E}(r_{2,gh}^2) \leq c \sup_{g,h} N_{gh}^2 \sup_{g} N_g^2 \sum_{g=1}^{G} \sigma_g^4.$$

Combining this with the bound (A.3), the result (A.10) for $m = 2$ follows by the first condition of Assumption 5. This completes the proof of (A.6) and hence of (A.2).

It remains to show the second part of Theorem 1. This follows directly from Lemma A.4 by application of Assumption 5 to the remainder terms.

A.2 Proof of Theorem 2

As in the proof of Theorem 1, we give the proof for τ_σ only, and we ignore the asymptotically irrelevant factors m_c and m_f. Under the conditions of Theorem 2, and specifically under Assumption 6, we find from (14) that $\mathbf{\Sigma}_\mathrm{c} = \sum_{g=1}^{G} \sigma_g^2$. However, it is important to note that, under the conditions of Theorem 2, $\sigma_g^2 = \mathrm{Var}(\zeta_g) \neq \sum_{h=1}^{M_g} \sigma_{gh}^2$.

We decompose the test statistic as follows:

$$\frac{\hat{\theta}}{\widehat{\mathrm{Var}}(\hat{\theta})^{1/2}} = \frac{\sum_{g=1}^{G} \sigma_g^2}{\widehat{\mathrm{Var}}(\hat{\theta})^{1/2}} \left(\frac{\hat{\theta} - \theta}{\sum_{g=1}^{G} \sigma_g^2} + \frac{\theta - \mathrm{E}(\theta)}{\sum_{g=1}^{G} \sigma_g^2} + \frac{\mathrm{E}(\theta)}{\sum_{g=1}^{G} \sigma_g^2} \right),$$

where we note that $\mathrm{E}(\theta) / \sum_{g=1}^{G} \sigma_g^2 = (\mathbf{\Sigma}_\mathrm{c} - \mathbf{\Sigma}_\mathrm{f})\mathbf{\Sigma}_\mathrm{c}^{-1}$ is non-zero in the limit under the alternative hypothesis, H_1 in (16). Thus, it suffices to prove that

$$\frac{\theta - \mathrm{E}(\theta)}{\sum_{g=1}^{G} \sigma_g^2} \xrightarrow{P} 0, \quad \frac{\hat{\theta} - \theta}{\sum_{g=1}^{G} \sigma_g^2} \xrightarrow{P} 0, \quad \text{and} \quad \frac{\widehat{\mathrm{Var}}(\hat{\theta})^{1/2}}{\sum_{g=1}^{G} \sigma_g^2} \xrightarrow{P} 0. \tag{A.11}$$

For the first result in (A.11), we prove convergence in mean square. The second moment

of the numerator is

$$\mathrm{E}\Big(\theta - \mathrm{E}(\theta)\Big)^2 = \mathrm{Var}(\theta) = \sum_{g=1}^{G} \mathrm{Var}\left(\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}^{M_g} \zeta_{gh_1}\zeta_{gh_2}\right) = \sum_{g=1}^{G} \mathrm{Var}\left(\zeta_g^2 - \sum_{h=1}^{M_g}\zeta_{gh}^2\right)$$
$$\leq c\sum_{g=1}^{G} N_g^4 \leq c \sup_g N_g^3 N, \tag{A.12}$$

where the second equality is by Assumption 6 and the penultimate inequality is by Lemma A.1 (applying the Cauchy-Schwarz inequality to the covariance terms). Hence, $\theta - \mathrm{E}(\theta)$ is $O_P(\sup_g N_g^{3/2} N^{1/2})$, which proves the first result in (A.11) by Assumption 7. The second result in (A.11) follows directly from Lemma A.2(ii) and Assumption 7. Finally, by the same methods as applied in the proof of (A.43), we find that

$$\widehat{\mathrm{Var}}(\hat{\theta}) = 2\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}^{M_g} \hat{\zeta}_{gh_1}^2\hat{\zeta}_{gh_2}^2 = O_P(\sup_{g,h} N_{gh}^2 \sup_g N_g N),$$

which proves the third result in (A.11) by Assumption 7.

A.3 Proof of Theorem 3

As in the proofs of Theorems 1 and 2, we give the proof for τ_σ only. The proof for τ_Σ is essentially the same but with slightly more complicated notation. The bootstrap probability measure is denoted P^*, and expectation under this measure is denoted E^*. We define the bootstrap contrast $\theta^* = \sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}^{M_g}\zeta_{gh_1}^*\zeta_{gh_2}^*$, and similarly the bootstrap variance estimator, and so on.

We prove the bootstrap analog of Theorem 1, but under the conditions of Theorem 3, which will establish the required result. Specifically, for all $x \in \mathbb{R}$ and all $\epsilon > 0$, we prove that

$$P^*\left(\frac{\hat{\theta}^*}{\sqrt{\mathrm{Var}^*(\theta^*)}} \leq x\right) \xrightarrow{P} \Phi(x) \quad \text{and} \quad P^*\left(\left|\frac{\widehat{\mathrm{Var}}(\hat{\theta}^*)}{\mathrm{Var}^*(\theta^*)} - 1\right| > \epsilon\right) \xrightarrow{P} 0, \tag{A.13}$$

as $N \to \infty$, where $\Phi(x)$ denotes the cumulative distribution function of the standard normal distribution. Clearly, (A.13) implies that $P^*(\tau_\sigma^* \leq x) \xrightarrow{P} \Phi(x)$. From Corollary 1 we have the result that $P_0(\tau_\sigma \leq x) \to \Phi(x)$. Because $\Phi(x)$ is everywhere continuous, the desired result then follows by application of the triangle inequality and Polya's Theorem.

Thus, we need to prove (A.13). We first note that, even though Assumption 1 is not imposed, it nonetheless holds by construction that, under the bootstrap probability measure P^*, the bootstrap data are clustered according to the fine structure in Assumption 1. Therefore, the proof of (A.13) largely follows that of Theorem 1. One main difference is that

$\sigma_g^2 = \text{Var}(\zeta_g) \neq \sum_{h=1}^{M_g} \sigma_{gh}^2$ because Assumption 1 is not imposed in Theorem 3.

We first establish the bootstrap equivalent of the lower bound in (A.3),

$$\text{Var}^*(\theta^*) \geq c\big(1 + o_P(1)\big) \sum_{g=1}^{G} \Big(\sum_{h=1}^{M_g} \sigma_{gh}^2 \Big)^2, \tag{A.14}$$

where we have used the fact that σ_g^4 in (A.3) needs to be replaced by $(\sum_{h=1}^{M_g} \sigma_{gh}^2)^2$ under the assumptions of Theorem 3. To prove (A.14), we first use $\zeta_{gh}^* = \hat{\zeta}_{gh} v_{gh}^*$, where v_{gh}^* is independent across both g and h, such that, c.f. (26) and (27),

$$\text{Var}^*(\theta^*) = \text{Var}^*\Big(\sum_{g=1}^{G} \sum_{h_1=1}^{M_g} \sum_{h_2 \neq h_1}^{M_g} \zeta_{gh_1}^* \zeta_{gh_2}^* \Big) = 2 \sum_{g=1}^{G} \sum_{h_1=1}^{M_g} \sum_{h_2 \neq h_1}^{M_g} \hat{\zeta}_{gh_1}^2 \hat{\zeta}_{gh_2}^2 = \widehat{\text{Var}}(\hat{\theta}).$$

The result in (A.14) now follows from Lemma A.4 by application of Assumption 5 to the remainder terms.

We next prove the following four results, which imply (A.13). For all $x \in \mathbb{R}$ and all $\epsilon > 0$,

$$P^*\left(\left| \frac{\hat{\theta}^* - \theta^*}{\sqrt{\text{Var}^*(\theta^*)}} \right| > \epsilon \right) \xrightarrow{P} 0, \tag{A.15}$$

$$P^*\left(\frac{\theta^*}{\sqrt{\text{Var}^*(\theta^*)}} \leq x \right) \xrightarrow{P} \Phi(x), \tag{A.16}$$

$$P^*\left(\left| \frac{\widehat{\text{Var}}(\theta^*)}{\text{Var}^*(\theta^*)} - 1 \right| > \epsilon \right) \xrightarrow{P} 0, \tag{A.17}$$

$$P^*\left(\left| \frac{\widehat{\text{Var}}(\hat{\theta}^*) - \widehat{\text{Var}}(\theta^*)}{\text{Var}^*(\theta^*)} \right| > \epsilon \right) \xrightarrow{P} 0, \tag{A.18}$$

as $N \to \infty$. The proofs of (A.15) and (A.16) are nearly identical to the corresponding proofs of (A.1) and (A.2). Similarly, the proofs of (A.17) and (A.18) are nearly identical to the corresponding proofs of (A.39) and (A.38). We therefore merely highlight the differences.

First, (A.15) follows by Markov's inequality and application of Lemma A.3, the lower bound (A.14), and Assumption 5.

Consider now (A.16). Under the bootstrap probability measure, $w_{gh}^* = 2\zeta_{gh}^* \sum_{j=1}^{h-1} \zeta_{gj}^*$ is a martingale difference sequence with respect to the filtration

$$\mathcal{F}_{gh}^* = \sigma\Big(\{v_{mn}^*\}_{m=1,\ldots,g-1, n=1,\ldots,M_m}, \{v_{gn}^*\}_{n=1,\ldots,h} \Big).$$

To verify the bootstrap equivalent of the Lyapunov condition, we apply the same proof as for (A.5). Replacing E with E*, the bounds (A.7)–(A.9) hold under the bootstrap measure with the right-hand sides being O_P of the indicated order by (A.24), (A.29), (A.30), and

Lemma A.1. Thus, in particular,

$$\sum_{g=1}^{G}\sum_{h=1}^{M_g}\mathrm{E}^*|w^*_{gh}|^{2\lambda} = O_P\Big(\sup_{g,h} N_{gh}^{3\lambda-1}\sup_g N_g N\Big),$$

which together with (A.14) and Assumption 5 verifies the Lyapunov condition for (A.16). For the proof of convergence of the conditional variance, we apply the same proof as for (A.6) with $r^*_{1,gh} = (\zeta^{*2}_{gh} - \hat\zeta^2_{gh})\sum_{j=h+1}^{M_g}\hat\zeta^2_{gj}$ and $r^*_{2,gh} = w^*_{gh}\sum_{j=h+1}^{M_g}\hat\zeta^2_{gj}$. For both terms we prove mean square convergence (because $\lambda \geq 2$). The arguments are nearly identical to those in the proof of (A.6), with all bounds being O_P of the indicated order, using (A.24), (A.29), (A.30), and Lemma A.1. This completes the proof of (A.16).

