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
Aarhus University

Matthew D. Webb
Carleton University

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

12-2022 (major revisions)
3-2023 (minor revisions)

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

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

Morten Ørregaard Nielsen
Aarhus University
mon@econ.au.dk

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

March 11, 2023

**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 the editor, Serena Ng, an anonymous associate editor, three anonymous referees, Yevgeniy Feyman, and participants at the 2018 Canadian Economics Association conference, 2018 Canadian Econometric Study Group conference, 2020 Econometric Society World Congress, 2021 WEA conference, 2021 APPAM conference, 2021 IAAE Annual Conference, Université de Montréal, University of Exeter, UCLA, UC Santa Barbara, Lakehead University, Michigan State University, and Copenhagen Business School for comments. MacKinnon and Webb thank the Social Sciences and Humanities Research Council of Canada (SSHRC grants 435-2016-0871 and 435-2021-0396) for financial support. Nielsen thanks the Danish National Research Foundation for financial support (DNRF Chair grant number DNRF154). Computer code, including a `Stata` ado file, for performing the testing procedures proposed here may be found at http://qed.econ.queensu.ca/pub/faculty/mackinnon/svtest/.

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

# 1 Introduction

Modern empirical econometrics often allows for correlation within clusters of observations, and this can have 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 tests 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) cover much of the literature up to a few years ago. Esarey and Menger (2019) and MacKinnon and Webb (2020) provide more recent surveys. Conley, Gonçalves, and Hansen (2018) deal with a broader class of methods for various types of dependent data. MacKinnon, Nielsen, and Webb (2023a) provide a thorough and detailed guide to empirical practice. 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; MacKinnon, Nielsen, and Webb 2023b); and inference with unbalanced clusters (Imbens and Kolesár 2016; Carter, Schnepel, and Steigerwald 2017; MacKinnon and Webb 2017; Djogbenou, MacKinnon, and Nielsen 2019; MacKinnon, Nielsen, and Webb 2022).

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) advocate clustering at the larger, more aggregate level. In the case of randomized experiments, Athey and Imbens (2017) recommend clustering at the level of randomization. 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 (2022) recommend clustering at the pair level rather than the group level. While these rules of thumb can sometimes be very helpful, they may or may not lead to the appropriate clustering level 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). 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 severe (MacKinnon and Webb 2017, 2018), and can also reduce power (MacKinnon and Webb 2020).

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, or SV, 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.

Tests for the appropriate level of clustering have also been proposed by Ibragimov and Müller (2016) and recently by Cai (2022). These tests are very different from our tests and very different from each other. We discuss them briefly in Section 3.4.

The model of interest is discussed in Section 2. Our score-variance tests are described in Section 3, including the bootstrap implementation, the sequential testing procedure, and the use of our tests as pre-tests for inference about regression coefficients. Section 4 provides asymptotic theory for the two test statistics, the bootstrap tests, and the sequential testing procedure. In Section 5, we consider the common situation in which the regressors that are not of primary interest have been partialed out prior to performing the test. The size and power of the proposed tests are analyzed by Monte Carlo simulations in Section 6. An empirical example that deals with clustering by classroom or school using the STAR dataset (Finn and Achilles 1990; Mosteller 1995) is discussed in Section 7. Finally, Section 8 concludes and offers some guidance for empirical researchers. All mathematical proofs are given in Appendix A.

## 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, respectively, $N \times 1$ vectors of observations and disturbances (or error terms), and $\boldsymbol{X}$ is an $N \times k$ matrix of regressors (or covariates). The $k \times 1$ parameter vector $\boldsymbol{\beta}$ contains the coefficients on the regressors.

Suppose that the data are divided into $G$ clusters, indexed by $g$, where the $g^{\text{th}}$ 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 ordinary least squares (OLS) estimator 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

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

where $\mathbb{I}(\cdot)$ denotes the indicator function and $\boldsymbol{\Sigma}_g$ is a $k \times k$ variance matrix. Although the properties of the $\boldsymbol{\Sigma}_g$ depend on the properties of the variance matrix of $\boldsymbol{u}$, we do not explicitly make any assumptions about the latter because our tests are concerned solely with the variances of the score vectors.

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

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \hat{\boldsymbol{\Sigma}} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}, \tag{4}$$

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

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{c}} = m_{\mathrm{c}} \sum_{g=1}^{G} \boldsymbol{X}_g^\top \hat{\boldsymbol{u}}_g \hat{\boldsymbol{u}}_g^\top \boldsymbol{X}_g = m_{\mathrm{c}} \sum_{g=1}^{G} \hat{\boldsymbol{s}}_g \hat{\boldsymbol{s}}_g^\top, \tag{5}$$

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 $m_{\mathrm{c}}$ is a finite-sample correction, the most commonly

employed factor being $m_{\text{c}} = G/(G-1) \times (N-1)/(N-k)$, which is designed to account for degrees of freedom. Using $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}_{\text{c}}$ in (4) yields $\text{CV}_1$, the most widely-used CRVE for $\hat{\boldsymbol{\beta}}$. Asymptotic inference on regression coefficients using $\text{CV}_1$ is studied by Djogbenou et al. (2019) and Hansen and Lee (2019).

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

$$\hat{\boldsymbol{\Sigma}}_{\text{het}} = \sum_{i=1}^{N} \hat{u}_i^2 \boldsymbol{X}_i^\top \boldsymbol{X}_i = \boldsymbol{X}^\top \text{diag}(\hat{u}_1^2, \dots, \hat{u}_N^2) \boldsymbol{X}, \tag{6}$$

where $\boldsymbol{X}_i$ is the $i^{\text{th}}$ row of the $\boldsymbol{X}$ matrix and $\hat{u}_i$ is the $i^{\text{th}}$ residual. The variance matrix obtained by setting $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}_{\text{het}}$ in (4) is the famous heteroskedasticity-consistent variance matrix estimator (HCCME) of Eicker (1963) and White (1980). Of course, the matrix $\hat{\boldsymbol{\Sigma}}_{\text{het}}$ can be modified in various ways to improve its finite-sample properties (MacKinnon and White 1985; MacKinnon 2013). The simplest is to multiply it by $m_{\text{het}} = N/(N-k)$, so that (4) becomes what is usually called $\text{HC}_1$. □

**Remark 2.** As Abadie, Athey, Imbens, and Wooldridge (2023) point out, when the object of interest is the average treatment effect in a finite population, cluster-robust standard errors based on (5) can be "unnecessarily conservative." Consequently, they develop an approach to inference that depends both on how the data were sampled and on how treatment was assigned. In this paper, however, we follow most of the literature on cluster-robust inference and rely on the traditional approach in which every sample is treated as a random outcome from a data-generating process (DGP). The objective is to draw inferences about the parameters of the DGP, which may be interpreted as features of an infinitely large population; see MacKinnon et al. (2023a) for additional details. □

# 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 (4) 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.

To make our testing procedure operational, we formulate it 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. 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 $g = 1, \ldots, G$,

$$\mathrm{E}(\boldsymbol{s}_{gh} \boldsymbol{s}_{gh'}^\top) = \mathbb{I}(h = h') \boldsymbol{\Sigma}_{gh}, \quad h, h' = 1, \ldots, M_g, \tag{7}$$

where each of the $\boldsymbol{\Sigma}_{gh}$ is a $k \times k$ matrix. Thus, (3) and (7) embody the assumption that the fine clustering structure is nested within the coarse one. Another possible design would have a one-way clustering structure nested within a two-way one; see Remark 9.

Now let $\boldsymbol{\Sigma}_\mathrm{c}$ and $\boldsymbol{\Sigma}_\mathrm{f}$ denote the matrix $\boldsymbol{\Sigma}$ under the coarse and fine clustering structures, respectively. From (3) and (7), these matrices are

$$\boldsymbol{\Sigma}_\mathrm{c} = \sum_{g=1}^{G} \boldsymbol{\Sigma}_g \quad \text{and} \quad \boldsymbol{\Sigma}_\mathrm{f} = \sum_{g=1}^{G} \sum_{h=1}^{M_g} \boldsymbol{\Sigma}_{gh}. \tag{8}$$

We consider the null and alternative hypotheses

$$\mathrm{H}_0 \colon \lim_{N \to \infty} \boldsymbol{\Sigma}_\mathrm{f} \boldsymbol{\Sigma}_\mathrm{c}^{-1} = \mathbf{I} \quad \text{and} \quad \mathrm{H}_1 \colon \lim_{N \to \infty} \boldsymbol{\Sigma}_\mathrm{f} \boldsymbol{\Sigma}_\mathrm{c}^{-1} \neq \mathbf{I}. \tag{9}$$

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 et al. (2019).

**Remark 3.** In (9), we are not directly testing the fine clustering condition in (7). 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 $\hat{\boldsymbol{\beta}}$ is given by (4) with $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}_\mathrm{f}$. □

**Remark 4.** An important null hypothesis is that no CRVE is needed because the HCCME considered in Remark 1, obtained by combining (4) and (6), is valid. In this case, each fine cluster has just one observation, so that $M_g = N_g$ and $N_{gh} = 1$ for all $g$ and $h$. □

**Remark 5.** In practical applications, the number of coefficients in regression models, and hence the size of the CRVE matrices, is often large, so that comparing these matrices directly can be impractical. Furthermore, it is usually only one coefficient, or a small subset of them, that is actually of interest. Many coefficients typically correspond to fixed effects and other conditioning variables that are not of primary interest. By partialing out the latter, it is possible to reduce the dimensionality of the test and focus on the parameter(s) of interest. We discuss this issue in Section 5. □

## 3.1 Test Statistics

Our score-variance, or SV, 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 $k = 1$. This could be after all regressors except one have been partialed out (Section 5), so that interest is focused on a particular coefficient that we are trying to make inferences about. This leads to a test statistic with 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 $\hat{\boldsymbol{s}}_{ghi}$ denote the empirical score for observation $i$ within fine cluster $h$ in coarse cluster $g$, and let $\hat{\boldsymbol{s}}_{gh} = \sum_{i=1}^{N_{gh}} \hat{\boldsymbol{s}}_{ghi}$ denote the empirical score for fine cluster $h$ in coarse cluster $g$, such that $\hat{\boldsymbol{s}}_{g} = \sum_{h=1}^{M_g} \hat{\boldsymbol{s}}_{gh}$. Under coarse clustering, the estimated $\boldsymbol{\Sigma}_{\mathrm{c}}$ matrix in (5) is

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{c}} = m_{\mathrm{c}} \sum_{g=1}^{G} \hat{\boldsymbol{s}}_{g} \hat{\boldsymbol{s}}_{g}^{\top} = m_{\mathrm{c}} \sum_{g=1}^{G} \left( \sum_{h=1}^{M_g} \hat{\boldsymbol{s}}_{gh} \right) \left( \sum_{h=1}^{M_g} \hat{\boldsymbol{s}}_{gh} \right)^{\top}. \tag{10}$$

Similarly, we can write, c.f. (7) and (8),

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{f}} = m_{\mathrm{f}} \sum_{g=1}^{G} \sum_{h=1}^{M_g} \hat{\boldsymbol{s}}_{gh} \hat{\boldsymbol{s}}_{gh}^{\top}, \tag{11}$$

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 $\boldsymbol{X}$ becomes the vector $\boldsymbol{x}$, and the empirical scores are scalars. Specifically, $\hat{s}_{ghi} = x_{ghi}\hat{u}_{ghi}$ and $\hat{s}_{gh} = \sum_{i=1}^{N_{gh}} \hat{s}_{ghi}$ denote the empirical scores for observation $i$ and fine cluster $h$, respectively. Then the

matrices (10) and (11) reduce to the scalars

$$\hat{\sigma}_{\mathrm{c}}^2 = m_{\mathrm{c}} \sum_{g=1}^{G} \left( \sum_{h=1}^{M_g} \hat{s}_{gh} \right)^2 \quad \text{and} \quad \hat{\sigma}_{\mathrm{f}}^2 = m_{\mathrm{f}} \sum_{g=1}^{G} \sum_{h=1}^{M_g} \hat{s}_{gh}^2. \tag{12}$$

The quantities given in (10), (11), and (12) are all defined in essentially the same way. They simply amount to different choices of empirical scores. If $N_{gh} = 1$, then $\hat{\sigma}_{\mathrm{f}}^2$ simplifies to

$$\hat{\sigma}_{\mathrm{het}}^2 = \sum_{g=1}^{G} \sum_{h=1}^{M_g} \sum_{i=1}^{N_{gh}} \hat{s}_{ghi}^2, \tag{13}$$

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 (12), namely,

$$\hat{\theta} = \hat{\sigma}_{\mathrm{c}}^2 - \hat{\sigma}_{\mathrm{f}}^2. \tag{14}$$

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,

$$\hat{\boldsymbol{\theta}} = \mathrm{vech}(\hat{\boldsymbol{\Sigma}}_{\mathrm{c}} - \hat{\boldsymbol{\Sigma}}_{\mathrm{f}}), \tag{15}$$

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 (14) and (15). To this end, suppose that we observe the (scalar) scores $s_{gh}$ for fine cluster $h$ in coarse cluster $g$. Then the analog of $\hat{\theta}$ is the contrast

$$\theta = \sum_{g=1}^{G} \sum_{h_1=1}^{M_g} \sum_{h_2 \neq h_1}^{M_g} s_{gh_1} s_{gh_2}. \tag{16}$$

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, $\theta$ clearly has mean zero by (7).

The variance of $\theta$ in (16) is, under the null hypothesis,

$$\mathrm{Var}(\theta) = \sum_{g=1}^{G} \sum_{h_1=1}^{M_g} \sum_{\ell_1=1}^{M_g} \sum_{h_2 \neq h_1}^{M_g} \sum_{\ell_2 \neq \ell_1}^{M_g} \mathrm{E}(s_{gh_1} s_{g\ell_1} s_{gh_2} s_{g\ell_2}). \tag{17}$$

The expectation of any product of scores can only be nonzero, under the null and (7), 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 (17) simplifies to

$$\mathrm{Var}(\theta) = 2 \sum_{g=1}^{G} \sum_{h_1=1}^{M_g} \sum_{h_2 \neq h_1}^{M_g} \sigma_{gh_1}^2 \sigma_{gh_2}^2, \tag{18}$$

where $\sigma_{gh}^2 = \mathrm{Var}(s_{gh})$ is used to denote $\boldsymbol{\Sigma}_{gh}$ in the scalar case.

The sample analog of the right-hand side of (18) is $2 \sum_{g=1}^{G} \sum_{h_1=1}^{M_g} \sum_{h_2 \neq h_1}^{M_g} \hat{s}_{gh_1}^2 \hat{s}_{gh_2}^2$; see (7). This suggests the variance estimator

$$\widehat{\mathrm{Var}}(\hat{\theta}) = 2 \sum_{g=1}^{G} \left( \sum_{h=1}^{M_g} \hat{s}_{gh}^2 \right)^2 - 2 \sum_{g=1}^{G} \sum_{h=1}^{M_g} \hat{s}_{gh}^4. \tag{19}$$

This equation avoids the triple summation in (18) by squaring the sums of squared empirical scores, which then requires that the second term be subtracted. In deriving (19), we have ignored the factors $m_\mathrm{c}$ and $m_\mathrm{f}$, which are asymptotically irrelevant. If instead we had retained them, there would be no cancellation when subtracting $\hat{\sigma}_\mathrm{f}^2$ from $\hat{\sigma}_\mathrm{c}$, leading to a much more complicated (and computationally burdensome) expression for $\widehat{\mathrm{Var}}(\hat{\theta})$. Combining (14) and (19) yields the studentized test statistic

$$\tau_\sigma = \frac{\hat{\theta}}{\sqrt{\widehat{\mathrm{Var}}(\hat{\theta})}}. \tag{20}$$

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

**Remark 6.** The statistic defined in (20) yields either a one-sided or a two-sided test. Upper-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 $\tau_\sigma^2$ may also be of interest. The asymptotic theory in Section 4 handles both cases. □

**Remark 7.** Consider again the special case in which the null is heteroskedasticity with no clustering. When the elements of $\boldsymbol{x}$ display little intra-cluster correlation, the contrast $\hat{\theta}$, and hence the absolute value of $\tau_\sigma$, 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 factor, the ratio of clustered to non-clustered standard errors (Moulton 1986), will be relatively small. Of course, the opposite will be true when the elements of $\boldsymbol{x}$ display a lot of intra-cluster correlation. Thus, all else equal, SV tests may well yield different results for different choices of $\boldsymbol{x}$. □

When $k > 1$, so that $\hat{\boldsymbol{\theta}}$ is a vector, the variance estimator analogous to (19) is

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}) = 2\sum_{g=1}^{G} \boldsymbol{H}_k \left( \sum_{h=1}^{M_g} \hat{\boldsymbol{s}}_{gh}\hat{\boldsymbol{s}}_{gh}^\top \otimes \sum_{h=1}^{M_g} \hat{\boldsymbol{s}}_{gh}\hat{\boldsymbol{s}}_{gh}^\top \right) \boldsymbol{H}_k^\top - 2\sum_{g=1}^{G}\sum_{h=1}^{M_g} \boldsymbol{H}_k \left( \hat{\boldsymbol{s}}_{gh}\hat{\boldsymbol{s}}_{gh}^\top \otimes \hat{\boldsymbol{s}}_{gh}\hat{\boldsymbol{s}}_{gh}^\top \right) \boldsymbol{H}_k^\top. \quad (21)$$

Here $\boldsymbol{H}_k$ is the so-called elimination matrix, which satisfies $\mathrm{vech}(\boldsymbol{S}) = \boldsymbol{H}_k \mathrm{vec}(\boldsymbol{S})$ 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

$$\tau_\Sigma = \hat{\boldsymbol{\theta}}^\top \widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})^{-1} \hat{\boldsymbol{\theta}}. \quad (22)$$

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

**Remark 8.** As pointed out by a referee, Cho and Phillips (2018) develop simple measures of the discrepancy between two positive definite symmetric matrices. When bootstrapped, these measures can be used as alternative test statistics for cases where $k > 1$. Preliminary simulations suggest that these bootstrap tests can work well, although not (in general) better than our proposed bootstrap tests based on (22). A full analysis is beyond the scope of this paper and is therefore left for future work. □

**Remark 9.** It is possible to use the test statistics (20) and (22) for testing one-way against two-way clustering. Suppose there are two alternative clustering dimensions, labeled A and B, and their intersection is labeled I. These could correspond to, say, state (A) and year (B), with the intersection denoting observations that correspond to the same year in the same state. Let $\hat{\boldsymbol{\Sigma}}_j$ denote the (one-way) CRVE in (5) under clustering dimension $j \in \{\mathrm{A}, \mathrm{B}, \mathrm{I}\}$. Then the two-way CRVE (Cameron, Gelbach, and Miller 2011) is given by (4) with

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{TW}} = \hat{\boldsymbol{\Sigma}}_{\mathrm{A}} + \hat{\boldsymbol{\Sigma}}_{\mathrm{B}} - \hat{\boldsymbol{\Sigma}}_{\mathrm{I}}. \quad (23)$$

If we test the null of one-way clustering by A against the alternative of clustering by both A and B, then $\hat{\boldsymbol{\Sigma}}_{\mathrm{c}} = \hat{\boldsymbol{\Sigma}}_{\mathrm{TW}}$ and $\hat{\boldsymbol{\Sigma}}_{\mathrm{f}} = \hat{\boldsymbol{\Sigma}}_{\mathrm{A}}$. Therefore, the vector of contrasts in (15) becomes

$$\hat{\boldsymbol{\theta}} = \mathrm{vech}(\hat{\boldsymbol{\Sigma}}_{\mathrm{TW}} - \hat{\boldsymbol{\Sigma}}_{\mathrm{A}}) = \mathrm{vech}(\hat{\boldsymbol{\Sigma}}_{\mathrm{B}} - \hat{\boldsymbol{\Sigma}}_{\mathrm{I}}). \quad (24)$$

The result in (24) shows that testing the null of one-way clustering by A against the alternative of two-way clustering by A and B must lead to the same test statistic as testing the null of one-way clustering by I against the alternative of one-way clustering by B.

Although it is straightforward to derive a test statistic based on (24), the asymptotic analysis of this statistic would be different from the analysis for testing nested one-way clustering in Section 4 below. For example, to derive the asymptotic null distribution of

a statistic based on (24) for testing clustering by I against clustering by B would mean analyzing it under the DGP that clustering is in fact by A. Therefore, we leave this analysis for future work.

Similarly, if there is two-way clustering under both the null and alternative hypotheses, it may be feasible to calculate score-variance statistics similar to (20) and (22). However, the analysis of the asymptotic null distribution would require completely different, and technically nontrivial, techniques (Davezies, D'Haultfœuille, and Guyonvarch 2021; MacKinnon, Nielsen, and Webb 2021; Menzel 2021; Chiang, Kato, and Sasaki 2022). □

In this section, we have proposed two score-variance tests of (9). They both involve comparing different variance estimates of the empirical scores, namely, the two scalars in (12) for the $\tau_\sigma$ test and the matrices in (10) and (11) for the $\tau_\Sigma$ test. The former is a special case of the latter, and it can be obtained for the same models as the latter by partialing out all regressors except one, as in Section 5. This special case is particularly interesting, because the $\tau_\sigma$ test can be directional, and also because many equations simplify neatly in the scalar case.

As we show in Section 6, 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 (20) and (22), which often perform much better in finite samples, as we also show in Section 6. 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 et al. 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 $b^{\text{th}}$ 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), Djogbenou et al. (2019), and Webb (2022).

The algorithm for a wild (cluster) bootstrap-based implementation of our tests is as follows. It applies to both $\tau_\sigma$ and $\tau_\Sigma$. For simplicity, the algorithm below simply refers to one test statistic, $\tau$. 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 $\tau_\Sigma$ and three with $k = 1$ based on different versions of $\tau_\sigma$.

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

1. Estimate model (1) by OLS to obtain the residuals $\hat{\boldsymbol{u}}$.
2. Compute the empirical score vector $\hat{\boldsymbol{s}}$ and use it to compute $\tau$.
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 $\boldsymbol{X}$, to compute $\tau^{*b}$, the bootstrap analog of $\tau$.
4. Compute the bootstrap $P$ value $\hat{P}^* = B^{-1}\sum_{b=1}^{B}\mathbb{I}(\tau^{*b} > \tau)$.

As usual, if $\alpha$ is the level of the test, then $B$ should be chosen so that $(1-\alpha)(B+1)$ 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 $\alpha$. Power increases in $B$, but it does so very slowly once $B$ exceeds a few hundred (Davidson and MacKinnon 2000).

**Remark 10.** When $\tau$ is defined as $\tau_\sigma$, Algorithm 1 yields a one-sided upper-tail test. When $\tau$ is defined as $|\tau_\sigma|$, $\tau_\sigma^2$, or $\tau_\Sigma$, it yields a two-sided test; see Remark 6. □

**Remark 11.** If desired, bootstrap critical values can be calculated as quantiles of the $\tau^{*b}$. For example, when $B = 999$ and the $\tau^{*b}$ are sorted from smallest to largest, the 0.05 critical value for a one-sided upper-tail test is $\tau^{*b'}$ for $b' = (1 - 0.05)(B + 1) = 950$. □

**Remark 12.** 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. □

## 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. Berger and Sinclair (1984) show 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 $\mathbf{\Sigma}$ matrices by $\mathbf{\Sigma}_0, \mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_p$; see (4) and (8). 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,

$$\hat{m} = \min\{m \in \{0, 1, \ldots, p\} \text{ such that } \mathrm{H}_0 \colon \operatorname*{plim}_{N\to\infty} \boldsymbol{\Sigma}_m \boldsymbol{\Sigma}_{m+1}^{-1} = \mathbf{I} \text{ is not rejected}\}. \tag{25}$$

Of course, the $\hat{m}$ resulting from Algorithm 2 and from (25) 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 (25) is (nearly) consistent. Precise asymptotic properties of the proposed sequential procedure are established in Section 4.3, and finite-sample performance is investigated by Monte Carlo simulation methods in Section 6.3.

## 3.4 Other Tests for the Level of Clustering

To our knowledge, only two other tests for the appropriate level of clustering have been proposed. Like our $\tau_\sigma$ test, these concern the standard error for a single coefficient. The best-known of them is due to Ibragimov and Müller (2016), and we refer to it as the IM test. It is a one-sided test that is derived from the procedure of Ibragimov and Müller (2010). The IM test is based on the assumption that $G$ is fixed while the number of observations tends to infinity. In this respect, it differs from our score-variance tests, for which the asymptotic theory in Section 4.1 requires that $G \to \infty$. However, see the discussion in Section 4.4.

There are two versions of the IM test. The first involves estimating the model separately for every coarse cluster. Unfortunately, this is impossible to do for models that involve treatment effects whenever the treatment is invariant within clusters. Even with treatment at the fine-cluster level, it may not be possible to estimate the model for every coarse cluster. This is the case, for example, in the empirical example of Section 7. Ibragimov and Müller (2016) therefore also propose a two-sample version of their test statistic that can be used for testing the level of clustering for treatment models when entire clusters are treated or not treated. Differences between estimates for treatment and control clusters can be used to estimate the treatment effects and perform a test of fine clustering.

Another test for the appropriate level of clustering for a single coefficient has very recently been proposed by Cai (2022). This test is based on randomization inference. Like the IM test, and unlike our $\tau_\sigma$ test, it is necessarily one-sided and treats $G$ as fixed. It also requires that the fine clusters be large, so that, unlike both the IM test and our tests, it cannot be

used to test the null hypothesis of independence at the observation level.