For the proof of (A.17), we follow the proof of (A.39) and obtain $q^*_{3,gh} = (\zeta^{*2}_{gh} - \hat\zeta^2_{gh})\sum_{j=1}^{h-1}\zeta^{*2}_{gh}$. We then apply the same proof as for $q_{3,gh}$ with $\lambda = 2$. Specifically, we find that there exists a set $\mathcal{A}^*$ with $P^*(\mathcal{A}^*) \xrightarrow{P} 1$, and on this set we have

$$\mathrm{Var}^*(q^*_{3,gh}) = O_P\bigg(N^4_{gh}\Big(\sum_{h-1}^{M_g}\sigma^2_{gh}\Big)^2\bigg),$$

where we used again (A.24), (A.29), (A.30), and Lemma A.1. Because $q^*_{3,gh}$ is a martingale difference sequence, the proof of (A.17) is concluded in the same way as that of (A.39).

Finally, we prove (A.18). As in (A.43)–(A.47), we write $\widehat{\mathrm{Var}}(\hat\theta^*) - \widehat{\mathrm{Var}}(\theta^*)$ as

$$2(\hat\beta^*_1 - \hat\beta_1)^2\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}^{M_g}(\hat\zeta^*_{gh_1}\hat\zeta^*_{gh_2} + \zeta^*_{gh_1}\zeta^*_{gh_2})\sum_{i=1}^{N_{gh_1}} z^2_{gh_1 i}\sum_{j=1}^{N_{gh_2}} z^2_{gh_2 j} \tag{A.19}$$

$$+ 8(\hat\beta^*_1 - \hat\beta_1)\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}^{M_g}\zeta^{*2}_{gh_1}\zeta^*_{gh_2}\sum_{i=1}^{N_{gh_2}} z^2_{gh_2 i} \tag{A.20}$$

$$- 4(\hat\beta^*_1 - \hat\beta_1)^2\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}^{M_g}\Big(\zeta^*_{gh_1}\sum_{i=1}^{N_{gh_2}} z^2_{gh_2 i} + \zeta^*_{gh_2}\sum_{j=1}^{N_{gh_1}} z^2_{gh_1 j}\Big)\zeta_{gh_1}\sum_{\ell=1}^{N_{gh_2}} z^2_{gh_2\ell} \tag{A.21}$$

$$- 4(\hat\beta^*_1 - \hat\beta_1)^2\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}^{M_g}\zeta^*_{gh_1}\Big(\sum_{j=1}^{N_{gh_1}} z^2_{gh_1 j}\Big)\Big(\sum_{i=1}^{N_{gh_2}} z^2_{gh_2 i}\Big)^2. \tag{A.22}$$

For (A.19), (A.21), and (A.22), we use (A.30) and (A.36) together with Lemma A.1, and find that

$$\mathrm{E}^*|(\text{A.19})| = O_P\bigg(N^{-1}\sup_{g,h} N_{gh}\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}^{M_g} N^2_{gh_1}N^2_{gh_2}\bigg) = O_P\Big(\sup_{g,h} N^3_{gh}\sup_g N_g\Big).$$

By the same argument, we also find the same bound for (A.21) and (A.22). Using (A.14) and the first condition of Assumption 5 shows the required result for these terms. For (A.20),

we apply the Cauchy-Schwarz inequality as in (A.48),

$$(\text{A.20})^2 \leq 64(\hat{\beta}_1 - \beta_{1,0})^2 \Big(\sup_{g,h} \sum_{i=1}^{N_{gh_1}} z_{gh_1 i}^2 \Big)^2 \Big(\sum_{g=1}^{G} \Big(\sum_{h_1=1}^{M_g} \zeta_{gh_1}^{*2} \Big)^2 \Big) \Big(\sum_{g=1}^{G} \Big(\sum_{h_2=1}^{M_g} \zeta_{gh_2}^{*} \Big)^2 \Big). \tag{A.23}$$

The first two factors on the right-hand side satisfy (A.36) and (A.30), respectively. The third factor is non-negative, and, under the bootstrap probability measure, it has a mean which is $O_P\big(\sum_{g=1}^{G} (\sum_{h=1}^{M_g} N_{gh}^2)^2 \big) = O_P(\sup_{g,h} N_{gh} \sup_g N_g N)$ using (A.24), (A.29), (A.30), and Lemma A.1. The last factor is non-negative and, under the bootstrap probability measure, it has a mean which is $O_P\big(\sum_{g=1}^{G} (\sum_{h=1}^{M_g} \sigma_{gh}^2)^2 \big)$ using again (A.24), (A.29), (A.30), and Lemma A.1. The proof for (A.23) is now completed in the same way as that of (A.48). This completes the proof of (A.18) and hence of Theorem 3.

A.4 Proof of Theorem 4

The result that $P(\hat{m} \leq m_0 - 1) \to 0$ is a direct consequence of Theorem 2 for the asymptotic tests and of Corollary 2(ii) for the bootstrap tests. In case (ii), where $m_0 = p$, there is nothing more to prove. In case (i) we have $m_0 \leq p - 1$. Because $P(\hat{m} \leq m_0 - 1) \to 0$, the sequential procedure will reach the test of the null hypothesis $m = m_0$ with probability converging to one. This is a test of a true null, so we find from Corollary 1 and Corollary 2(i) that $P(\hat{m} = m_0) \to 1 - \alpha$, which proves the theorem.

A.5 Auxiliary Lemmas

Lemma A.1. *Let Assumption 2 be satisfied. Then*

$$\sup_{g,h} N_{gh}^{-\xi} \mathrm{E} \|\boldsymbol{s}_{gh}\|^{\xi} = O(1) \quad \textit{and} \quad \sup_{g} N_{g}^{-\xi} \mathrm{E} \|\boldsymbol{s}_{g}\|^{\xi} = O(1) \quad \textit{for } 1 \leq \xi \leq 2\lambda.$$

Proof. This is Lemma A.2 of Djogbenou, MacKinnon, and Nielsen (2019). □

Lemma A.2. *Let Assumptions 2–4 be satisfied. Let $\hat{\boldsymbol{\theta}}$ be defined by (23) and also define $\boldsymbol{\theta} = \sum_{g=1}^{G} \sum_{h_1=1}^{M_g} \sum_{h_2 \neq h_1}^{M_g} \mathrm{vech}(\boldsymbol{\zeta}_{gh_1} \boldsymbol{\zeta}_{gh_2}^{\top})$; c.f. (24). Then*

(i) Under Assumption 1, $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| = O_P(\sup_{g,h} N_{gh} \sup_g N_g)$.

(ii) Under Assumption 6, $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| = O_P(N^{1/2} \sup_g N_g^{3/2})$.

Proof. We give the proof in the scalar case only. The proof for the multivariate case is nearly identical but with more complicated notation.

First recall that $\boldsymbol{z}$ and $\boldsymbol{X}_2$ are orthogonal by construction, such that

$$\hat{\zeta}_{ghi} = z_{ghi}\hat{u}_{ghi} = z_{ghi}u_{ghi} - z_{ghi}^2(\hat{\beta}_1 - \beta_{1,0}) = \zeta_{ghi}(1 + o_P(1)) - z_{ghi}^2(\hat{\beta}_1 - \beta_{1,0}), \tag{A.24}$$

where the second equality is by (17). Consequently, $\hat{\zeta}_{gh} = \zeta_{gh}(1+o_P(1)) - \sum_{i=1}^{N_{gh}} z_{ghi}^2(\hat{\beta}_1 - \beta_{1,0})$. In all applications of (A.24), we will omit the factor $(1 + o_P(1))$ since it is asymptotically irrelevant. From (20), (22), and (24), we then find the difference

$$\hat{\theta} - \theta = \sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}^{M_g}\Big(\sum_{i=1}^{N_{gh_1}} z_{gh_1i}^2(\hat{\beta}_1 - \beta_{1,0})\Big)\Big(\sum_{i=1}^{N_{gh_2}} z_{gh_2i}^2(\hat{\beta}_1 - \beta_{1,0})\Big) \tag{A.25}$$

$$- 2\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}^{M_g}\zeta_{gh_1}\sum_{i=1}^{N_{gh_2}} z_{gh_2i}^2(\hat{\beta}_1 - \beta_{1,0}). \tag{A.26}$$

Using (17) we find that $\hat{\beta}_1 - \beta_{1,0} = (\boldsymbol{z}^\top\boldsymbol{z})^{-1}\sum_{g=1}^{G}\sum_{h=1}^{M_g} z_{gh}u_{gh} = (\boldsymbol{z}^\top\boldsymbol{z})^{-1}\boldsymbol{q}^\top\sum_{g=1}^{G}\sum_{h=1}^{M_g}\boldsymbol{s}_{gh}$. Under Assumption 1 we have

$$\operatorname{Var}\Big(\sum_{g=1}^{G}\sum_{h=1}^{M_g}\boldsymbol{s}_{gh}\Big) = \sum_{g=1}^{G}\sum_{h=1}^{M_g}\operatorname{Var}(\boldsymbol{s}_{gh}) \leq c\sum_{g=1}^{G}\sum_{h=1}^{M_g}N_{gh}^2 \leq cN\sup_{g,h}N_{gh} \tag{A.27}$$

using Assumption 2 and Lemma A.1. Similarly, under Assumption 6,

$$\operatorname{Var}\Big(\sum_{g=1}^{G}\sum_{h=1}^{M_g}\boldsymbol{s}_{gh}\Big) = \sum_{g=1}^{G}\operatorname{Var}\Big(\sum_{h=1}^{M_g}\boldsymbol{s}_{gh}\Big) = \sum_{g=1}^{G}\operatorname{Var}(\boldsymbol{s}_g) \leq c\sum_{g=1}^{G}N_g^2 \leq cN\sup_{g}N_g. \tag{A.28}$$

Hence, using also Assumption 3,

$$\begin{aligned} |\hat{\beta}_1 - \beta_{1,0}| &= O_P\Big(N^{-1/2}\sup_{g,h}N_{gh}^{1/2}\Big) \quad \text{under Assumption 1,} \\ |\hat{\beta}_1 - \beta_{1,0}| &= O_P\Big(N^{-1/2}\sup_{g}N_{g}^{1/2}\Big) \quad \text{under Assumption 6.} \end{aligned} \tag{A.29}$$

We also need the simple bounds

$$\begin{aligned} &\sup_{g,h}N_{gh}^{-1}\sum_{i=1}^{N_{gh}}z_{ghi}^2 = \sup_{g,h}N_{gh}^{-1}\sum_{i=1}^{N_{gh}}\boldsymbol{a}^\top\boldsymbol{X}_{ghi}^\top\boldsymbol{X}_{ghi}\boldsymbol{a}(1+o_P(1)) = O_P(1) \quad \text{and} \\ &\sup_{g}N_g^{-1}\sum_{h=1}^{M_g}\sum_{i=1}^{N_{gh}}z_{ghi}^2 = O_P(1), \end{aligned} \tag{A.30}$$

which follow from (10) and the uniform moment bound in Assumption 3. Using (A.30), we

find that the absolute value of the right-hand side of (A.25) is bounded by

$$(\hat{\beta}_1-\beta_{1,0})^2\sum_{g=1}^{G}\Big(\sum_{h=1}^{M_g}\sum_{i=1}^{N_{gh}}z_{ghi}^2\Big)^2=(\hat{\beta}_1-\beta_{1,0})^2O_P\Big(\sum_{g=1}^{G}N_g^2\Big)=(\hat{\beta}_1-\beta_{1,0})^2O_P\Big(N\sup_g N_g\Big), \quad \text{(A.31)}$$

which proves the result for (A.25) using (A.29).