Cai (2022) presents results from a number of simulation experiments for his test, the IM test, and the bootstrap version of our $\tau_\sigma$ test. They suggest that the IM test and our test are much more similar to each other than they are to Cai's test. The IM test always rejects more often than the bootstrap version of our $\tau_\sigma$ test, both when the null hypothesis is true and when it is false. It can over-reject quite severely in some cases, especially when the ratio of fine to coarse clusters is small. In the first version of this paper, we presented some figures comparing rejection frequencies for the IM test and the $\tau_\sigma$ test. In the interests of space, however, we have omitted these results, because they are broadly similar to those from the experiments in Cai (2022).

At this point, what is known about the properties of our tests, the IM test, and Cai's test suggests that none of them is to be preferred in every case. They can all provide useful information about the appropriate level at which to cluster. Two attractive features of our tests, which are not shared by the other two, are that the $\tau_\sigma$ test can be either one-sided or two-sided and that the $\tau_\Sigma$ test is based on more than one coefficient of interest.

## 3.5 Inference about Regression Coefficients

The ultimate purpose of using any test for the appropriate level of clustering is to make more reliable inferences about the coefficient(s) of interest, that is, some element(s) of $\boldsymbol{\beta}$ in (1). This may or may not involve some sort of formal pre-testing or model averaging procedure.

For simplicity, suppose we are attempting to construct a confidence interval for $\beta_1$, 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. With a testing procedure like the ones proposed in this paper, an investigator can instead choose the level of clustering based on the outcome of a test. This involves choosing a level $\alpha$ for the test and deciding whether to use a one-sided or a two-sided test. We then form the interval based on coarse clustering when the 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. 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 usually 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 usually too short. In Section 6.4, we report the results of some simulation experiments that compare confidence

intervals based on several alternative procedures.

As Ibragimov and Müller (2016) point out, it probably makes sense to report confidence intervals for $\beta_1$ based on all clustering levels that appear plausible. The resulting inferences are then explicitly conditional on the level of clustering. Because our tests, like the other ones discussed in Section 3.4, provide evidence on the plausibility of each level of clustering, they can reduce the number of intervals that need to be reported. For example, if the hypothesis of independence is strongly rejected against one or more clustering structures, then it would not be necessary to report a confidence interval based on a heteroskedasticity-robust standard error. But if the $P$ value for the hypothesis of fine clustering against coarse clustering is neither extremely small nor very large, then it might well seem reasonable to report confidence intervals based on both levels. Whatever intervals an investigator chooses to report, tests for the appropriate clustering level can provide valuable information about which ones are empirically more plausible. These tests may thus be thought of as robustness checks (Cai 2022).

# 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 $\mathrm{Var}(\boldsymbol{s}_{gh}) = \boldsymbol{\Sigma}_{gh}$. 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 and maximum eigenvalues 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 $\lambda$ 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 (9) 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.

The denominators of both terms in Assumption 5 show that these conditions trade off intra-cluster dependence and cluster-size heterogeneity. That is, the greater the amount of intra-cluster correlation, the larger are the denominators in Assumption 5, which allows larger clusters without dominating the limit; a similar tradeoff was found in Djogbenou et al. (2019). Furthermore, more homogeneity in cluster sizes allows for fewer and larger clusters. We illustrate these tradeoffs in the following remarks.

**Remark 13.** Under Assumptions 1 and 4, both denominators 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

$$\sup_{g,h} N_{gh}^2 \left(\frac{\sup_g N_g}{\inf_g N_g}\right)\left(\frac{\sup_g N_g}{N}\right) \longrightarrow 0 \quad \text{and} \quad \left(\frac{\sup_g N_g}{\inf_g N_g}\right)^{\lambda}\left(\frac{\sup_{g,h} N_{gh}^{3\lambda-1}}{N^{\lambda-1}}\right) \longrightarrow 0. \tag{26}$$

If the cluster sizes are bounded under the alternative, i.e. $\sup_g N_g < \infty$, then (26) 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. Suppose, on the other hand, that Assumption 4 were strengthened to assume that $\inf_{g,h} N_{gh}^{-2}\omega_{\min}(\boldsymbol{\Sigma}_{gh}) > 0$, as would be the case if a random-

effects or factor-type model were assumed under the null. In that case, Assumption 5 and (26) could be weakened substantially. □

**Remark 14.** It is interesting to consider a setup for clusters that are relatively homogeneous, but possibly unbounded, in size. To make this concrete, suppose the coarse clusters have size $N_g = O(N^\alpha)$ for $g = 1, \dots, G$, where $\alpha \in [0, 1)$ and '$O(\cdot)$' is to be understood as an exact rate subject to $N_g$ being an integer. Because $\sum_{g=1}^{G} N_g = N$, it then holds that $G = O(N^{1-\alpha})$. Similarly, for each $g$, the fine clusters have size $N_{gh} = O(N_g^\gamma)$ for $h = 1, \dots, M_g = O(N_g^{1-\gamma})$. That is, when $\alpha$ is large (small), there are few large (many small) coarse clusters. Similarly, when $\gamma$ is large (small), there are few large (many small) fine clusters per coarse cluster. Under this setup, (26) is satisfied if $\alpha(2\gamma + 1) < 1$ and $\alpha\gamma < (\lambda - 1)/(3\lambda - 1)$.

The important implication of this setup is that the implied restrictions on the cluster sizes in Assumption 5 are very weak. In fact, if we assume that the fine cluster sizes are bounded (i.e., $\gamma = 0$), which applies, for example, in the important special case in which the scores are independent but heteroskedastic under the null, then we can allow $G = O(N^{1-\alpha})$ for any $\alpha < 1$. That is, the number of coarse clusters can be arbitrarily close to $O(1)$. For example, we allow $G = O(N^{0.1})$ and $N_g = O(N^{0.9})$, which corresponds to very few and very large coarse clusters. In this sense, our asymptotic framework can nearly accommodate the fixed-$G$ setup; see Section 4.4. □

## 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 (9). □

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, we do not make that assumption. 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

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

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 $\boldsymbol{\Sigma}_g = \sum_{h_1=1}^{M_g} \sum_{h_2=1}^{M_g} \mathrm{E}(\boldsymbol{s}_{gh_1} \boldsymbol{s}_{gh_2}^\top)$ to be sufficiently large, in aggregate. This condition is embodied in Assumption 7.

**Remark 16.** As in Remark 13, there is a tradeoff between cluster size heterogeneity and intra-cluster 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

$$\frac{\sup_g N_g^3}{N} \longrightarrow 0. \tag{27}$$

Suppose instead that Assumption 4 were strengthened to assume that $\inf_g N_g^{-2} \omega_{\min}(\boldsymbol{\Sigma}_g) > 0$ (as in Remark 13, 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

$$\frac{\sup_g N_g^3}{\inf_g N_g^2 N} \longrightarrow 0. \tag{28}$$

With relatively homogeneous coarse clusters as in Remark 14, i.e. coarse clusters where $\sup_g N_g$ and $\inf_g N_g$ are of the same order of magnitude, the condition (28) 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 (9) is not true. Then, as $N \to \infty$, it holds that*

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

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.

Of course, power will depend in a complicated way on many aspects of the model and DGP, including the number of large clusters and their sizes, because these will affect the number of correlations that need to be estimated between fine clusters within coarse clusters; see (16). Power will also depend on the true values of these correlations. If they are mostly non-zero and non-trivial, then power will be higher with larger coarse clusters.

## 4.2 Theory for Bootstrap Tests

We now demonstrate the asymptotic validity of the bootstrap implementation of our tests. To this end, let $\tau$ denote either of our statistics, and let the cumulative distribution function of $\tau$ under $\mathrm{H}_0$ be denoted $P_0(\tau \leq x)$. The corresponding bootstrap statistic is denoted $\tau^*$. As usual, let $P^*$ denote the bootstrap probability measure, conditional on a given sample, and let $\mathrm{E}^*$ 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 $\mathrm{H}_0$. 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

The next theorem provides theoretical justification for the sequential testing procedure given in Algorithm 2.

**Theorem 4.** *Let $\hat{m}$ be defined in Algorithm 2 or (25). Suppose Assumption 1 is satisfied when the "fine" clustering level in (9) is $m = m_0 \in \{0, 1, \ldots, p\}$ (and hence also when $m > m_0$), and suppose $H_0$ in (9) is not true for clustering levels $m < m_0$. Suppose also that Assumptions 2–5 and 7 are satisfied, and let $\alpha$ 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 (25) 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 "underestimate" the clustering level, at least asymptotically, because $\hat{m} < m_0$ with probability converging to 0.

## 4.4 Fixed-$G$ Asymptotic Theory

In the literature on cluster-robust inference, a few authors have considered an alternative asymptotic framework, referred to as fixed-$G$ asymptotics, in which the number of clusters is fixed as $N \to \infty$ while cluster sizes diverge; key early papers are Ibragimov and Müller (2010) and Bester, Conley, and Hansen (2011). However, fixed-$G$ asymptotics are proven under the very strong assumption that a central limit theorem applies to the normalized scores for each cluster. This assumption seriously limits the amount of intra-cluster dependence. For example, it rules out common models such as many types of random-effects and factor models. See MacKinnon et al. (2023a) for a detailed discussion.

Nonetheless, we now briefly consider an asymptotic framework in which the number of coarse clusters, $G$, is fixed, but there are many fine clusters within each coarse cluster, i.e. $M_g \to \infty$ for all $g$. For simplicity, we consider the scalar case with $k = 1$. Let $\sigma_g^2 = \text{Var}(\sum_{h=1}^{M_g} s_{gh}) = \sum_{h=1}^{M_g} \sigma_{gh}^2$ (under the null) and define the weights $w_g^2 = \lim_{M_g \to \infty} \sigma_g^2 / \text{Var}(\theta)^{1/2}$, where $\text{Var}(\theta)$ is given in (18). Then suppose, for all $g$, that (i) $\sigma_g^{-1} \sum_{h=1}^{M_g} s_{gh} \xrightarrow{d} \text{N}(0,1)$, (ii) $\sigma_g^{-1} \sum_{h=1}^{M_g} s_{gh}^2 \xrightarrow{P} 1$, and (iii) $w_g^2 \in [0, \infty)$. The high-level condition (i) is typical of the fixed-$G$ literature and imposes very strong limitations on the amount of intra-cluster dependence that is allowed. Condition (ii) is a homogeneity assumption, and condition (iii) ensures that one cluster does not dominate the sum in the limit. Under the null hypothesis and these conditions, it can be proven that

$$\frac{\theta}{\sqrt{\text{Var}(\theta)}} \xrightarrow{d} \sum_{g=1}^{G} w_g^2 (\chi_{1,g}^2 - 1), \tag{29}$$

where $\chi^2_{1,g}$ for $g = 1, \ldots, G$ denote independent $\chi^2_1$ random variables. Under suitable additional regularity conditions, we conjecture that $\tau_\sigma = \hat{\theta}/\sqrt{\widehat{\text{Var}}(\hat{\theta})}$ has the same asymptotic distribution as in (29).

The limiting distribution in (29) is a weighted sum of independent $\chi^2_1$ random variables. Because the weights $w_g^2$ depend on unknown parameters, the distribution is non-pivotal and hence cannot be used for inference. Under the extreme homogeneity condition that the $w_g^2$ are the same for all $g$, the distribution simplifies to $(\chi^2_G - G)/\sqrt{2G}$, which is a centered and normalized $\chi^2_G$. This distribution is pivotal and could be used for inference, although the conditions under which it is derived are extraordinarily strong.

Continuing with this type of fixed-$G$ asymptotic argument, we could instead assume that $N_{gh} \to \infty$ for all $g, h$, while the $M_g$ and $G$ are fixed. That is, the number of observations within each fine cluster diverges, but there are only a fixed number of fine and coarse clusters. This setup is quite similar to the previous one. We conjecture that the asymptotic distribution would again be a weighted sum of $\chi^2_1$ random variables similar to the one in (29), but the summation would extend over $\sum_{g=1}^{G} M_g = G_\text{f}$ elements.

In either case, if the weights $w_g^2$ are not too heterogeneous, the fixed-$G$ limiting distributions could be well approximated by a standard normal distribution, at least when the number of clusters is not very small. In the setup with $M_g \to \infty$, this would be the number of coarse clusters, $G$. In the setup with $N_{gh} \to \infty$, it would be the number of fine clusters, $G_\text{f}$. Thus, in the end, the normal limit theory obtained under large-$G$ asymptotics in Theorem 1 and Corollary 1 may also provide a good approximation under fixed-$G$ asymptotics. A full analysis of fixed-$G$ asymptotic theory for our model and test statistics would be interesting, but it is beyond the scope of this paper and is consequently left for future work.

# 5 Dimension Reduction by Partialing Out

As discussed in Remark 5, it is commonly the case in empirical work that the number of regressors is very large and that most of the regressors are not of primary interest. Comparing large-dimensional CRVE matrices by the methods in Section 3 is impractical. Fortunately, it is easy to solve this problem by partialing out the regressors that are not of primary interest prior to performing our tests.

Suppose the full set of regressors is partitioned as $\boldsymbol{X} = [\boldsymbol{X}_1, \ \boldsymbol{X}_2]$, where $\boldsymbol{X}_1$ denotes the $N \times k_1$ matrix of the regressors of interest and $\boldsymbol{X}_2$ denotes the $N \times k_2$ matrix of other regressors, with $k = k_1 + k_2$. 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 $k_1 \times 1$ parameter vector $\boldsymbol{\beta}_1$ and the rest 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 $k_1$ regressions. The model (1) can then be rewritten as

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\delta} + \boldsymbol{u}, \qquad \boldsymbol{Z} = \boldsymbol{M}_{\boldsymbol{X}_2}\boldsymbol{X}_1, \tag{30}$$

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 (30) have exactly the same explanatory power and the same disturbances, $\boldsymbol{u}$. The coefficient $\boldsymbol{\beta}_1$ in (30) is identical to the one defined in the previous paragraph, but the coefficient $\boldsymbol{\delta}$ 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),

$$\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{y} = \boldsymbol{\beta}_{1,0} + (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\sum_{g=1}^{G}\boldsymbol{Z}_g^\top\boldsymbol{u}_g, \tag{31}$$

where $\boldsymbol{\beta}_{1,0}$ is the true value of $\boldsymbol{\beta}_1$. The relationship between $\boldsymbol{Z}$ and $\boldsymbol{X}$ can be written as

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{Q} \quad \text{with} \quad \boldsymbol{Q} = [\mathbf{I}_{k_1},\ -\boldsymbol{X}_1^\top\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{X}_2)^{-1}]^\top. \tag{32}$$

Therefore, the score for $\boldsymbol{\beta}_1$ is $\boldsymbol{Z}_g^\top\boldsymbol{u}_g = \boldsymbol{Q}^\top\boldsymbol{X}_g^\top\boldsymbol{u}_g = \boldsymbol{Q}^\top\boldsymbol{s}_g$. Thus, from (31) and (32), we obtain the following sandwich formula, c.f. (4),

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_1) = (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Q}^\top\hat{\boldsymbol{\Sigma}}\boldsymbol{Q}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}. \tag{33}$$

Under Assumption 3, $\boldsymbol{Q} \xrightarrow{P} [\mathbf{I}_{k_1},\ -\boldsymbol{\Xi}_{12}\boldsymbol{\Xi}_{22}^{-1}]^\top = \boldsymbol{A}$, say, so that the middle matrix in (33) is clearly an estimator of $\boldsymbol{A}^\top\boldsymbol{\Sigma}\boldsymbol{A}$.

The matrix $\boldsymbol{Q}$ in (32) and its limit $\boldsymbol{A}$ can be viewed as mechanisms for dimension reduction. They transform the problem from one involving the $k \times k$ matrix $\boldsymbol{\Sigma}$ to one involving the $k_1 \times k_1$ matrix $\boldsymbol{A}^\top\boldsymbol{\Sigma}\boldsymbol{A}$. The latter is the variance of $\boldsymbol{A}^\top\boldsymbol{X}^\top\boldsymbol{u}$, and it depends on the clustering structure in the same way as $\boldsymbol{\Sigma}$. For the model (30), we consequently replace the hypotheses in (9) with

$$\mathrm{H}_0\colon \lim_{N\to\infty}(\boldsymbol{A}^\top\boldsymbol{\Sigma}_\mathrm{f}\boldsymbol{A})(\boldsymbol{A}^\top\boldsymbol{\Sigma}_\mathrm{c}\boldsymbol{A})^{-1} = \mathbf{I} \quad \text{and} \quad \mathrm{H}_1\colon \lim_{N\to\infty}(\boldsymbol{A}^\top\boldsymbol{\Sigma}_\mathrm{f}\boldsymbol{A})(\boldsymbol{A}^\top\boldsymbol{\Sigma}_\mathrm{c}\boldsymbol{A})^{-1} \neq \mathbf{I}. \tag{34}$$

Furthermore, from (32), we see that we can use the same algebra for the model in (30) as for the model in (1) to define the test statistics, i.e. (10), (11), and so on, but now with empirical scores $\boldsymbol{Q}^\top\hat{\boldsymbol{s}}_g$ and $\boldsymbol{Q}^\top\hat{\boldsymbol{s}}_{gh}$ instead of $\hat{\boldsymbol{s}}_g$ and $\hat{\boldsymbol{s}}_{gh}$, respectively. This also applies to

the bootstrap implementation in Algorithm 1. Of course, degrees-of-freedom corrections like the factor $m_c$ in (5) need to reflect the total number of estimated coefficients.

Since very few regression models in economics contain just one regressor, the $\tau_\sigma$ test will almost always involve partialing out. It seems likely that the $\tau_\Sigma$ test will also involve partialing out in the vast majority of cases, so that the dimension of the vector $\hat{\boldsymbol{\theta}}$ upon which the $\tau_\Sigma$ test is based will be $k_1$ rather than $k$.

**Remark 17.** The empirical scores $\boldsymbol{Q}^\top \hat{s}_{gh}$ and $\boldsymbol{Q}^\top \hat{\boldsymbol{s}}_{gh}$ depend on the matrix $\boldsymbol{Z} = \boldsymbol{M}_{\boldsymbol{X}_2}\boldsymbol{X}_1$, which is the residual matrix from regressing $\boldsymbol{X}_1$ on $\boldsymbol{X}_2$. Therefore, different choices for $\boldsymbol{X}_1$ will yield different empirical scores, and hence different test statistics; see Remark 7. This is also reflected in the hypotheses in (34), where different choices for $\boldsymbol{X}_1$ will yield a different $\boldsymbol{A}$ matrix and hence different null and alternative hypotheses. □

**Remark 18.** Theorems 1–4 continue to hold with the new definitions given in this section, with $\boldsymbol{\beta}_1$ replacing $\boldsymbol{\beta}$, $\boldsymbol{Z}$ replacing $\boldsymbol{X}$, and $k_1$ replacing $k$. Because the matrix $\boldsymbol{Q} \xrightarrow{P} \boldsymbol{A}$ under Assumption 3, it acts only as a fixed constant in all asymptotic arguments; that is, $\boldsymbol{Q}^\top \boldsymbol{s}_{gh} = \boldsymbol{A}^\top \boldsymbol{s}_{gh}(1 + o_P(1))$. Thus, the same proofs apply with $\boldsymbol{s}_{gh}$ replaced by $\boldsymbol{A}^\top \boldsymbol{s}_{gh}$. □

**Remark 19.** Careful inspection of the proofs shows that, in the setup of this section, we can replace $\boldsymbol{\Sigma}_{gh}$ with $\boldsymbol{A}^\top \boldsymbol{\Sigma}_{gh} \boldsymbol{A}$ in Assumption 5. This could be attractive in some cases. Suppose, for example, that $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are (asymptotically) orthogonal, such that $\boldsymbol{A}^\top \boldsymbol{\Sigma} \boldsymbol{A}$ is equal to the diagonal block of $\boldsymbol{\Sigma}$ corresponding to $\boldsymbol{X}_1^\top \boldsymbol{u}$. Suppose also that $\boldsymbol{X}_1$ and $\boldsymbol{u}$ are both finely clustered, but the $\boldsymbol{X}_2$ are independent. Then $\boldsymbol{A}^\top \boldsymbol{\Sigma}_{gh} \boldsymbol{A}$ satisfies the condition in Remark 13, while $\boldsymbol{\Sigma}_{gh}$ only satisfies the corresponding condition in Assumption 4, and hence using $\boldsymbol{A}^\top \boldsymbol{\Sigma}_{gh} \boldsymbol{A}$ in Assumption 5 would lead to a weaker condition. □

# 6 Simulation Experiments

Most of the papers cited in the second paragraph of Section 1 employ simulation experiments to study the finite-sample properties of 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 $\xi_g$, which affects every observation within that cluster *equally*. This yields disturbances that are equi-correlated within each cluster.

Although this type of DGP 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. Because the fixed effects completely explain the $\xi_g$, the residuals are always uncorrelated. Thus, for models with cluster fixed effects, it is always valid to use

heteroskedasticity-robust (HR) standard errors whenever the intra-cluster correlation of the disturbances arises solely from a random-effects model. In such cases, 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 7. 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 DGP for which cluster fixed effects do not remove all of the intra-cluster correlation. To this end, we generate both the regressors and the disturbances in our experiments using factor models of the form

$$\begin{aligned} z_{gi} &= \rho^{1/2}\xi_g^1 + (1-\rho)^{1/2}\zeta_{gi} \quad \text{if } i \text{ is odd} \\ z_{gi} &= \rho^{1/2}\xi_g^2 + (1-\rho)^{1/2}\zeta_{gi} \quad \text{if } i \text{ is even.} \end{aligned} \tag{35}$$

Here $\xi_g^1$ and $\xi_g^2$ are random effects, distributed as standard normal, which apply respectively to the odd-numbered and even-numbered observations within the $g^{\text{th}}$ cluster. The $\zeta_{gi}$ are also distributed as standard normal. Under the DGP (35), the $z_{gi}$ have variance one, and the intra-cluster correlation of the odd (or even) observations is $\rho \geq 0$.

The DGP (35) can be interpreted in a variety of ways, depending on the nature of the data. The idea is that there are two types of observations within each cluster, and all the intra-cluster correlation is within each type. For example, with clustering at the geographical level, there might be two sub-regions. With clustering at the industry level, there might be two types of firm. The key assumption is that the researcher knows which cluster an observation belongs to, but not which type. Including cluster fixed effects explains some of the intra-cluster correlation by estimating an average of $\xi_g^1$ and $\xi_g^2$ for each cluster, but it does not explain all of it. Thus cluster-robust inference is still needed, and our tests should still have power.

In practice, of course, there might be more than than two types within each cluster, and the numbers of observations in each would almost certainly not be the same. It would be easy to make the DGP (35) more complicated. However, our objective is not to mimic any actual dataset, but simply to generate data in a way that allows cluster fixed effects to be combined with cluster-robust standard errors.

The DGP (35) makes no reference to fine and coarse clusters. It could be used to generate either finely or coarsely clustered data. The regressors $\boldsymbol{X}_1$ (that is, the ones whose coefficients are of interest; see Section 5) are generated using (35), and they are always

coarsely clustered. This ensures that, if the disturbances are either independent ($\rho = 0$), finely clustered, or coarsely clustered, the scores are also independent, finely clustered, or coarsely clustered, respectively.

In all experiments, each of the regressors in $\boldsymbol{X}_1$ is generated independently. This implies that there is no correlation among the coefficient estimates. It might seem that the extent of any such correlation would be important for the properties of the $\tau_\Sigma$ tests. However, that is not the case. We find numerically that the $\tau_\Sigma$ 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 columns of $\boldsymbol{X}_1$ independently.

## 6.1 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

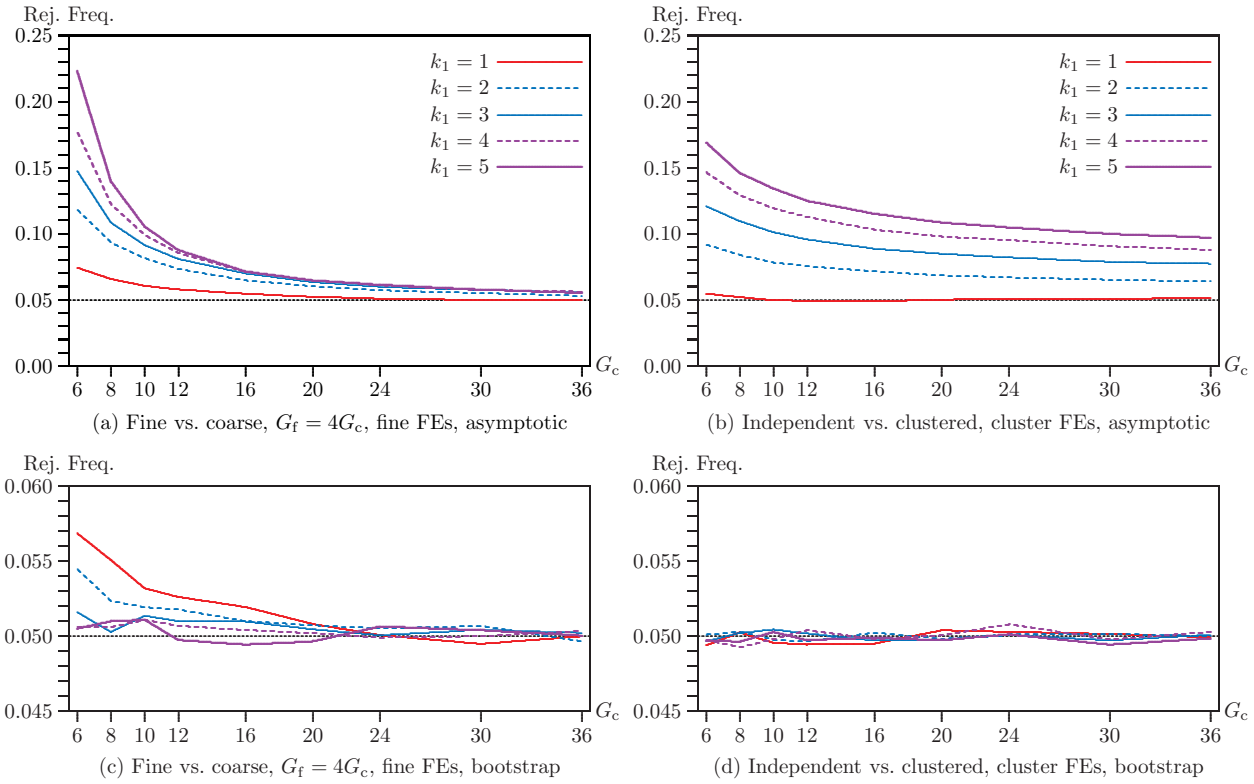
$$y_{ghi} = \sum_{\ell=1}^{k_1} \beta_\ell X_{ghi}^\ell + \boldsymbol{X}_{gh}^2 \boldsymbol{\delta} + u_{ghi}, \tag{36}$$