Next, we write (A.26) as

$$\text{(A.26)} = -2\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\zeta_{gh_1}\sum_{h_2=1}^{M_g}\sum_{i=1}^{N_{gh_2}}z_{gh_2i}^2(\hat{\beta}_1-\beta_{1,0}) \quad \text{(A.32)}$$

$$+2\sum_{g=1}^{G}\sum_{h=1}^{M_g}\zeta_{gh}\sum_{i=1}^{N_{gh}}z_{ghi}^2(\hat{\beta}_1-\beta_{1,0}). \quad \text{(A.33)}$$

Under Assumption 1, (A.27), (A.29), and (A.30) show that $|\text{(A.32)}| = O_P(\sup_{g,h} N_{gh} \sup_g N_g)$ and that $|\text{(A.33)}| = O_P(\sup_{g,h} N_{gh}^2)$, noting in both cases that (A.27) also holds with $\boldsymbol{s}_{gh}$ replaced by $|\zeta_{gh}|$. This proves the result for (A.26) under Assumption 1.

Under Assumption 6, we apply the Cauchy-Schwarz inequality,

$$|\text{(A.32)}| \le 2|\hat{\beta}-\beta_0|\Big(\sum_{g=1}^{G}\zeta_g^2\Big)^{1/2}\Big(\sum_{g=1}^{G}\Big(\sum_{h=1}^{M_g}\sum_{i=1}^{N_gh}z_{ghi}^2\Big)^2\Big)^{1/2},$$

$$|\text{(A.33)}| \le 2|\hat{\beta}-\beta_0|\Big(\sum_{g=1}^{G}\sum_{h=1}^{M_g}\zeta_{gh}^2\Big)^{1/2}\Big(\sum_{g=1}^{G}\sum_{h=1}^{M_g}\Big(\sum_{i=1}^{N_gh}z_{ghi}^2\Big)^2\Big)^{1/2}.$$

For both (A.32) and (A.33) we apply (A.29) to the first factor, Lemma A.1 to the second factor, and (A.30) to the third factor on the right-hand sides. This shows that (A.32) is $O_P(\sup_g N_g^{3/2}N^{1/2})$ and that (A.33) is $O_P(\sup_{g,h} N_{gh} \sup_g N_g^{1/2}N^{1/2})$, which proves the results for (A.32) and (A.33), and hence for (A.26), under Assumption 6. □

Lemma A.3. *Let Assumptions 2–4 and 6 be satisfied. Let* $\hat{\boldsymbol{\theta}}^* = \sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}^{M_g}\text{vech}(\boldsymbol{\zeta}_{gh_1}^*\boldsymbol{\zeta}_{gh_2}^{*\top})$ *and* $\boldsymbol{\theta}^* = \sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}^{M_g}\text{vech}(\boldsymbol{\zeta}_{gh_1}^*\boldsymbol{\zeta}_{gh_2}^{*\top})$. *Then*

$$\mathrm{E}^*\|\hat{\boldsymbol{\theta}}^*-\boldsymbol{\theta}^*\| = O_P\Big(\sup_{g,h}N_{gh}\sup_g N_g\Big).$$

Proof. The proof is very similar to that of Lemma A.2. Again, we give the proof in the scalar case only since the multivariate case is nearly identical but with more complicated notation.

We first write

$$\hat{\theta}^* - \theta^* = \sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}^{M_g}\Big(\sum_{i=1}^{N_{gh_1}} z_{gh_1i}^2(\hat{\beta}^* - \hat{\beta})\Big)\Big(\sum_{i=1}^{N_{gh_2}} z_{gh_2i}^2(\hat{\beta}^* - \hat{\beta})\Big) \tag{A.34}$$

$$- 2\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}^{M_g}\zeta_{gh_1}^*\sum_{i=1}^{N_{gh_2}} z_{gh_2i}^2(\hat{\beta}^* - \hat{\beta}). \tag{A.35}$$

As in (A.27), we find that

$$\mathrm{Var}^*\Big(\sum_{g=1}^{G}\sum_{h=1}^{M_g}\zeta_{gh}^*\Big) = \mathrm{Var}^*\Big(\sum_{g=1}^{G}\sum_{h=1}^{M_g}\hat{\zeta}_{gh}v_{gh}^*\Big) = \sum_{g=1}^{G}\sum_{h=1}^{M_g}\hat{\zeta}_{gh}^2 = O_P\Big(N\sup_{g,h}N_{gh}\Big),$$

where the second equality uses independence of v_{gh}^* across g and h and the third equality uses (A.24), (A.29), (A.30), and Lemma A.1. It follows that

$$\mathrm{Var}^*(\hat{\beta}^* - \hat{\beta}) = O_P\Big(N^{-1}\sup_{g,h}N_{gh}\Big). \tag{A.36}$$

Using (A.30) and (A.36), we find that $\mathrm{E}^*|(\text{A.34})| = O_P(\sup_{g,h}N_{gh}\sup_g N_g)$ as in (A.31). By the same argument, see also (A.32)–(A.33), we find that Var^* (A.35) $= O_P(\sup_{g,h}N_{gh}^2\sup_g N_g^2)$. □

Lemma A.4. *Let Assumptions 2–4 be satisfied and let* $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})$ *be given by (29). Suppose also that either (i) Assumption 1 or (ii) Assumption 6 and* $\lambda \geq 2$ *is satisfied. Then, for an arbitrary conforming, non-zero vector* $\boldsymbol{\delta}$ *and* $\boldsymbol{\delta}^\top \boldsymbol{H}_k = \boldsymbol{b}^\top = [\boldsymbol{b}_1^\top \otimes \boldsymbol{b}_2^\top]$,

$$\widehat{\mathrm{Var}}(\boldsymbol{\delta}^\top\hat{\boldsymbol{\theta}}) - 2\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}^{M_g}\boldsymbol{b}_1^\top\boldsymbol{\Sigma}_{gh_1}\boldsymbol{b}_1\boldsymbol{b}_2^\top\boldsymbol{\Sigma}_{gh_2}\boldsymbol{b}_2 = O_P\Big(N^{1/\lambda}\sup_g N_g\sup_{g,h}N_{gh}^{3-1/\lambda}\Big)$$
$$+ O_P\Big(\sup_{g,h}N_{gh}^2\sup_g N_g^2\Big) + O_P\Bigg(\sup_{g,h}N_{gh}\sup_g N_g\Big(\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2=1}^{M_g}\boldsymbol{b}_1^\top\boldsymbol{\Sigma}_{gh_1}\boldsymbol{b}_1\boldsymbol{b}_2^\top\boldsymbol{\Sigma}_{gh_2}\boldsymbol{b}_2\Big)^{1/2}\Bigg)$$

and

$$\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}^{M_g}\boldsymbol{b}_1^\top\boldsymbol{\Sigma}_{gh_1}\boldsymbol{b}_1\boldsymbol{b}_2^\top\boldsymbol{\Sigma}_{gh_2}\boldsymbol{b}_2 \geq c\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2=1}^{M_g}\boldsymbol{b}_1^\top\boldsymbol{\Sigma}_{gh_1}\boldsymbol{b}_1\boldsymbol{b}_2^\top\boldsymbol{\Sigma}_{gh_2}\boldsymbol{b}_2.$$

Proof. We give the proof of the first result for the univariate case, where $\widehat{\mathrm{Var}}(\hat{\theta})$ is given by (27), and we show that

$$\widehat{\mathrm{Var}}(\hat{\theta}) - 2\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}^{M_g}\sigma_{gh_1}^2\sigma_{gh_2}^2 = O_P\Big(N^{1/\lambda}\sup_g N_g\sup_{g,h}N_{gh}^{3-1/\lambda}\Big)$$
$$+ O_P\Big(\sup_{g,h}N_{gh}^2\sup_g N_g^2\Big) + O_P\Bigg(\sup_{g,h}N_{gh}\sup_g N_g\Big(\sum_{g=1}^{G}\Big(\sum_{h=1}^{M_g}\sigma_{gh}^2\Big)^2\Big)^{1/2}\Bigg). \tag{A.37}$$

The proof for the multivariate case is nearly identical, but with more complicated notation.

We decompose the left-hand side of (A.37) as

$$2\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}^{M_g}(\hat{\zeta}_{gh_1}^2\hat{\zeta}_{gh_2}^2-\zeta_{gh_1}^2\zeta_{gh_2}^2) \tag{A.38}$$

$$+2\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}^{M_g}(\zeta_{gh_1}^2\zeta_{gh_2}^2-\sigma_{gh_1}^2\sigma_{gh_2}^2). \tag{A.39}$$

We first prove the result for (A.39) under Assumption 1. We use (26) and (27) to write

$$\text{(A.39)}=4\sum_{g=1}^{G}\sum_{h=1}^{M_g}(q_{1,gh}+q_{3,gh}), \tag{A.40}$$

where $q_{1,gh}=\sigma_{gh}^2\sum_{j=1}^{h-1}(\zeta_{gj}^2-\sigma_{gj}^2)$ and $q_{3,gh}=(\zeta_{gh}^2-\sigma_{gh}^2)\sum_{j=1}^{h-1}\zeta_{gj}^2$. Under Assumption 1 we have already proven in (A.10) that $q_{1,gh}=O_P(\sup_{g,h}N_{gh}^{3/2}\sup_g N_g^{1/2}(\sum_{g=1}^{G}\sigma_g^4)^{1/2})$. The sequence $q_{3,gh}$ is a martingale difference with respect to the filtration $\mathcal{F}_{gh}$ defined just below (A.4). When $1<\lambda<2$, we prove convergence in L_λ-norm. By the von Bahr-Esseen inequality, $\mathrm{E}\big|\sum_{g=1}^{G}\sum_{h=1}^{M_g}q_{3,gh}\big|^\lambda\leq 2\sum_{g=1}^{G}\sum_{h=1}^{M_g}\mathrm{E}|q_{3,gh}|^\lambda$, where $\mathrm{E}|q_{3,gh}|^\lambda\leq\mathrm{E}|\zeta_{gh}|^{2\lambda}\mathrm{E}\big|\sum_{j=1}^{h-1}\zeta_{gj}^2\big|^\lambda$, which was analyzed in (A.7). The remainder of the proof for $q_{3,gh}$ with $1<\lambda<2$ is identical to that of the Lyapunov condition in (A.5), showing that $\sum_{g=1}^{G}\sum_{h=1}^{M_g}q_{3,gh}=O_P(N^{1/\lambda}\sup_g N_g\sup_{g,h}N_{gh}^{3-1/\lambda})$.

Next, suppose $\lambda\geq 2$. We find that $\sum_{j=1}^{h-1}\zeta_{gj}^2\leq\sum_{j=1}^{M_g}\zeta_{gj}^2$ is a non-negative random variable, and hence is of order $O_P(\mathrm{E}\sum_{j=1}^{M_g}\zeta_{gj}^2)=O_P(\sum_{h=1}^{M_g}\sigma_{gh}^2)$. That is, there exists a constant $K<\infty$ and a set $\mathcal{A}$ with $P(\mathcal{A})\to 1$ on which $\sum_{j=1}^{h-1}\zeta_{gj}^2\leq K\sum_{h=1}^{M_g}\sigma_{gh}^2$. Then, on the set $\mathcal{A}$,

$$\mathrm{E}(q_{3,gh}^2|\mathcal{F}_{g,h-1})=\mathrm{Var}(\zeta_{gh}^2)\Big(\sum_{j=1}^{h-1}\zeta_{gj}^2\Big)^2\leq K^2\Big(\sum_{h=1}^{M_g}\sigma_{gh}^2\Big)^2\mathrm{Var}(\zeta_{gh}^2), \tag{A.41}$$

and therefore

$$\mathrm{Var}(q_{3,gh})\leq cN_{gh}^4\Big(\sum_{h=1}^{M_g}\sigma_{gh}^2\Big)^2 \tag{A.42}$$

by Lemma A.1. Using (A.42) and the fact that $q_{3,gh}$ is a martingale difference sequence, it follows that, on the set $\mathcal{A}$,

$$\mathrm{Var}\left(\sum_{g=1}^{G}\sum_{h=1}^{M_g}q_{3,gh}\right)=\sum_{g=1}^{G}\sum_{h=1}^{M_g}\mathrm{Var}(q_{3,gh})\leq c\sup_{g,h}N_{gh}^3\sup_g N_g\Big(\sum_{h=1}^{M_g}\sigma_{gh}^2\Big)^2.$$

This shows the required result for $q_{3,gh}$ on the set $\mathcal{A}$ when $\lambda\geq 2$. Because $P(\mathcal{A})\to 1$, this completes the proof for (A.39) under Assumption 1.

We now prove the result for (A.39) under Assumption 6 and $\lambda \geq 2$. We again apply the decomposition in (A.40). Define $Q_{m,g} = \sum_{h=1}^{M_g} q_{m,gh}$ for $m = 1, 3$, which are both independent across g by Assumption 6. For $Q_{1,g}$ we note that $\sum_{j=h+1}^{M_g} \sigma_{gj}^2 \leq \sum_{j=1}^{M_g} \sigma_{gj}^2$ and apply the Cauchy-Schwarz inequality such that

$$\mathrm{E}(Q_{1,g}^2) \leq \Big(\sum_{h=1}^{M_g} \sigma_{gh}^2 \Big)^2 \mathrm{E}\Big(\sum_{h=1}^{M_g} |\zeta_{gh}^2 - \sigma_{gh}^2| \Big)^2 \leq \Big(\sum_{h=1}^{M_g} \sigma_{gh}^2 \Big)^2 \Big(\sum_{h=1}^{M_g} (\mathrm{E}(\zeta_{gh}^2 - \sigma_{gh}^2)^2)^{1/2} \Big)^2,$$

where last factor on the right-hand side is $O(\sup_{g,h} N_{gh}^2 \sup_g N_g^2)$ by Lemma A.1. Because $Q_{1,g}$ has mean zero and is independent across g, it follows that

$$\mathrm{Var}\Big(\sum_{g=1}^{G} Q_{1,g} \Big) = \sum_{g=1}^{G} \mathrm{E}(Q_{1,g}^2) \leq c \sup_{g,h} N_{gh}^2 \sup_g N_g^2 \sum_{g=1}^{G} \Big(\sum_{h=1}^{M_g} \sigma_{gh}^2 \Big)^2,$$

which proves the result for $Q_{1,g}$. For $Q_{3,g}$ we note that there exists a constant $K < \infty$ and a set $\mathcal{A}$ with $P(\mathcal{A}) \to 1$ such that, on $\mathcal{A}$, it holds that $\sum_{j=1}^{h-1} \zeta_{gj}^2 \leq K \sum_{j=1}^{M_g} \sigma_{gj}^2$. We can then apply the same proof as for $Q_{1,g}$. This completes the proof for (A.39) under Assumption 6.

To prove the result for (A.38), we use (27) and (A.24) to write

$$\begin{aligned}
\text{(A.38)} &= 2 \sum_{g=1}^{G} \sum_{h_1=1}^{M_g} \sum_{h_2 \neq h_1}^{M_g} (\hat{\zeta}_{gh_1} \hat{\zeta}_{gh_2} + \zeta_{gh_1} \zeta_{gh_2})(\hat{\zeta}_{gh_1} \hat{\zeta}_{gh_2} - \zeta_{gh_1} \zeta_{gh_2}) \\
&= 2(\hat{\beta}_1 - \beta_{1,0})^2 \sum_{g=1}^{G} \sum_{h_1=1}^{M_g} \sum_{h_2 \neq h_1}^{M_g} (\hat{\zeta}_{gh_1} \hat{\zeta}_{gh_2} + \zeta_{gh_1} \zeta_{gh_2}) \sum_{i=1}^{N_{gh_1}} z_{gh_1 i}^2 \sum_{j=1}^{N_{gh_2}} z_{gh_2 j}^2 \qquad \text{(A.43)} \\
&\quad + 4(\hat{\beta}_1 - \beta_{1,0}) \sum_{g=1}^{G} \sum_{h_1=1}^{M_g} \sum_{h_2 \neq h_1}^{M_g} (\hat{\zeta}_{gh_1} \hat{\zeta}_{gh_2} + \zeta_{gh_1} \zeta_{gh_2}) \zeta_{gh_1} \sum_{i=1}^{N_{gh_2}} z_{gh_2 i}^2. \qquad \text{(A.44)}
\end{aligned}$$

By another application of (A.24) followed by straightforward application of (A.29), (A.30), and Lemma A.1, it follows that (A.43) is of order $O_P(\sup_{g,h} N_{gh}^2 \sup_g N_g^2)$.

For (A.44), we apply again (A.24) and write

$$\begin{aligned}
\text{(A.44)} &= 8(\hat{\beta}_1 - \beta_{1,0}) \sum_{g=1}^{G} \sum_{h_1=1}^{M_g} \sum_{h_2 \neq h_1}^{M_g} \zeta_{gh_1}^2 \zeta_{gh_2} \sum_{i=1}^{N_{gh_2}} z_{gh_2 i}^2 \qquad \text{(A.45)} \\
&\quad - 4(\hat{\beta}_1 - \beta_{1,0})^2 \sum_{g=1}^{G} \sum_{h_1=1}^{M_g} \sum_{h_2 \neq h_1}^{M_g} \Big(\zeta_{gh_1} \sum_{i=1}^{N_{gh_2}} z_{gh_2 i}^2 + \zeta_{gh_2} \sum_{j=1}^{N_{gh_1}} z_{gh_1 j}^2 \Big) \zeta_{gh_1} \sum_{\ell=1}^{N_{gh_2}} z_{gh_2 \ell}^2 \qquad \text{(A.46)} \\
&\quad - 4(\hat{\beta}_1 - \beta_{1,0})^3 \sum_{g=1}^{G} \sum_{h_1=1}^{M_g} \sum_{h_2 \neq h_1}^{M_g} \zeta_{gh_1} \Big(\sum_{j=1}^{N_{gh_1}} z_{gh_1 j}^2 \Big) \Big(\sum_{i=1}^{N_{gh_2}} z_{gh_2 i}^2 \Big)^2. \qquad \text{(A.47)}
\end{aligned}$$

Direct application of (A.29), (A.30), and Lemma A.1 shows that (A.46) is $O_P(\sup_{g,h} N_{gh}^2 \sup_g N_g^2)$

and that (A.47) is $O_P(N^{-1/2}\sup_{g,h} N_{gh}^2 \sup_g N_g^{5/2}) = O_P(\sup_{g,h} N_{gh}^2 \sup_g N_g^2)$. Finally, for the right-hand side of (A.45), we apply the Cauchy-Schwarz inequality,

$$\begin{aligned}(\text{A.45})^2 &\leq 64(\hat{\beta}_1 - \beta_{1,0})^2 \Bigg(\sum_{g=1}^{G} \Bigg(\sum_{h_1=1}^{M_g} \zeta_{gh_1}^2 \sum_{i=1}^{N_{gh_1}} z_{gh_1 i}^2 \Bigg)^2 \Bigg) \Bigg(\sum_{g=1}^{G} \Bigg(\sum_{h_2=1}^{M_g} \zeta_{gh_2} \Bigg)^2 \Bigg) \\ &\leq 64(\hat{\beta}_1 - \beta_{1,0})^2 \Bigg(\sup_{g,h} \sum_{i=1}^{N_{gh_1}} z_{gh_1 i}^2 \Bigg)^2 \Bigg(\sum_{g=1}^{G} \Bigg(\sum_{h_1=1}^{M_g} \zeta_{gh_1}^2 \Bigg)^2 \Bigg) \Bigg(\sum_{g=1}^{G} \Bigg(\sum_{h_2=1}^{M_g} \zeta_{gh_2} \Bigg)^2 \Bigg). \qquad \text{(A.48)}\end{aligned}$$

As in (A.41), we find that the penultimate factor on the right-hand side of (A.48) is bounded by a constant times $\sum_{g=1}^{G}(\sum_{h=1}^{M_g} \sigma_{gh}^2)^2$ on a set $\mathcal{A}$ with $P(\mathcal{A}) \to 1$. The last factor on the right-hand side of (A.48) is a non-negative random variable and hence is of order $O_P(\mathrm{E}\sum_{g=1}^{G}(\sum_{h=1}^{M_g} \zeta_{gh})^2) = O_P(\sum_{g=1}^{G} \mathrm{E}\zeta_g^2) = O_P(N \sup_g N_g)$ by Lemma A.1. Combining these results and using (A.29) and (A.30), we find that

$$\text{(A.45)} = O_P\Bigg(\sup_g N_g \sup_{g,h} N_{gh} \Bigg(\sum_{g=1}^{G} \Bigg(\sum_{h=1}^{M_g} \sigma_{gh}^2 \Bigg)^2 \Bigg)^{1/2} \Bigg),$$

which proves the required result for (A.45), and hence for (A.44) and (A.38).