where the regressors $X_{ghi}^\ell$ are generated independently across $\ell$ by (35) at the coarse level with $\rho = 0.5$. The additional regressors in $\boldsymbol{X}_{gh}^2$ are either a constant term or a set of cluster fixed effects. When testing fine against coarse clustering, the fixed effects are at the fine level, and the disturbances are finely clustered with $\rho = 0.1$. When testing independence against (coarse) clustering, the fixed effects are at the coarse level, and the disturbances are independent. The number of coarse clusters, which in this section we denote by $G_\mathrm{c}$, 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 plots rejection frequencies at the 0.05 level for $\tau_\Sigma$ tests against $G_\mathrm{c}$, which varies from 6 to 36. We started at $G_\mathrm{c} = 6$ to avoid singularities when $k_1 = 5$ and stopped at $G_\mathrm{c} = 36$ because the results were hardly changing at that point. The values $k_1 = 1, \ldots, 5$ imply that the number of degrees of freedom for the tests is 1, 3, 6, 10, or 15. Panels (a) and (c) concern tests of fine clustering against coarse clustering, and panels (b) and (d) concern tests of independence against clustering. The top two panels report rejection frequencies for asymptotic tests at the 0.05 level, and the bottom two report comparable ones for bootstrap tests. Notice that the vertical axes for the asymptotic tests are much longer than the ones for the bootstrap tests, because the latter work very much better.

One striking feature of Figure 1 is that, for the asymptotic tests, over-rejection increases sharply with $k_1$. This should not have been a surprise in view of the fact that, like the

Figure 1: Rejection frequencies for $\tau_\Sigma$ tests at 0.05 level, $G_c$ varying



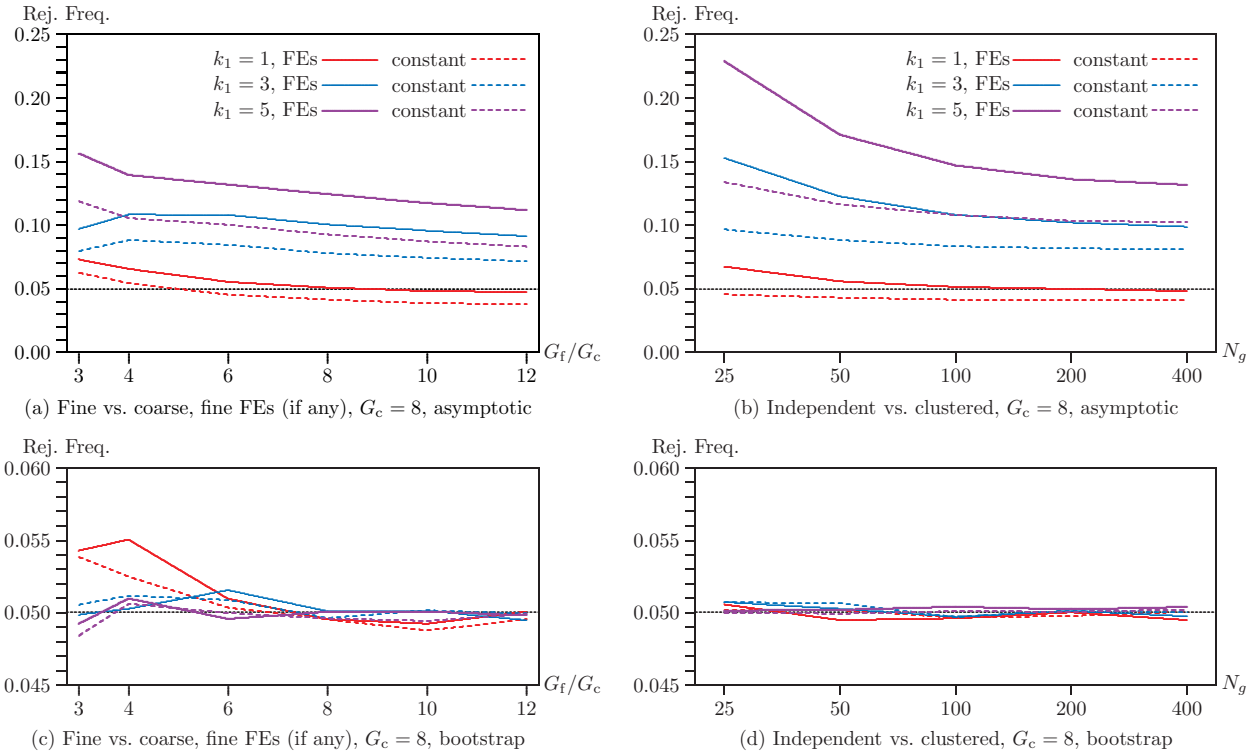

**Notes:** The regressors are generated by (35) with $\rho = 0.5$ and $1 \leq k_1 \leq 5$. The regressand is generated by (36). The disturbances are independent standard normals in Panels (b) and (d) and finely clustered with $\rho = 0.1$ in Panels (a) and (c). $G_c$ denotes the number of coarse clusters. Each coarse cluster contains 400 observations, so that $N = 400G_c$. In Panels (a) and (c), there are $G_f = 4G_c$ fine clusters, each containing 100 observations. Bootstrap tests employ $B = 399$. Panel (c) uses the wild cluster bootstrap, and Panel (d) uses the ordinary wild bootstrap. There are 400,000 replications.

information matrix test (White 1982), the $\tau_\Sigma$ test has degrees of freedom that are $O(k_1^2)$. 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 the number of coefficients being tested.

When $G_c$ is small, asymptotic tests of fine against coarse clustering, in Panel (a), over-reject more severely than tests of independence, in Panel (b). When $k_1 = 1$, there is almost no over-rejection for the tests of independence in Panel (b). For $k_1 \geq 2$, there is also more over-rejection in Panel (a) than in Panel (b) when $G_c = 6$, but the over-rejection diminishes much more rapidly as $G_c$ increases in Panel (a) than in Panel (b).

The bootstrap versions of the tests perform very much better than the asymptotic ones. There is slight over-rejection in Panel (c) for smaller values of $G_c$, which is really only noticeable for $k_1 = 1$ and $k_1 = 2$. In Panel (d), the bootstrap tests of independence work

Figure 2: Rejection frequencies for $\tau_\Sigma$ tests at 0.05 level, $G_\mathrm{f}$ or $N_g$ varying




(a) Fine vs. coarse, fine FEs (if any), $G_\mathrm{c} = 8$, asymptotic

(b) Independent vs. clustered, $G_\mathrm{c} = 8$, asymptotic

(c) Fine vs. coarse, fine FEs (if any), $G_\mathrm{c} = 8$, bootstrap

(d) Independent vs. clustered, $G_\mathrm{c} = 8$, bootstrap

**Notes:** The regressors are generated as in Figure 1, but only for $k_1 = 1$, 3, and 5. There are $G_\mathrm{c} = 8$ coarse clusters. In Panels (a) and (c), there are between 3 and 12 fine clusters per coarse cluster, each with 100 observations. In Panels (b) and (d), there are just coarse clusters, with between 25 and 400 observations per coarse cluster. Bootstrap tests employ $B = 399$. Panel (c) uses the wild cluster bootstrap, and Panel (d) uses the ordinary wild bootstrap. There are 400,000 replications.

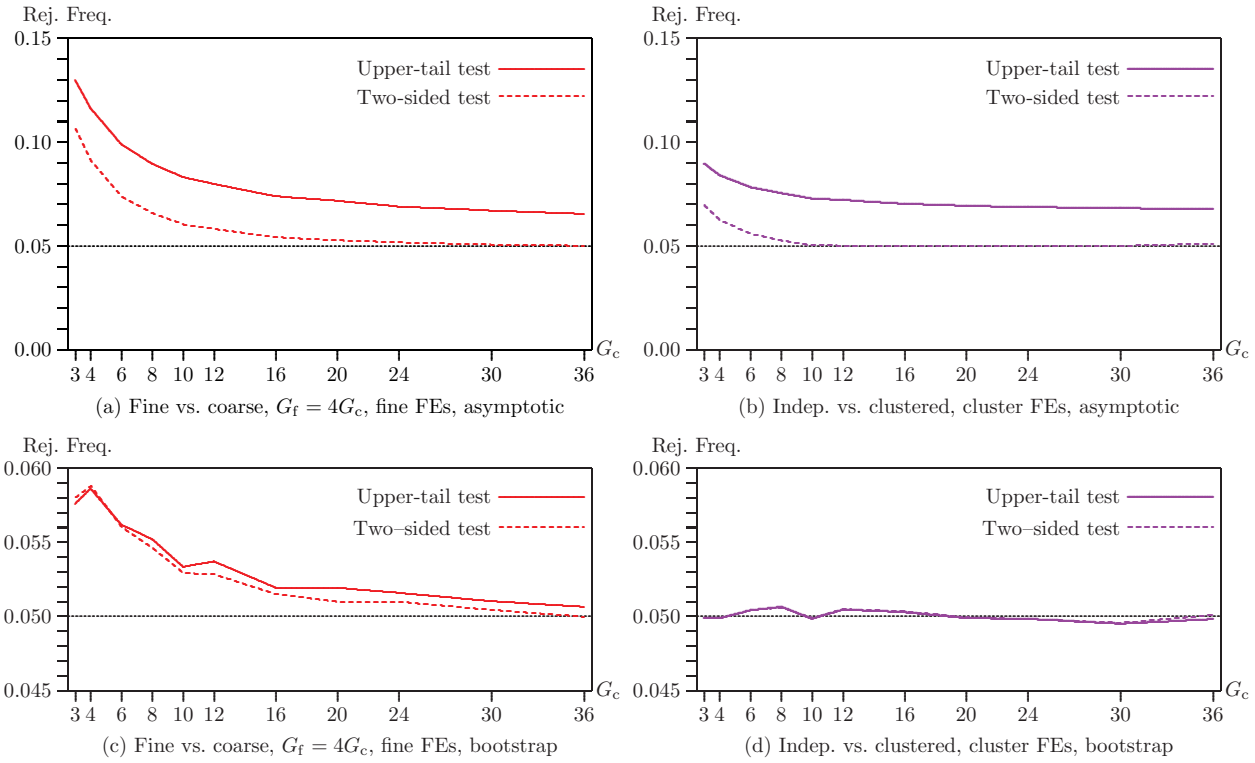
perfectly, except for experimental errors.

The bootstrap tests can be computationally demanding when the sample size is large, particularly for larger values of $k_1$. This is especially true for tests where the null hypothesis is no clustering, because the calculations in (10), (11), and (21) 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 $G_\mathrm{f}$ for tests of fine clustering.

In Figure 2, we hold the number of coarse clusters constant at $G_\mathrm{c} = 8$ and allow either the number of fine clusters per coarse cluster or the $N_g$ to vary. Results are shown for two specifications of (36). For the first of these, there are fine fixed effects when the null is fine clustering and cluster fixed effects when the null is independence, as in Figure 1. For the second, there is just a constant term. To make the figure readable, results are shown only for $k_1 = 1$, 3, and 5.

In Panels (a) and (c), the horizontal axis shows the number of fine clusters per coarse

Figure 3: Rejection frequencies for $\tau_\sigma$ tests at 0.05 level, $G_c$ varying

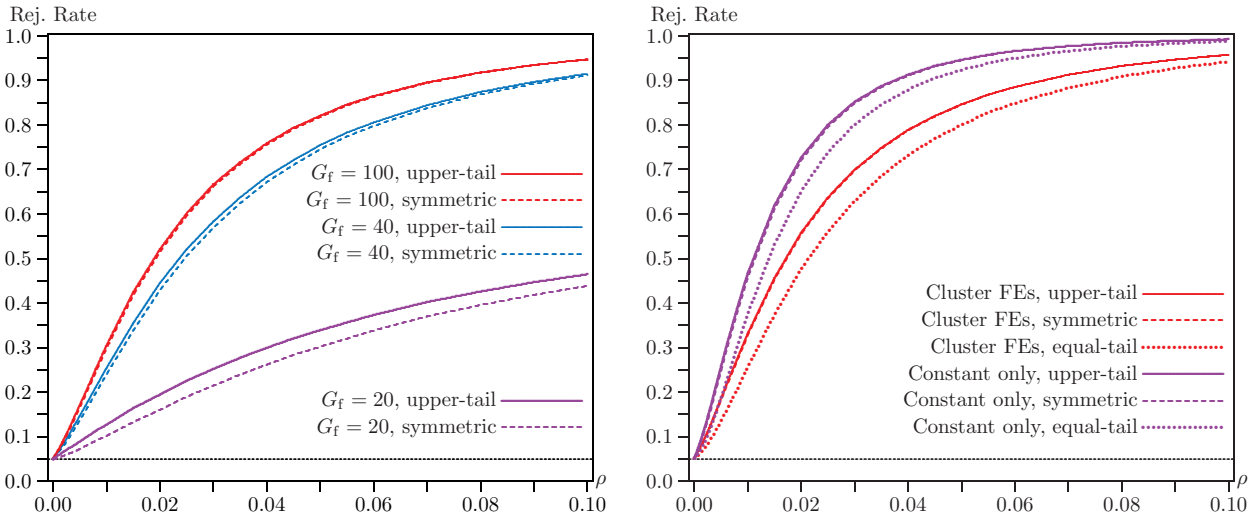



**Notes:** There is one regressor, which is generated by (35) with $\rho = 0.5$. In Panels (a) and (c), the disturbances are finely clustered with $\rho = 0.1$ and $G_f = 4G_c$ fine clusters, each with 100 observations. In Panels (b) and (d), they are independent standard normals. $G_c$ denotes the number of coarse clusters, each of which contains 400 observations, so that $N = 400G_c$. Bootstrap tests employ $B = 399$. Panel (c) uses the wild cluster bootstrap, and Panel (d) uses the ordinary wild bootstrap. There are 400,000 replications.

cluster, which varies between 3 and 12, so that the total number of fine clusters varies between 24 and 96. The rejection frequencies for asymptotic tests of fine against coarse clustering drop somewhat as $G_f/G_c$ increases. For the model with fixed effects, the asymptotic tests for $k_1 = 1$ work almost perfectly for $G_f/G_c \geq 8$, and all the bootstrap tests work almost perfectly for $G_f/G_c \geq 6$. The asymptotic tests always reject less often for the model with a constant term than for the model with fixed effects. For $k_1 = 1$, the former actually under-reject modestly for larger values of $G_f/G_c$.

In Panels (b) and (d), the horizontal axis shows the number of observations per coarse cluster, which varies between 25 and 400, on a log scale. It is evident that the asymptotic tests of independence perform better as the clusters become larger, although the curves are pretty flat at $N_g = 400$. The asymptotic tests with just a constant over-reject much less than the tests with fixed effects. When $k_1 = 1$, these tests under-reject for all values of $N_g$. All the bootstrap tests work essentially perfectly.

Figure 4: Power of bootstrap $\tau_\sigma$ tests at 0.05 level when there is coarse clustering




(a) Fine vs coarse, $G_c = 10$, $N = 5000$, coarse FEs, bootstrap

(b) Indep. vs. clustered, $G_c = 10$, $N = 5000$, bootstrap

**Notes:** The data are generated by (36) with coarse fixed effects and coarse (or no) clustering. There are 5000 observations, 10 coarse clusters, 20, 40, or 100 fine clusters, 400,000 replications, and 999 bootstraps.

Figure 3 shows rejection frequencies for both upper-tail and two-sided $\tau_\sigma$ tests. The experimental design is essentially the same as for Figure 1, except that, since $k_1 = 1$, results for $G_c = 3$ and $G_c = 4$ are included. The asymptotic upper-tail tests over-reject noticeably more often than the asymptotic two-sided tests. In contrast, the bootstrap upper-tail and two-sided tests perform identically (and extremely well). Thus it seems to be valuable to bootstrap both types of $\tau_\sigma$ test, but particularly important to bootstrap upper-tail tests.

## 6.2 The Power of Bootstrap Tests

In the next set of experiments, we turn our attention to power, focusing on the special case of the $\tau_\sigma$ test for a single coefficient. The data are generated by (36), with one regressor and coarse fixed effects. As usual, the regressor is generated by (35) with coarse clustering and $\rho = 0.5$. The disturbances are generated by the same model, with $\rho$ varying between 0.00 and 0.10. We report results only for bootstrap tests with $B = 999$. Using 999 instead of 399 reduces the, already quite small, power loss caused by using a finite number of bootstrap samples (Davidson and MacKinnon 2000).

Figure 4 shows the power of either two or three types of bootstrap $\tau_\sigma$ tests against coarse clustering as a function of the value of $\rho$ for the disturbances. The three types are upper-tail, symmetric, and equal-tail. As can be seen in both panels, all tests reject extremely close to 5% of the time when the null hypothesis is true. In Panel (a), the null hypothesis

is fine clustering for three different values of $G_{\text{f}}$. Power increases greatly when the number of fine clusters goes from 20 to 40. It increases further, but much more modestly, when $G_{\text{f}}$ goes from 40 to 100. The upper-tail tests are more powerful than the symmetric ones, but only slightly more when $G_{\text{f}} = 100$. To avoid making the figure unreadable, Panel (a) omits the equal-tail tests, which have much less power than the other two types of tests.

In Panel (b), the null hypothesis is independence, and the alternative is clustering with 10 (coarse) clusters. There are three bootstrap tests for a model with just a constant term and three tests for a model with cluster fixed effects. As expected, the tests are more powerful when there is just a constant term, since the fixed effects explain some of the intra-cluster correlation. For each set of tests, the upper-tail test is slightly more powerful than the symmetric test, which in turn is substantially more powerful than the equal-tail test.

The results in Panel (b) illustrate the fact that it generally makes no sense to use equal-tail bootstrap SV tests. These tests are designed to reject equally often in each tail under the null hypothesis. Since the mean of the test statistics under the null is positive in our experiments, the equal-tail test implicitly uses asymmetric critical values, with the positive one being larger in absolute value than the negative one. This reduces its power against $\sigma_{\text{c}}^2 > \sigma_{\text{f}}^2$, which is precisely the alternative we want SV tests to have power against.

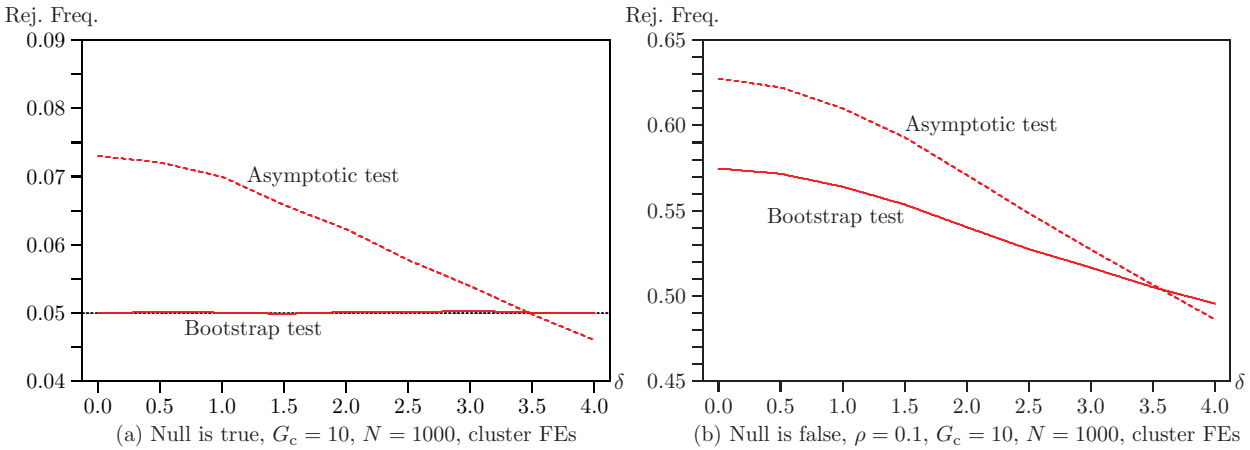
Up to this point, all the simulations have involved equal-sized clusters. There are many ways in which coarse-cluster sizes, fine-cluster sizes, and the numbers of fine clusters per coarse cluster could vary. We next allow cluster sizes to vary for one level of clustering. Figure 5 considers $\tau_\sigma$ tests of no clustering and plots rejection frequencies against a measure of cluster size variation. The $N$ observations are allocated among $G$ clusters using the equation

$$N_g = \left[ \frac{N \exp(\delta g/G)}{\sum_{j=1}^{G} \exp(\delta j/G)} \right] \quad \text{for } g = 1, \ldots, G-1, \tag{37}$$

where $\delta \geq 0$, $[\cdot]$ denotes the integer part of its argument, and $N_G = N - \sum_{g=1}^{G-1} N_g$. This scheme has been used in MacKinnon and Webb (2017), Djogbenou et al. (2019), and several other papers. In the experiments of Figure 5, $G = 10$ and $N = 1000$. When $\delta = 0$, $N_g = 100$ for all $g$. For $\delta = 1$, the $N_g$ range from 61 to 155; for $\delta = 2$, from 34 to 213; and for $\delta = 4$, from 9 to 340. There is one regressor and 10 cluster fixed effects.

In Panel (a) of Figure 5, the null hypothesis of no clustering is true. The upper-tail asymptotic test over-rejects noticeably for small values of $\delta$, but rejection frequencies decline as $\delta$ increases, and they are less than 0.05 for $\delta = 4$. In contrast, the upper-tail bootstrap test rejects almost exactly 5% of the time for all values of $\delta$. In Panel (b), the null hypothesis is false. Both tests have substantial power when $\delta$ is small, but it falls as $\delta$ increases. This makes sense, because the total number of off-diagonal elements in all the clusters increases

Figure 5: Cluster size variation and the performance of upper-tail $\tau_\sigma$ tests




(a) Null is true, $G_c = 10$, $N = 1000$, cluster FEs

(b) Null is false, $\rho = 0.1$, $G_c = 10$, $N = 1000$, cluster FEs

**Notes:** The regressor is generated by (35) with $\rho = 0.5$. The disturbances are generated by (35) with $\rho = 0.0$ in Panel (a) and $\rho = 0.1$ in Panel b). There are 10 clusters, 1000 observations, and cluster fixed effects. Cluster sizes vary according to (37). All tests are at the nominal 0.05 level. There are 400,000 replications and 399 bootstraps.

with $\delta$, causing the number of terms in the variance (18) to increase. The asymptotic test has noticeably more power than the bootstrap test for small values of $\delta$, but it has less power for the largest values, where it under-rejects under the null. The power differences almost certainly just reflect the size distortions of the asymptotic tests.

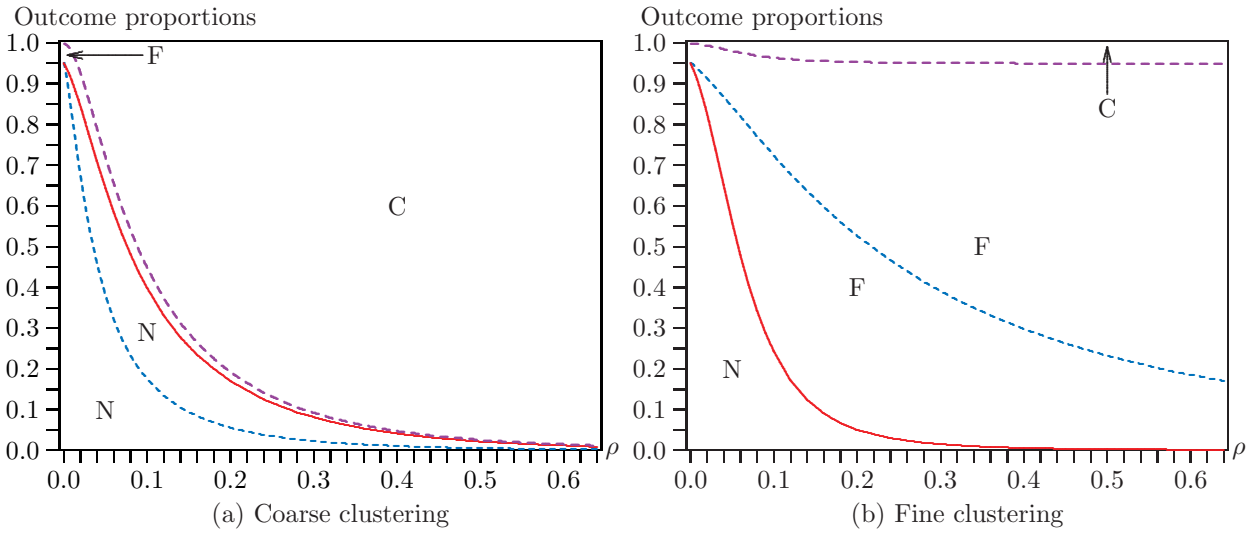
The results in Figure 5 suggest that the finite-sample performance of SV tests inevitably depends on the pattern of cluster sizes, although probably much less for bootstrap tests than for asymptotic ones.