To prove the second result of the lemma we write the left-hand side as

$$\begin{aligned}&2\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2=1}^{M_g} \boldsymbol{b}_1^\top \boldsymbol{\Sigma}_{gh_1} \boldsymbol{b}_1 \boldsymbol{b}_2^\top \boldsymbol{\Sigma}_{gh_2} \boldsymbol{b}_2 - 2\sum_{g=1}^{G}\sum_{h=1}^{M_g} \boldsymbol{b}_1^\top \boldsymbol{\Sigma}_{gh} \boldsymbol{b}_1 \boldsymbol{b}_2^\top \boldsymbol{\Sigma}_{gh} \boldsymbol{b}_2 \\ &\qquad = 2\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2=1}^{M_g} \boldsymbol{b}_1^\top \boldsymbol{\Sigma}_{gh_1} \boldsymbol{b}_1 \boldsymbol{b}_2^\top \boldsymbol{\Sigma}_{gh_2} \boldsymbol{b}_2 \left(1 - \frac{\sum_{h=1}^{M_g} \boldsymbol{b}_1^\top \boldsymbol{\Sigma}_{gh} \boldsymbol{b}_1 \boldsymbol{b}_2^\top \boldsymbol{\Sigma}_{gh} \boldsymbol{b}_2}{\sum_{h_1=1}^{M_g}\sum_{h_2=1}^{M_g} \boldsymbol{b}_1^\top \boldsymbol{\Sigma}_{gh_1} \boldsymbol{b}_1 \boldsymbol{b}_2^\top \boldsymbol{\Sigma}_{gh_2} \boldsymbol{b}_2} \right) \\ &\qquad \geq 2\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2=1}^{M_g} \boldsymbol{b}_1^\top \boldsymbol{\Sigma}_{gh_1} \boldsymbol{b}_1 \boldsymbol{b}_2^\top \boldsymbol{\Sigma}_{gh_2} \boldsymbol{b}_2 \left(1 - \frac{\sup_h \boldsymbol{b}_2^\top \boldsymbol{\Sigma}_{gh} \boldsymbol{b}_2}{\sum_{h=1}^{M_g} \boldsymbol{b}_2^\top \boldsymbol{\Sigma}_{gh} \boldsymbol{b}_2} \right),\end{aligned}$$

where the inequality is due to $\sum_{h=1}^{M_g} \boldsymbol{b}_1^\top \boldsymbol{\Sigma}_{gh} \boldsymbol{b}_1 \boldsymbol{b}_2^\top \boldsymbol{\Sigma}_{gh} \boldsymbol{b}_2 \leq (\sup_h \boldsymbol{b}_2^\top \boldsymbol{\Sigma}_{gh} \boldsymbol{b}_2) \sum_{h=1}^{M_g} \boldsymbol{b}_1^\top \boldsymbol{\Sigma}_{gh} \boldsymbol{b}_1$. The result follows because $\sup_{g,h} \boldsymbol{b}_2^\top \boldsymbol{\Sigma}_{gh} \boldsymbol{b}_2 / (\boldsymbol{b}_2^\top \sum_{h=1}^{M_g} \boldsymbol{\Sigma}_{gh} \boldsymbol{b}_2) \leq \sup_{g,h} \omega_{\max}(\boldsymbol{\Sigma}_{gh}(\sum_{h=1}^{M_g} \boldsymbol{\Sigma}_{gh})^{-1}) < 1$ by Assumption 4. □

Appendix B: The IM Test

A few simulation results for the IM test of Ibragimov and Müller (2016) were presented in Section 5.4. In this appendix, we describe that test. Let G denote the number of coarse clusters for which it is possible to estimate β, the coefficient of interest, on a cluster-by-cluster basis. Note that this may well be smaller than the original number of coarse clusters. Since the $\hat{\beta}_g$ are estimated separately for every coarse cluster, there are effectively coarse-level

fixed effects. Suppose that cluster-by-cluster estimation yields estimates $\hat{\beta}_g$ for $g = 1, \ldots, G$. Then a natural estimate of the variance of the average of the cluster-by-cluster estimates is

$$V_\beta = \frac{1}{G-1} \sum_{g=1}^{G} (\hat{\beta}_g - \bar{\beta})^2, \quad \text{where } \bar{\beta} = \frac{1}{G} \sum_{g=1}^{G} \hat{\beta}_g. \tag{B.1}$$

This is the usual estimate of the variance of a sample average based on G observations.

When there is either no clustering or clustering at a finer level, the variances of the $\hat{\beta}_g$ can be estimated using either an HCCME or a CRVE for each coarse cluster separately. If we suppose for simplicity that all of the k regressors except the one of interest have been partialed out, leaving just one regressor, then the CRVE estimate based on fine clustering is

$$\hat{\sigma}^2_{g\text{f}} = \frac{M_g}{M_g - 1} \frac{N_g - 1}{N_g - k} (\boldsymbol{x}_g^\top \boldsymbol{x}_g)^{-2} \sum_{h=1}^{M_g} \boldsymbol{x}_{gh}^\top \hat{\boldsymbol{u}}_{gh} \hat{\boldsymbol{u}}_{gh}^\top \boldsymbol{x}_{gh}, \tag{B.2}$$

where $\boldsymbol{x}_g$ is the vector of observations on the regressor for coarse cluster g, $\boldsymbol{x}_{gh}$ is the subvector of $\boldsymbol{x}_g$ for fine cluster h, and $\hat{\boldsymbol{u}}_{gh}$ is the corresponding subvector of the residual vector $\hat{\boldsymbol{u}}_g$. Similarly, the heteroskedasticity-robust estimate would be

$$\hat{\sigma}^2_{g\,\text{het}} = \frac{N_g}{N_g - k} (\boldsymbol{x}_g^\top \boldsymbol{x}_g)^{-2} \sum_{i=1}^{N_g} x_{ghi}^2 \hat{u}_{ghi}^2, \tag{B.3}$$

where x_{ghi} and $\hat{u}_{ghi}$ are elements of $\boldsymbol{x}_{gh}$ and $\hat{\boldsymbol{u}}_{gh}$, respectively. Of course, in practice, $\hat{\sigma}^2_{g\text{f}}$ and $\hat{\sigma}^2_{g\,\text{het}}$ would usually be obtained by taking the appropriate diagonal elements of either a CRVE or an HCCME for the model estimated using data for coarse cluster g.

Let $\hat{\sigma}_g$ denote the square root of either $\hat{\sigma}^2_{g\,\text{het}}$ or $\hat{\sigma}^2_{g\text{f}}$, depending on whether the null hypothesis is no clustering or fine clustering. Then generate a large number of realizations, say S, of G independent random variates z_{gs} that follow the standard normal distribution. For each simulation, compute

$$Y_{gs} = \hat{\sigma}_g z_{gs}, \; g = 1, \ldots, G, \quad \bar{Y}_s = \frac{1}{G} \sum_{g=1}^{G} Y_{gs}, \quad \text{and} \quad V_s = \frac{1}{G-1} \sum_{g=1}^{G} (Y_{gs} - \bar{Y}_s)^2. \tag{B.4}$$

Notice that V_s is one realization of a random variable that might reasonably be expected to have approximately the same distribution as the variance of $\hat{\beta}$ when there is either no clustering or fine clustering.

The IM test is based on comparing V_β from (B.1) with the empirical distribution of the V_s from (B.4). The P value for a one-sided test is simply

$$P_{\text{IM}} = \frac{1}{S} \sum_{s=1}^{S} \mathbb{I}(V_s > V_\beta). \tag{B.5}$$

Thus the test will reject whenever V_β, the direct estimate of the variance of $\hat{\beta}$ from the cluster-by-cluster estimates, is larger than most of the realizations of V_s, which depend on the estimates $\hat{\sigma}_g$ that assume either fine clustering or no clustering. Although the IM test requires simulation to calculate the V_s, it is not expensive to compute, because the model is not re-estimated during the simulations.

There are two serious practical limitations of the IM test. The first is that, in many cases, it is not possible to compute the $\hat{\beta}_g$ for every coarse cluster. The second is that, even when they can all be computed, the full model may not be estimable for each coarse cluster, so the interpretation of the $\hat{\beta}_g$ may differ across clusters. For instance, when including fixed effects for categorical variables, not all types may be found within each coarse cluster. Both these problems are encountered in the empirical example of Section 6.

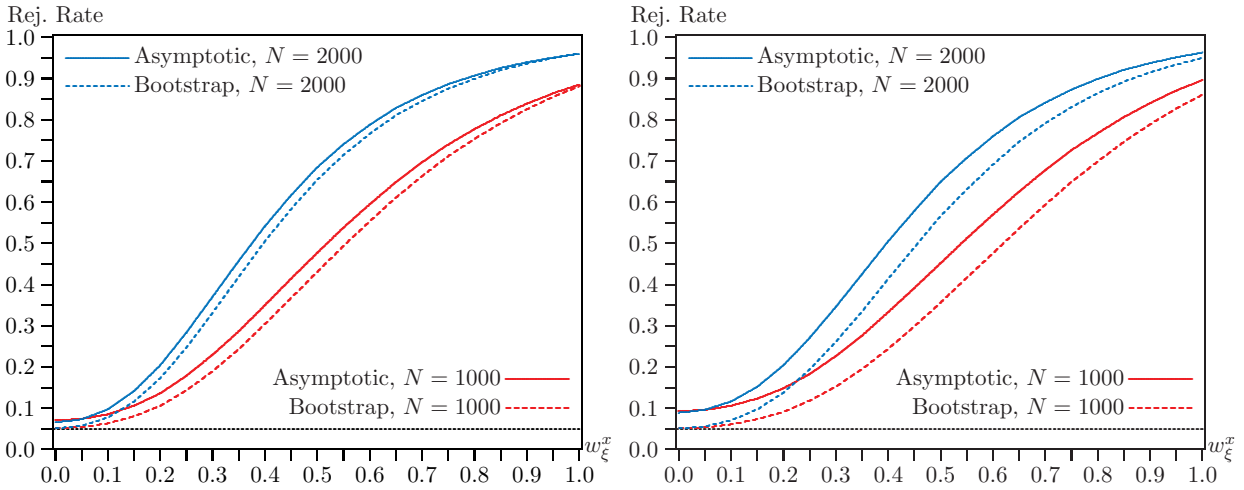
Appendix C: Additional Simulations

In this appendix, we present the results of a number of additional simulation experiments. In Appendix C.1, we compare the power of bootstrap and asymptotic score-variance tests. The former appear to have less power only because the latter over-reject under the null hypothesis. Appendices C.2 and C.3 deal with rejection frequencies under the null for asymptotic, bootstrap, and IM tests. In the former subsection, we study the effects of heteroskedasticity. In the latter, we study the effects of cluster sizes that vary, sometimes to an extreme extent. In such cases, the performance of the three tests can differ markedly. Next, we perform a few simulation experiments in Appendix C.4 that use the actual clusters and regressors for the STAR example, in order to explain some of the empirical results in Section 6. Finally, in Appendix C.5 we perform some experiments where the score-variance test is used as a pre-test to select the level of clustering prior to conducting inference on regression coefficients.