## 6.3 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 8 coarse clusters, 48 fine clusters, and 2400 observations. As in Section 6.2, there are 999 bootstrap samples. The model always contains coarse-level fixed effects, and all of the tests are at the 0.05 level. The figure shows the outcomes of sequential, upper-tail $\tau_\sigma$ tests as $\rho$, the intra-cluster correlation for each set of disturbances generated by (35), varies within either coarse or fine clusters.

In Panel (a) of Figure 6, there is coarse clustering in the DGP, except when $\rho = 0$. In that 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

Figure 6: Outcomes for sequential upper-tail bootstrap tests at 0.05 level




**Notes:** There is one regressor, generated by (35) with $\rho = 0.5$, plus cluster fixed effects. The regressand is generated by (36) with clustered disturbances at either the coarse level (left panel) or the fine level (right panel), for $\rho$ between 0.0 and 0.64. There are 8 coarse clusters, 48 fine clusters, and 2400 observations. Bootstrap tests use $B = 999$, and there are 400,000 replications. The solid red and dashed purple curves separate the three outcomes of the sequential procedure; the red curve separates N from F, and the purple curve F from C. The dashed blue curve shows the outcome of a direct test of N against C.

not inflate the Type I error. In this case, the true null is rejected almost exactly $\alpha$% 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 $\alpha$% of the time.

As $\rho$ increases, the procedure chooses N or F less and less often. For very small values of $\rho$, it chooses N or F more often than C, but that changes quickly as $\rho$ increases. The gap between the solid red and dashed purple curves shows the fraction of the time that F is (incorrectly) chosen. This gap is always small, and it vanishes as $\rho$ becomes large.

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 purple dashed curve 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 6, there is fine clustering in the DGP, except when $\rho = 0$. The sequential procedure again works very well. As $\rho$ increases, it incorrectly chooses no clustering a rapidly diminishing fraction of the time. For larger values of $\rho$, it incorrectly chooses coarse clustering about 5.2% of the time, because the bootstrap SV tests over-reject slightly with only 8 coarse and 48 fine clusters. 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 null hypothesis that the disturbances are not clustered. This is not surprising, since the alternative involves clustering at a coarser level than the DGP.

## 6.4 Making Inferences about a Regression Coefficient

In Section 3.5, we discussed several procedures for making inferences about a single regression coefficient when clustering may be either fine or coarse. We now investigate some of these procedures, notably pre-test ones based on SV tests. There are four simulation experiments, each involving 12 coarse clusters. In two of them, we pre-test the null of no clustering, and in the other two we pre-test the null of fine clustering with 96 fine clusters.

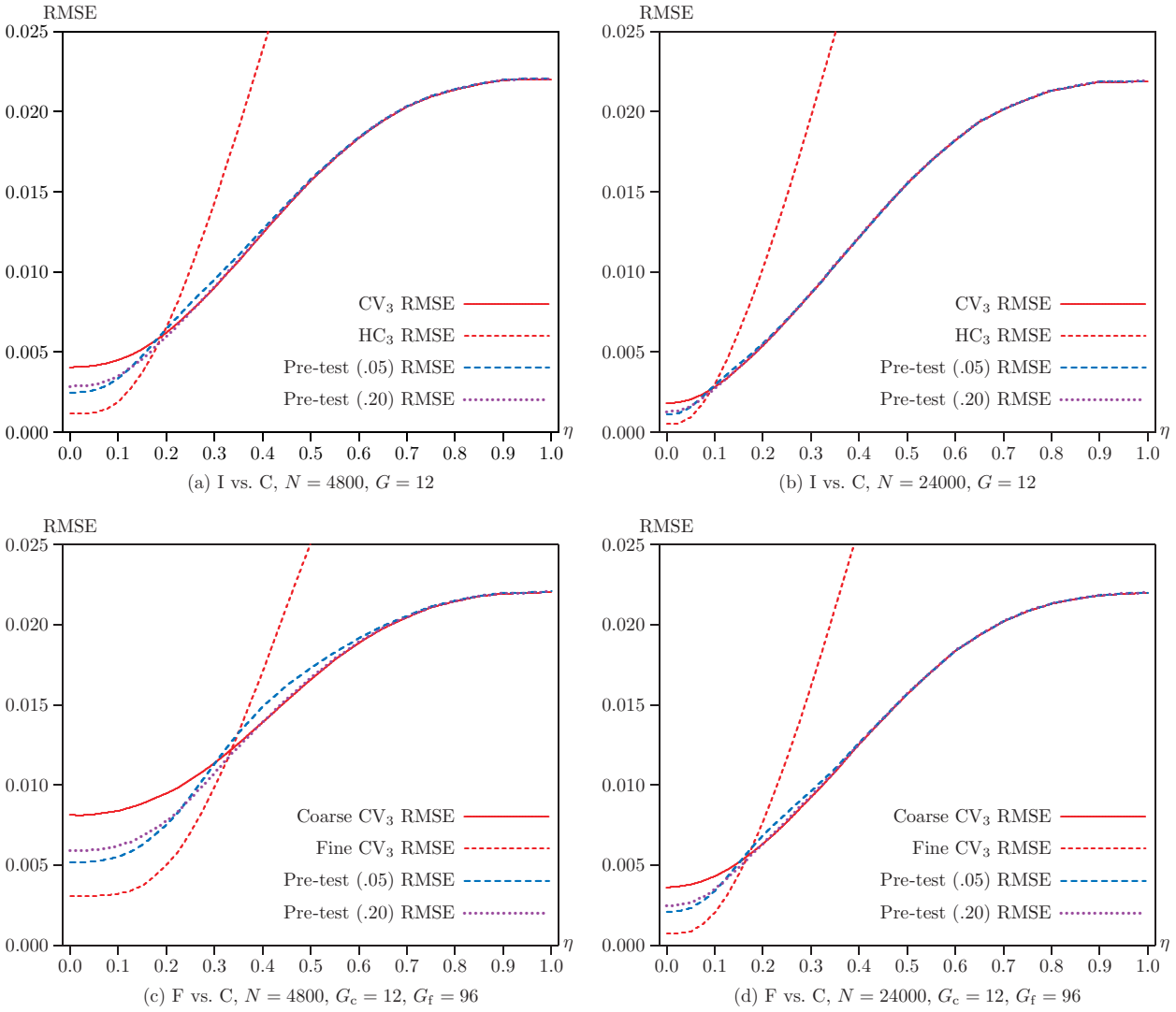
The model is a variant of (36), with eight regressors plus coarse-level fixed effects, so that $k = G_c + 8 = 20$. The regressors are generated by (35) with $\rho = 0.5$. The disturbances $u_{ghi}$ are generated as a convex combination of two disturbances, $\epsilon^c_{gi}$ and $\epsilon^f_{ghi}$, with weights $\eta$ and $1 - \eta$ respectively, rescaled so that the $u_{ghi}$ have unit variance. The $\epsilon^c_{gi}$ are generated by (35) with $\rho = 0.25$. When the pre-test null hypothesis is fine clustering, the $\epsilon^f_{ghi}$ are generated in the same way as the $\epsilon^c_{gi}$, but for 96 fine clusters instead of 12 coarse ones. When the pre-test null hypothesis is no clustering, the $\epsilon^f_{ghi}$ are i.i.d. normal.

The parameter $\eta$ determines the amount of correlation within coarse clusters. The pre-test null hypotheses are true when $\eta = 0$, so that there is either no intra-cluster correlation or only correlation within the fine clusters. The pre-test null hypotheses are false when $\eta > 0$, and the DGP moves further away from the pre-test null as $\eta$ increases. In the experiments, we vary $\eta$ from 0 to 1.

There are several asymptotically valid standard errors for coarse clustering, fine clustering, and no clustering. The best-known variance matrix estimator with clustering, often referred to as $\text{CV}_1$, is the usual sandwich estimator (4) with $\hat{\mathbf{\Sigma}}_c$ given by (5) or (10). However, recent work (Hansen 2022; MacKinnon et al. 2023b, 2022) suggests that the cluster jackknife, or $\text{CV}_3$, estimator usually performs better than $\text{CV}_1$, so we use the former for inference about the regression coefficient. For the case of no clustering, we use the $\text{HC}_3$ standard error of MacKinnon and White (1985), which is a jackknife estimator analogous to $\text{CV}_3$.

We focus on inference about $\beta_1$, one of the $\beta_\ell$ in (36). The pre-test estimators that we

Figure 7: Root mean squared errors of four standard error estimates




(a) I vs. C, $N = 4800$, $G = 12$

(b) I vs. C, $N = 24000$, $G = 12$

(c) F vs. C, $N = 4800$, $G_c = 12$, $G_f = 96$

(d) F vs. C, $N = 24000$, $G_c = 12$, $G_f = 96$

**Notes:** The regressors are generated by (35) with coarse clustering and $\rho = 0.5$, and the disturbances are generated as discussed in the second paragraph of this subsection. When $\eta = 0$, there is either no clustering (top panels) or fine clustering (bottom panels), depending on the pre-test null hypothesis. When $\eta > 0$, there is coarse clustering. The pre-test estimators are based on upper-tail $\tau_\sigma$ tests. There are 400,000 replications.

study are based on upper-tail $\tau_\sigma$ tests. Upper-tail tests are more powerful than two-sided tests, so that the former make fewer Type II errors; see Figure 4. Moreover, even when the difference between $\mathrm{Var_c}(\hat{\beta}_1)$ and $\mathrm{Var_f}(\hat{\beta}_1)$ is positive, $\widehat{\mathrm{Var}}_\mathrm{c}(\hat{\beta}_1)$ can be smaller than $\widehat{\mathrm{Var}}_\mathrm{f}(\hat{\beta}_1)$. This happens frequently in our experiments when $\eta$ is greater than 0 but small. Thus, investigators who do not wish to reject fine clustering in favor of coarse clustering when the coarse standard error is smaller than the fine one will choose to employ upper-tail pre-tests.

The choice among various standard errors is an estimation problem. Thus, it seems

reasonable to compare them on the basis of root mean squared error (RMSE). When the pre-test null hypothesis is no clustering, the standard error is based on $HC_3$, $CV_3$, or the one chosen by pre-tests at either the 0.05 or 0.20 level. When the pre-test null is fine clustering, the standard error is based on fine $CV_3$, coarse $CV_3$, or the one chosen by pre-tests at the same two levels. Figure 7 shows the RMSEs associated with each of these standard errors. In Panels (a) and (b), the pre-test null hypothesis is no clustering. In Panels (c) and (d), it is fine clustering, with 96 clusters. There are 4800 observations in Panels (a) and (c) and 24,000 in Panels (b) and (d).