C.1 Power of Asymptotic and Bootstrap Score-Variance Tests

In Figures 3 and 4, we reported simulation results for the power of the bootstrap versions of our score-variance tests. In this section, we perform some additional experiments in which we study the power of both the asymptotic and bootstrap versions. We also change the experimental design. The DGP for the disturbances is now the same in all experiments (see the notes to Figure C.1), but the DGP for the regressor of interest varies. For this regressor, the weight on the random factors in (36), which we denote w_ξ^x, is allowed to vary between 0 and 1. When it is 0, the null hypothesis is true, because the scores are uncorrelated within clusters, even though the disturbances are correlated. As w_ξ^x increases, the scores become

Figure C.1: Power of two-sided τ_σ tests when there is coarse clustering



(a) No clustering vs. coarse

(b) Fine vs. coarse

Notes: The data are generated by (38). There are 5 coarse clusters, 20 fine clusters, fine-level fixed effects, and either 1000 or 2000 observations. The disturbances have coarse clustering, with $\rho = 0.5$ and weight $w_\xi = 0.7$. The regressor of interest also has coarse clustering, with $\rho^x = 0.5$ and w_ξ^x varying. There are 400,000 replications and 999 bootstraps.

more correlated, and the power of the tests increases.

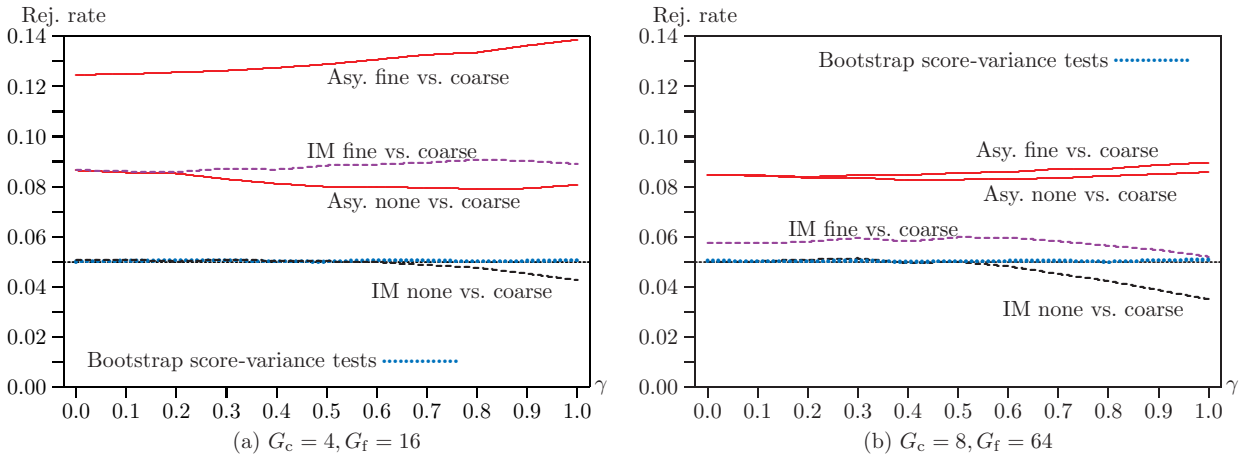
Figure C.1 shows power functions for asymptotic and bootstrap tests of two hypotheses for two sample sizes. Because there are only five coarse clusters, the bootstrap tests perform very much better under the null and consequently appear to have less power. This is particularly noticeable in Panel (b), where the asymptotic tests of fine against coarse clustering over-reject rather severely. Of course, the additional power of the asymptotic tests is entirely spurious; see Davidson and MacKinnon (2006).

Not surprisingly, all the tests have more power when $N = 2000$ than when $N = 1000$ because, under the current experimental design, additional observations contribute information even within clusters. The gaps between the bootstrap and asymptotic power functions are also a bit smaller for the larger sample size. However, based on the theory of Sections 4.1 and 4.2, the bootstrap and asymptotic power functions do not necessarily coincide as $N \to \infty$ unless the numbers of coarse and fine clusters also tend to infinity.

C.2 Heteroskedasticity

Although Theorems 1 and 3 explicitly allow for heteroskedasticity of unknown form, all the simulations up to this point have involved disturbances that are homoskedastic conditional on the regressors. One might reasonably worry that the finite-sample performance of our

Figure C.2: Rejection frequencies for one-sided tests with heteroskedasticity



Notes: The data are generated by (C.1). There are 6400 observations and either 4 or 8 coarse clusters. The regressors have coarse clustering, with $\rho^x = 0.5$ and $w_\xi^x = 0.7$. There are 400,000 replications and 399 bootstraps. IM tests use 999 simulations.

score-variance tests may depend on the extent and pattern of heteroskedasticity. In this section, we provide evidence that this seems to be the case only to a very limited extent.

In these experiments, the data are generated using a modified version of (38),

$$y_{ghi} = \beta_0 + \beta_1 X_{1ghi} + \beta_2 X_{2ghi} + \boldsymbol{D}_{gh}\boldsymbol{\delta} + \exp\Big(\gamma(\beta_0 + \beta_1 X_{1ghi} + \beta_2 X_{2ghi})\Big)u_{ghi}, \tag{C.1}$$

where the regressors are X_{jghi}, for $j = 1, 2$, $\boldsymbol{D}_{gh}$ are fixed effects at the fine cluster level, and the disturbances u_{ghi} are generated as before. In our experiments, we set $\beta_0 = \beta_1 = \beta_2 = 1$. The values of these parameters did not matter previously. We also set $\boldsymbol{\delta} = \mathbf{0}$, so that the DGP in (C.1) does not include cluster fixed effects, but the model that is actually estimated does include them at the fine level. The regressors are generated with the same pattern of coarse clustering as in Figure 1, and the disturbances are independent. When $\gamma = 0$, the DGP in (C.1) reduces to a special case of (38) with no fixed effects. As γ increases, the heteroskedasticity becomes stronger.

Figure C.2 shows rejection frequencies for several tests as functions of γ, which varies between 0 (homoskedasticity) and 1 (substantial heteroskedasticity). In Panel (a), the numbers of coarse and fine clusters are very small, at just 4 and 16. In Panel (b), they are somewhat larger, at 8 and 64. For comparability with the IM test, and also for readability, the figure reports results for one-sided tests. Two-sided asymptotic tests perform considerably better than one-sided ones, but their dependence on γ is similar. In Panel (b), the former perform almost as well as the IM test for fine against coarse clustering.

Figure C.2 shows that rejection frequencies for asymptotic tests depend on γ, but the dependence is quite modest. In Panel (b), where there are 64 fine clusters, both of the IM tests show considerably greater dependence on γ than do the asymptotic score-variance tests. In contrast, the bootstrap score-variance tests seem to perform perfectly in all cases. At least in these experiments, their excellent performance seems to be completely invariant to γ, G_{c}, and G_{f}.

C.3 Varying Cluster Sizes

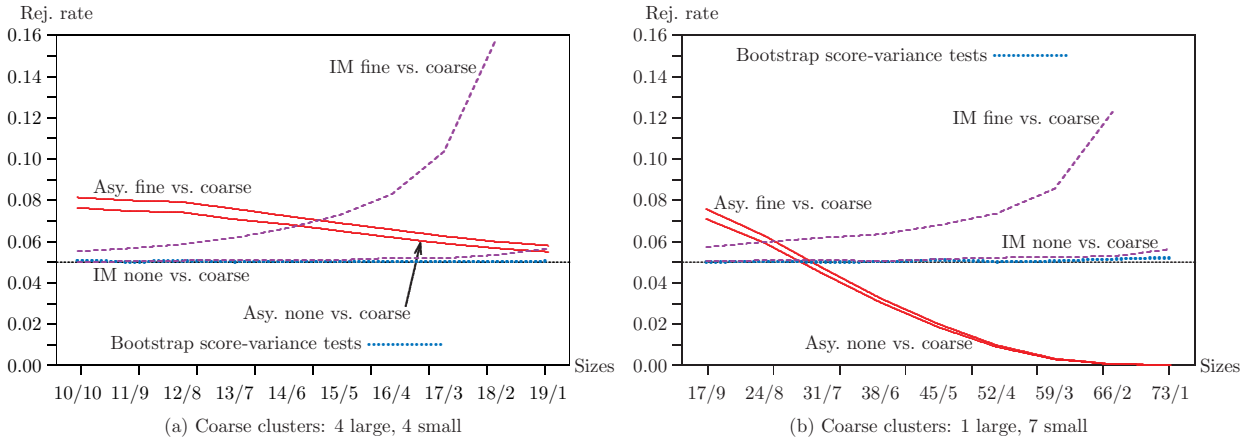
The assumptions of Theorems 1 and 3 explicitly rule out too much variability in cluster sizes. However, it is not clear just how much variability can be tolerated in finite samples. In this section, we study the performance of our score-variance tests and of the IM test when cluster sizes are not constant by allowing the number of fine clusters per coarse cluster to vary.

In our experiments, there are always 80 fine clusters, each of which contains 50 observations, so that there are 4000 observations in total. There are 8 coarse clusters, with the number of fine clusters per coarse cluster allowed to vary. The number of ways in which 80 fine clusters can be allocated among 8 coarse clusters is very large, and so we restrict attention to two special cases. Our score-variance tests can be computed even when some of the coarse clusters contain just one fine cluster. This may be important in practice, because, for example, some countries may be divided into states or regions and others may not. In contrast, IM tests of fine versus coarse clustering are impossible to compute in this case, because doing so would require the computation of a fine-cluster CRVE for each coarse cluster, which is impossible when there is just one fine cluster.

Figure C.3 shows rejection frequencies for one-sided tests with coarse-level fixed effects. Only results for tests of no clustering against coarse clustering and fine against coarse clustering are shown, because the way in which fine clusters are allocated among coarse clusters has no impact on tests of no clustering against fine clustering. In Panel (a), there are always four large coarse clusters and four small ones. The number of fine clusters per large coarse cluster ranges from 10 to 19, so that the number of observations per large coarse cluster varies from 500 to 950. The number of fine clusters per small coarse cluster therefore ranges from 1 to 10, and the corresponding number of observations from 50 to 500. In Panel (b), there is just one large cluster and seven small clusters. The size of the large cluster varies from 850 to 3650 out of the 4000 total observations.

In Panel (a), the asymptotic score-variance tests over-reject moderately when all clusters are the same size, but the over-rejection becomes less serious as the large and small clusters differ more in size. The bootstrap score-variance tests always work perfectly, even when

Figure C.3: Rejection frequencies for one-sided tests with varying coarse cluster sizes



(a) Coarse clusters: 4 large, 4 small

(b) Coarse clusters: 1 large, 7 small

Notes: The data are generated by (38). There are 4000 observations and 80 fine clusters. The regressors have coarse clustering, with $\rho^x = 0.5$ and $w_\xi^x = 0.7$. There are 400,000 replications and 399 bootstraps. IM tests use 999 simulations.

the small clusters contain just one fine cluster. Note that the figure actually shows results for both bootstrap tests (none vs. coarse and fine vs. coarse), but they are impossible to distinguish. The IM test of no clustering generally works very well, but it over-rejects a little as the variability of cluster sizes becomes more extreme. In contrast, the IM test of fine clustering always over-rejects, and it does so very severely in the right-hand side of the figure. The curve for this test ends early, because the test cannot be computed when each coarse cluster contains just one fine cluster.

When the single large cluster is not too large, all the tests work about as well in Panel (b) as in Panel (a). Even when it is very large, the IM tests actually work somewhat better in Panel (b) than in Panel (a), although the patterns in both panels are similar. In contrast, the asymptotic score-variance tests under-reject very severely when the large cluster is very large. In fact, they never reject at all in the most extreme case, when the large cluster contains 91.25% of the observations. Amazingly, the bootstrap score-variance tests always perform extremely well. In the very worst case, for no clustering against coarse clustering, the bootstrap test rejects 5.21% of the time.

Based on these results, we tentatively conclude that bootstrap score-variance tests can be used safely even when cluster sizes vary enormously. In contrast, asymptotic score-variance tests seem to be much more sensitive to variation in cluster sizes. However, having few clusters (as in Panel (a) of Figure C.2) seems to be more harmful to the asymptotic test than having cluster sizes that vary within reason. IM tests of no clustering against coarse clustering also work well even when cluster sizes vary extremely, but IM tests of fine against

Table C.1: Rejection Rates for Monte Carlo Experiments Using STAR Data

		Without School FE		With School FE	
		Asymptotic	Bootstrap	Asymptotic	Bootstrap
No clustering in DGP:					
small	H_N vs H_R	5.11	5.02	63.62	5.13
	H_N vs H_S	4.75	4.98	5.52	4.98
	H_R vs H_S	5.39	5.01	48.97	8.11
aide	H_N vs H_R	5.15	5.00	70.71	5.12
	H_N vs H_S	4.80	5.01	5.67	5.02
	H_R vs H_S	5.38	4.98	48.81	7.31
both	H_N vs H_R	4.82	4.99	66.77	5.09
	H_N vs H_S	5.05	5.02	6.51	5.04
	H_R vs H_S	5.74	5.03	68.81	9.00
Classroom clustering in DGP:					
small	H_R vs H_S	5.39	4.98	53.57	9.07
aide	H_R vs H_S	5.48	5.05	51.43	7.65
both	H_R vs H_S	5.80	5.06	73.06	9.95

Notes: There are 3,989 observations and either 330 (classroom) or 75 (school) clusters. All numbers are either asymptotic or bootstrap rejection rates for tests at the 0.05 level, expressed as percentages. For the τ_σ tests, denoted "small" and "aide," asymptotic P values are two-sided and based on the $N(0,1)$ distribution. For the τ_Σ tests, denoted "both," they are based on the $\chi^2(3)$ distribution. Bootstrap tests use $B = 399$, and there are 400,000 replications.

coarse clustering tend to over-reject seriously in such cases.

C.4 Simulations Using the STAR Dataset

Some of the empirical results for the STAR model in Section 6 suggest that score-variance tests can be unreliable when key regressors vary at the fine-cluster level and there are fixed effects at the coarse level. In order to investigate this phenomenon, we perform some additional Monte Carlo experiments. All of the regressors and the two clustering structures (by classroom and school) are identical to the ones in the empirical example. The only thing that varies across replications is the disturbance term, to which we add the fitted value to generate the dependent variable.

In the first set of experiments, the disturbance term is a vector of 3989 independent standard normal random variates. Thus the hypothesis of no clustering (H_N) is correct. Rejection rates are shown in the first part of Table C.1. When there are no school fixed effects, all the asymptotic tests work very well, and all the bootstrap tests work extremely

well. However, when school fixed effects are included in the regression, the asymptotic tests for H_N against H_R and for H_R against H_S over-reject severely. The bootstrap tests, on the other hand, work very much better than the asymptotic ones. For the tests of no clustering, they work essentially perfectly. However, for the tests of H_R against H_S, even the bootstrap tests over-reject somewhat. In the worst case, for the test of both coefficients jointly, they reject 9% of the time.

In the second set of experiments, the DGP has classroom-level (fine) clustering. The disturbances are generated by a random-effects model at the classroom level (recall that the fixed effects are at the school level), with parameters chosen so that the average correlation within each classroom is 0.20. This is actually a bit smaller than the average correlation estimated directly from the residuals for the model (39), which is 0.23.[4] Not surprisingly, the tests of H_N against both forms of clustering reject essentially all the time, so we do not report results for them. The last part of Table C.1 contains rejection rates for the tests of H_R against H_S, for which the null hypothesis is true. For the model with fixed effects, the rejection rates in this part are always somewhat higher than the corresponding ones for the first experiment. However, the patterns across the various tests are very similar.

It is not hard to see why some of the tests perform poorly when there are school fixed effects. Because of the fixed effects, the residuals for each school must sum to zero. With only 4.4 classrooms per school, on average, this creates substantial negative correlation for the residuals within each classroom. This negative correlation apparently interacts with the aide and small regressors, after they have been projected off the fixed effects and the other regressors, to create negative intra-classroom correlation in the empirical scores. Consequently, the τ_σ statistics for H_N against H_R have means that are large and negative (-2.23 for small and -2.41 for aide). The same phenomenon causes the τ_σ statistics for H_R against H_S to have means that are large and positive (1.95 for both small and aide). This accounts for the over-rejection by the asymptotic tests.

As can be seen from the last column of Table C.1, the ordinary wild bootstrap for tests of H_N against H_R does an excellent job of compensating for the poorly-centered distribution of the asymptotic test statistics. However, the wild cluster bootstrap for the tests of H_R against H_S does not work nearly as well. The problem is apparently that, with an average of only 4.4 classrooms per school, the wild cluster bootstrap has difficulty mimicking the pattern of within-cluster correlations of the empirical scores.

These arguments suggest that the tests which do not perform well in Table C.1 would perform better if the schools were larger, i.e. had more classrooms. We investigated this

[4]This estimate is for the model with just a constant term. For the model with fixed effects, the estimate is much lower (0.078), but it is surely biased downwards.

conjecture by making the schools about twice as large, on average. All of the rejection rates that were not already near 5% improved dramatically. For example, the asymptotic rejection rate for the τ_Σ tests of H_R against H_S dropped from 68.81% in the table to 17.68%, and the bootstrap rejection rate dropped from 9.00% to 5.43%. Having more classrooms per school reduces the intra-cluster correlations of the empirical scores very substantially. This makes the asymptotic tests perform better. It also allows the wild cluster bootstrap to mimic those correlations more accurately, which improves the performance of the bootstrap tests.

C.5 Making Inferences about a Regression Coefficient

In Section 3.4, we briefly discussed pre-test procedures for making inferences about a single regression coefficient when clustering may be either fine or coarse. In this subsection, we investigate these procedures by means of simulation experiments. For simplicity, and because it is a common case, the level of fine clustering is no clustering at all, so that the corresponding variance matrix is robust only to heteroskedasticity of unknown form.

The model is a variant of (38), with two regressors plus cluster fixed effects, so that $K = G + 2$. The values of ω_ξ and ρ for the regressors are 0.8 and 0.5. The disturbances are generated by

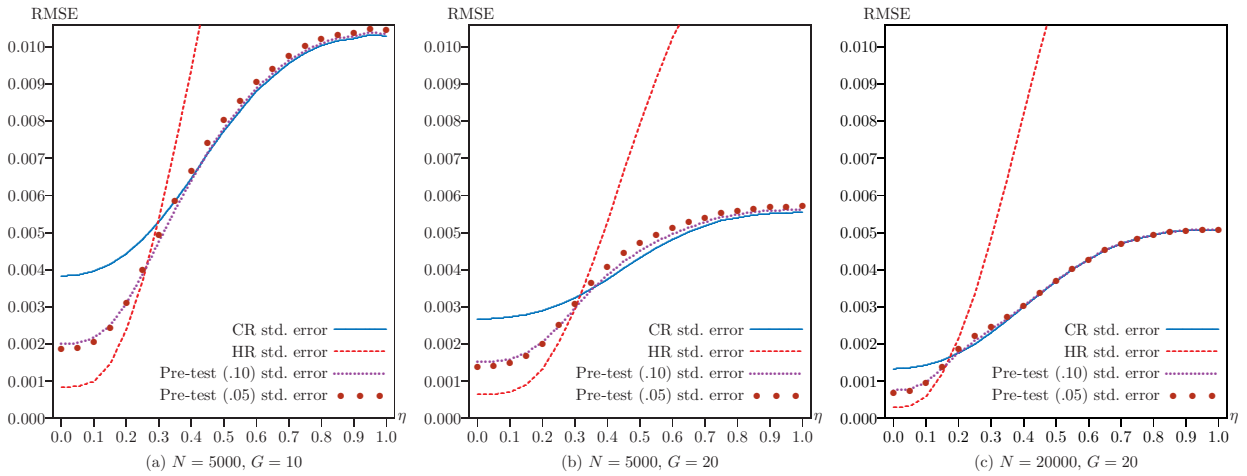
$$u_{gi} = (\eta^2 + (1-\eta)^2)^{-1/2}\Big((1-\eta)\epsilon_{gi} + \eta\epsilon^c_{gi}\Big), \tag{C.2}$$

where the ϵ_{gi} are i.i.d. standard normal and the ϵ^c_{gi} are generated in the same way as the regressors, but with $\omega_\xi = 0.4$, rescaled to have variance 1. The parameter η determines the amount of intra-cluster correlation. When $\eta = 0$, the disturbances are not clustered. The initial scaling factor ensures that $\mathrm{Var}(u_{gi}) = 1$ for all η.

In the experiments, we vary η from 0 to 1. When $\eta = 0$, the correct variance estimator to use is the heteroskedasticity-robust (HR) one with $\hat{\boldsymbol{V}} = \hat{\boldsymbol{V}}_{\text{het}}$ in (5) and (7), multiplied as usual by the factor $N/(N-K)$. For any other value of η, there is intra-cluster correlation, so that we should use the cluster-robust (CR) one with $\hat{\boldsymbol{V}} = \hat{\boldsymbol{V}}_{\text{c}}$; see (5) and (6). Of course, the investigator is assumed not to know η, or indeed (C.2), and it may therefore be attractive to employ a pre-test estimator.

For two reasons, the two pre-test estimators we study are based on one-sided tests. The first reason is that one-sided tests are more powerful than two-sided tests when there actually is clustering, so that the former make fewer Type II errors. We did obtain results for two-sided pre-test estimators, but for larger values of η they were clearly inferior to the one-sided ones that we report. The second reason is that, even when the difference between $\mathrm{Var}_{\text{c}}(\hat{\beta}_1)$ and $\mathrm{Var}_{\text{het}}(\hat{\beta}_1)$, the variances based on the true variance matrices of the sums of scores, is positive, it is quite possible for the variance estimate $\widehat{\mathrm{Var}}_{\text{c}}(\hat{\beta}_1)$ to be smaller than $\widehat{\mathrm{Var}}_{\text{het}}(\hat{\beta}_1)$.

Figure C.4: Root mean squared errors of various standard error estimates



(a) $N = 5000$, $G = 10$ (b) $N = 5000$, $G = 20$ (c) $N = 20000$, $G = 20$

Notes: The regressors are generated by (38) with coarse clustering, and the disturbances are generated by (C.2). There is no clustering when $\eta = 0$. There are 5000 or 20,000 observations and 10 or 20 clusters. The pre-test estimators are based on one-sided tests. There are 400,000 replications. The "true" standard errors relative to which the RMSEs are computed are based on 400,000 estimates of $\hat{\beta}_1$.

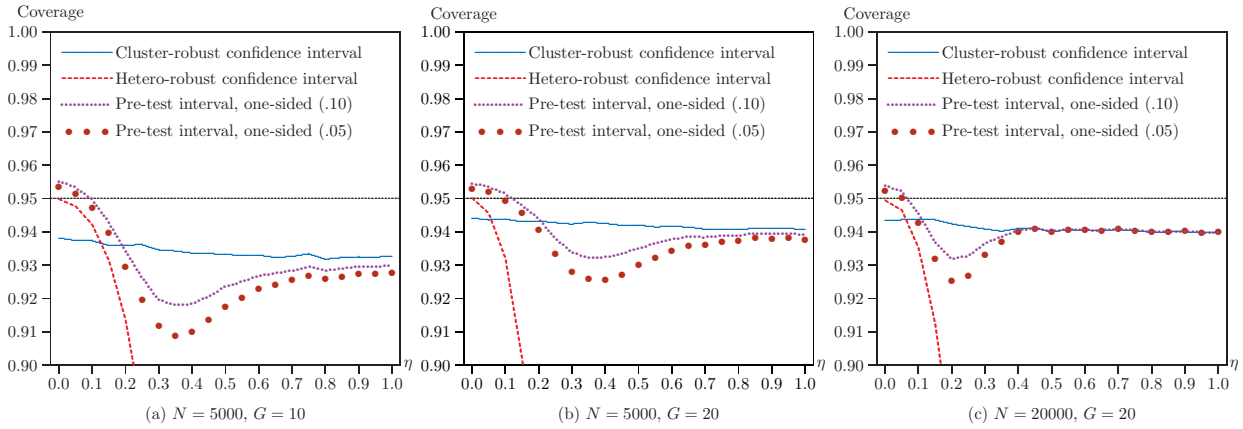
Indeed, this happens frequently in our experiments when η is small. Thus we believe that investigators will rarely wish to reject HR in favor of CR when the CR standard error is smaller than the HR one.

It is natural to think of the choice among the HR standard error of $\hat{\beta}_1$, the CR standard error, or a pre-test estimator of the standard error, as an estimation problem. Thus, it seems reasonable to compare them on the basis of root mean squared error (RMSE). Figure C.4 shows the RMSEs of HR, CR, and two pre-test estimators of the standard error of $\hat{\beta}_1$, the first coefficient in (38). These are based on experiments with 400,000 replications for two values of G (10 and 20) and two values of N (5000 and 20,000).

The results in Figure C.4 are striking. In all three panels, the RMSE of the HR standard error is the smallest of the four for $\eta \approx 0$, while the RMSE of the CR standard error is the largest. However, the HR RMSE rises very sharply for η greater than about 0.1 or 0.2, and it rapidly becomes so large than it cannot be plotted on the same axes as the other RMSEs. Note that it does have an "S" shape, like the other estimators, although this cannot be seen in the figure. The CR RMSE also increases with η, but to a much more moderate extent.

Both bias and variance contribute to the RMSEs of all the estimators. For HR, the variance is always fairly small, and the bias is zero when $\eta = 0$, but the bias becomes large and negative for moderate to large values of η, accounting for most of the RMSE. In contrast, the variance of CR is always much larger than the variance of HR, and the bias is also larger for small values of η. But, although both bias and variance increase with η, it is always the

Figure C.5: Coverage of various confidence intervals



(a) $N = 5000$, $G = 10$ (b) $N = 5000$, $G = 20$ (c) $N = 20000$, $G = 20$

Notes: The regressors are generated by (38) with coarse clustering, and the disturbances are generated by (C.2), so that there is no clustering when $\eta = 0$. There are 2 regressors plus G fixed effects. There are 5000 or 20,000 observations and 10 or 20 clusters. The pre-test estimators are based on one-sided tests. There are 400,000 replications.

latter that dominates, and the RMSE increases much more slowly than that of HR.

The two pre-test estimators perform as we might expect them to. For small values of η, their RMSEs are much smaller than the CR ones, but substantially larger than the HR ones, because of Type I error. In this case, the pre-test at the .05 level performs better than the pre-test at the .10 level, because it necessarily makes fewer Type I errors. In contrast, for large values of η, the pre-test standard errors perform very slightly worse than the CR ones, but very much better than the HR ones. Since this is caused by Type II error, the pre-test at the .05 level inevitably performs slightly worse than the pre-test at the .10 level.

The extent to which the pre-test standard errors have larger RMSEs than the HR ones for small values of η seems to be about the same in the three panels, but the range of values of η over which this occurs is considerably smaller in Panel (c) than in the other two panels because N is four times as large. In contrast, the extent to which the pre-test standard errors have slightly larger RMSEs than the CR ones for large values of η varies across the three panels. It is greatest in Panel (b), where the score-variance tests evidently have less power for $G = 20$ and $N = 5000$ than they do for either a smaller value of G, in Panel (a), or a larger value of N, in Panel (c). In the latter case, the CR and pre-test standard error estimators are effectively identical for $\eta \geq 0.4$, presumably because the tests reject almost all the time.

Figure C.5 shows the coverage of confidence intervals based on the four standard errors (HR, CR, and pre-test at two levels) and the same cases studied in Figure C.4. The CR intervals always under-cover somewhat, especially in Panel (a) where $G = 10$, because the CR standard errors are biased downwards. The under-coverage, like the bias, becomes slightly

worse as η increases. Coverage would have been closer to 95% if we had used the wild cluster bootstrap, but that would have been computationally very demanding to simulate. On the other hand, the coverage of the HR intervals is almost exactly 95% when $\eta = 0$, but they always under-cover for $\eta > 0$, and the under-coverage is very severe for most values of η.

The pre-test intervals over-cover very slightly when $\eta = 0$, which is a consequence of Type I errors in the pre-tests. However, they under-cover slightly more than the CR intervals for many values of η because of Type II errors. This is exactly what we would expect to see in view of the results in Figure C.4. The results in Figure C.5 suggest that, as is often the case for pre-testing, it may be desirable for the level of the pre-test to be quite high, perhaps even higher than .10.

By construction, the pre-test intervals are never shorter than the HR intervals, and the CR intervals are longer than the HR intervals on average. However, with small values of η, it is quite common to encounter samples for which the CR interval is shorter than the HR and pre-test intervals. This means that CR intervals can be misleadingly short, especially when G is small. However, that is not the reason for the under-coverage by CR intervals that is evident in Figure C.5, which is caused by the downward bias of the CR standard errors. The wild cluster bootstrap could undoubtedly be used to obtain more reliable CR intervals (Djogbenou et al. 2019; Roodman et al. 2019). It could almost certainly also be used to obtain more reliable pre-test intervals, and this will be the subject of future work.

References

Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge (2017). When should you adjust standard errors for clustering? Working Paper 24003, National Bureau of Economic Research.

Athey, S. and G. W. Imbens (2017). The econometrics of randomized experiments. In A. V. Banerjee and E. Duflo (Eds.), *Handbook of Economic Field Experiments*, pp. 73–140. Elsevier.

Berger, R. L. and D. F. Sinclair (1984). Testing hypotheses concerning unions of linear subspaces. *Journal of the American Statistical Association 79*, 158–163.

Bertrand, M., E. Duflo, and S. Mullainathan (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics 119*, 249–275.

Brown, B. M. (1971). Martingale central limit theorems. *Annals of Mathematical Statistics 42*, 59–66.

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.

Conley, T. G., S. Gonçalves, and C. B. Hansen (2018). Inference with dependent data in accounting and finance applications. *Journal of Accounting Research 56*, 1139–1203.

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

Davidson, R. and J. G. MacKinnon (1992). A new form of the information matrix test. *Econometrica 60*, 145–157.

Davidson, R. and J. G. MacKinnon (2000). Bootstrap tests: How many bootstraps? *Econometric Reviews 19*, 55–68.

Davidson, R. and J. G. MacKinnon (2006). The power of bootstrap and asymptotic tests. *Journal of Econometrics 133*, 421–441.

de Chaisemartin, C. and J. Ramirez-Cuellar (2020). At what level should one cluster standard errors in paired experiments, and in stratified experiments with small strata? Working Paper 27609, National Bureau of Economic Research.

Djogbenou, A. A., J. G. MacKinnon, and M. Ø. Nielsen (2019). Asymptotic theory and wild bootstrap inference with clustered errors. *Journal of Econometrics 212*, 393–412.

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

Finn, J. D. and C. M. Achilles (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal 27*, 557–577.

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

Harville, D. A. (1997). *Matrix Algebra From a Statistician's Perspective*. New York: Springer-Verlag.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica 46*, 1251–1272.

Ibragimov, R. and U. K. Müller (2010). t-statistic based correlation and heterogeneity robust inference. *Journal of Business & Economic Statistics 28*, 453–468.

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

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

Krueger, A. B. (1999). Experimental estimates of education production functions. *Quarterly*

Journal of Economics 114, 497–532.

Leeb, H. and B. M. Pötscher (2005). Model selection and inference: Facts and fiction. *Econometric Theory 21*, 21–59.

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

MacKinnon, J. G. (2013). Thirty years of heteroskedasticity-robust inference. In X. Chen and N. R. Swanson (Eds.), *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, pp. 437–461. Springer.

MacKinnon, J. G. (2019). How cluster-robust inference is changing applied econometrics. *Canadian Journal of Economics 52*, 851–881.

MacKinnon, J. G., M. O. . Nielsen, and M. D. Webb (2020). Cluster-robust inference: A guide to empirical practice. Qed working paper, Queen's University.

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

MacKinnon, J. G. and M. D. Webb (2018). The wild bootstrap for few (treated) clusters. *Econometrics Journal 21*, 114–135.

MacKinnon, J. G. and M. D. Webb (2020). When and how to deal with clustered errors in regression models. QED Working Paper 1421, Queen's University.

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The Future of Children 5*, 113–127.

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

Racine, J. S. and J. G. MacKinnon (2007). Simulation-based tests that can use any number of simulations. *Communications in Statistics–Simulation and Computation 36*, 357–365.

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.

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

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

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*, 1–25.

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