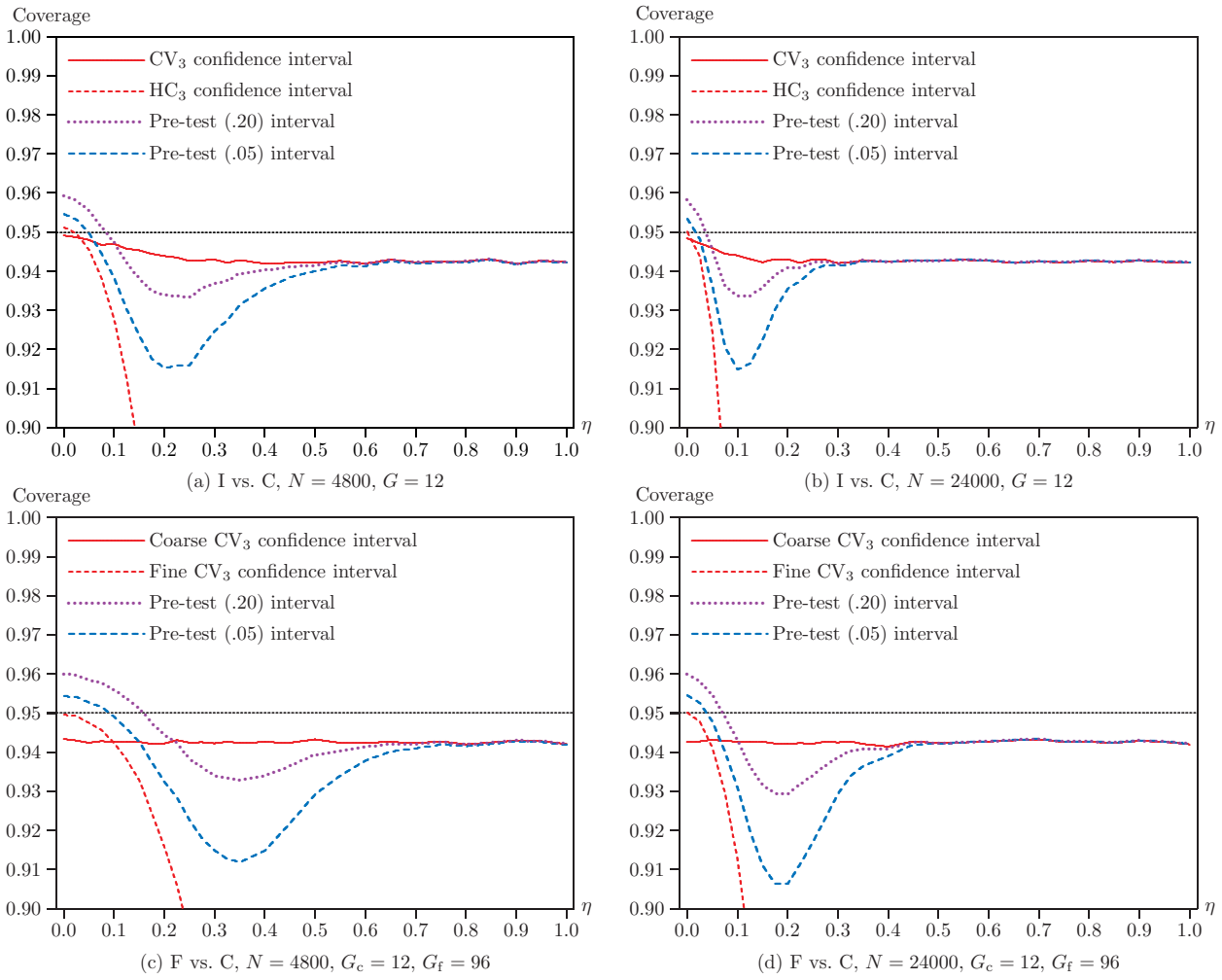
The $HC_3$ or fine $CV_3$ standard errors are the most accurate when $\eta = 0$, and they continue to be the most accurate for small values of $\eta$. However, for larger values of $\eta$, they are by far the least accurate, because they are severely biased. In contrast, the coarse $CV_3$ standard errors are the least accurate when $\eta$ is small, but for moderate and larger values of $\eta$ they are the most accurate. The two pre-test standard errors are substantially more accurate than the coarse $CV_3$ ones for small values of $\eta$ and almost identical to the latter for large values of $\eta$. In between, there is always a region where the pre-test standard errors are slightly less accurate than the coarse $CV_3$ ones. This is barely noticeable for pre-tests at the 0.20 level, but it is quite noticeable for pre-tests at the 0.05 level, especially in Panel (c), where the SV tests have the least power.

In our view, the 0.20 pre-test standard errors in Figure 7 perform substantially better than any of the others. They are much more accurate than coarse $CV_3$ standard errors for small values of $\eta$, slightly less accurate for some intermediate values, and essentially identical for larger values. Since using a more accurate standard error yields a confidence interval that provides a better sense of how reliable a coefficient estimate is, it seems reasonable to base confidence intervals on 0.20 pre-test standard errors.

Of course, using a more accurate standard error does not guarantee better coverage. Figure 8 shows the coverage of confidence intervals using the four standard errors in Figure 7. The coarsely-clustered intervals always under-cover to some extent. With only 12 clusters, that is not surprising. If we had used $CV_1$ instead of $CV_3$ to construct the intervals, they would have under-covered to a somewhat greater extent. On the other hand, coverage would almost certainly have been closer to 95% if we had used the wild cluster bootstrap (MacKinnon et al. 2023b), but that would have been computationally very demanding to simulate. The coverage using $HC_3$ and the finely-clustered $CV_3$ 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 $\eta$. Indeed, their coverage always rapidly drops below 0.90, the lower limit of the vertical axis.

The pre-test intervals over-cover slightly when $\eta = 0$, which is a consequence of Type I errors in the pre-tests. However, they under-cover more than the coarsely-clustered $CV_3$

Figure 8: Coverage of four 95% confidence intervals




**Notes:** These results are for the same experiments as in Figure 7.

intervals for intermediate values of $\eta$ because of Type II errors. The under-coverage is much more pronounced for pre-tests at the 0.05 level than for pre-tests at the 0.20 level. Because the sample size is five times larger in Panels (b) and (d) than in Panels (a) and (c), the pre-tests are more powerful, and the pre-test intervals converge more rapidly to the coarsely clustered $CV_3$ interval as $\eta$ increases.

To save computer time and programming effort, we use asymptotic SV tests in these experiments. In consequence, the levels of the pre-tests are not exactly 0.05 and 0.20. In particular, the actual levels of tests at the 0.20 level are noticeably lower than 0.20, and the ones for tests at the 0.05 level are somewhat higher than 0.05. If we had used bootstrap pre-tests, the under-coverage for moderate values of $\eta$ would have been a bit smaller for tests at the 0.20 level and a bit larger for tests at the 0.05 level. But all the curves for pre-

test confidence intervals would have looked very similar. They would also have looked very similar if we had used $\mathrm{CV}_1$ and $\mathrm{HC}_1$ instead of $\mathrm{CV}_3$ and $\mathrm{HC}_3$.

# 7 Empirical Example

We now illustrate the use of our score-variance tests in a realistic empirical setting. We employ 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:

$$\text{read-one}_{sri} = \alpha + \beta_s \text{small-class}_{sr} + \beta_a \text{aide-class}_{sr} + \boldsymbol{x}_{sri}^\top \boldsymbol{\delta} + u_{sri}. \tag{38}$$

The outcome variable $\text{read-one}_{sri}$ is the reading score in grade one of student $i$ in classroom $r$ in school $s$. We are interested in $\beta_s$ and $\beta_a$, 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 a teacher's aide. Additional control variables are collected in the vector of regressors $\boldsymbol{x}_{sri}$. 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 (38) are presented in the top half of Table 1. Two variants of the model are estimated. In the left panel, there is just a constant term. In the right panel, there are school fixed effects. 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. For each set, the first column reports results that are heteroskedasticity-robust (HR), using $\mathrm{HC}_3$ standard errors. The next two columns report results that are cluster-robust (CR) at either the classroom (R) level or the school (S) level, using $\mathrm{CV}_3$ standard errors. As in Section 6.4, we employ $\mathrm{HC}_3$ and $\mathrm{CV}_3$, instead of the more commonly-used $\mathrm{HC}_1$ and $\mathrm{CV}_1$ estimators, because the former tend to yield more

Table 1: STAR Example

| Estimates | | Without School FE | | | With School FE | | |
|---|---|---|---|---|---|---|---|
| | | $HC_3$(N) | $CV_3$(R) | $CV_3$(S) | $HC_3$(N) | $CV_3$(R) | $CV_3$(S) |
| small | $\hat{\beta}_s$ | 9.211 | 9.211 | 9.211 | 8.095 | 8.095 | 8.095 |
| | s.e. | 1.633 | 3.273 | 3.253 | 1.556 | 3.028 | 3.120 |
| | $t$-stat. | 5.640 | 2.814 | 2.831 | 5.203 | 2.673 | 2.595 |
| aide | $\hat{\beta}_a$ | 6.245 | 6.245 | 6.245 | 4.170 | 4.170 | 4.170 |
| | s.e. | 1.664 | 3.343 | 2.847 | 1.587 | 2.814 | 2.429 |
| | $t$-stat. | 3.752 | 1.868 | 2.194 | 2.628 | 1.482 | 1.717 |

| Cluster tests | | Without School FE | | | With School FE | | | |
|---|---|---|---|---|---|---|---|---|
| | | SV stat. | asy. $P$ | boot $P$ | SV stat. | asy. $P$ | boot $P$ | 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 | 0.540 | 0.543 | 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 | 0.961 | 0.973 | 1.871 | 0.031 | 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 clusters (denoted R for "room") or 75 school clusters (denoted S). The null hypotheses of no clustering, classroom clustering, and school clustering are called $H_N$, $H_R$, and $H_S$, respectively. Values of the $\tau_\sigma$ statistic (for "small" and "aide") or the $\tau_\Sigma$ statistic (for "both") are shown under "SV stat." All other numbers in the lower panel are $P$ values. For the $\tau_\sigma$ tests, asymptotic $P$ values are upper-tail and based on the $N(0,1)$ distribution. For the $\tau_\Sigma$ tests, they are based on the $\chi^2(3)$ distribution. Upper-tail bootstrap $P$ values use $B = 99{,}999$. IM tests use $S = 9{,}999$. Data and `Stata` files may be found at http://qed.econ.queensu.ca/pub/faculty/mackinnon/svtest/.

reliable inferences. The HR results would have been very similar if we had used $HC_1$ instead of $HC_3$. However, some of the CR results would have been noticeably different if we had used $CV_1$ instead of $CV_3$. The reason for this is interesting, and we discuss it below.

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; even more so if assignment was not entirely random.

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.

For the model without school fixed effects, the estimated impact on test scores of being in a small class is $\hat{\beta}_s = 9.211$. Based on an HR standard error of 1.63, the $t$-statistic for the null hypothesis that $\beta_s = 0$ is 5.64. When we instead use CR standard errors clustered at the classroom level, the standard error for $\beta_s$ increases to 3.23, and the $t$-statistic decreases to 2.81. Using CR standard errors clustered at the school level yields almost identical results; the standard error is 3.25, and the $t$-statistic is 2.83. In this case, the level at which we cluster makes no qualitative difference. For the model with school fixed effects, the estimate of the small-class effect is somewhat lower at $\hat{\beta}_s = 8.095$. The HR $t$-statistic is now 5.20, the classroom-level CR $t$-statistic is 2.67, and the school-level CR $t$-statistic is 2.57. Once again, the level at which we cluster does not change the conclusions.

The estimated effect on test scores of being in a class with an aide is $\hat{\beta}_a = 6.245$ without school fixed effects and $\hat{\beta}_a = 4.170$ with them. Based on the HR $t$-statistics, there seems to be fairly strong evidence that $\beta_a \neq 0$ for both models. However, when we cluster at the classroom level, we cannot reject this null hypothesis at the 0.05 level for either specification. When we cluster at the school level, we can do so for the model without fixed effects ($P = 0.031$), but not for the model with fixed effects.

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

For testing $\mathrm{H_N}$ against $\mathrm{H_R}$, the SV 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. For testing $\mathrm{H_N}$ against $\mathrm{H_S}$, the SV tests also very strongly reject the null in all cases. This is not surprising. 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 $\mathrm{H_N}$ when tested against $\mathrm{H_S}$.

IM tests can be computed when testing against $\mathrm{H_S}$, but only for the model with school fixed effects. For both coefficients, the IM tests suggest that $\mathrm{H_N}$ 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; see below for further discussion.

The results for testing $H_R$ against $H_S$ differ depending on the model, the coefficient(s) of interest, and the testing procedure. Consider first the model with no fixed effects. Here, both $\tau_\sigma$ statistics are negative, so of course upper-tail tests do not reject the null. This reflects the fact that, for both coefficients, the CR standard errors for school clustering are smaller than those for classroom clustering. The $\tau_\Sigma$ test for both coefficients jointly is always two-sided. With $P$ values of 0.157 (asymptotic) and 0.171 (bootstrap), it also fails to reject the null hypothesis. Thus we conclude that the classroom level is the right one at which to cluster for the model with just a constant term.

Consider next the model with school fixed effects. As we noted in Remarks 7 and 17, the "correct" level of clustering may be different for different hypotheses. This is what we find here. For $\hat{\beta}_s$, all three SV tests reject the null hypotheses and consequently suggest that school clustering is appropriate. In contrast, for $\hat{\beta}_a$, the SV tests suggest quite clearly (at least when using bootstrap $P$-values) that classroom clustering is appropriate.

Closer examination reveals that, for the model with school fixed effects, the asymptotic and bootstrap tests for $H_R$ against $H_S$ always yield quite different $P$ values. This is easily seen for $\beta_a$, where the bootstrap $P$ value of 0.344 is more than ten times the asymptotic $P$ value of 0.031. But it is also true for the other two tests. For $\beta_s$, the $\tau_\sigma$ test statistic of 4.366 has an asymptotic $P$ value of 0.000006 and a bootstrap $P$ value of 0.0044. For the joint test of both coefficients, the $\tau_\Sigma$ test statistic of 28.673 has an asymptotic $P$ value of 0.000003 and a bootstrap $P$ value of 0.0109. In the latter two cases, the bootstrap $P$ values are small, but they are many times larger than the asymptotic ones.

The differences between asymptotic and bootstrap $P$ values for SV tests of classroom against school clustering in the model with school fixed effects arise because there are only a few classrooms per school. The average is 4.4, and most schools have just 3 or 4 classrooms. Because the residuals are orthogonal to the school fixed effects, they must add to zero over all classrooms in each school. This mechanically creates negative correlation between the residuals across classrooms within each school, even if the disturbances are uncorrelated across classrooms. The negative correlation of the residuals leads to spurious correlation of the empirical scores whenever a regressor of interest, after being projected off the fixed effects and the other regressors, is correlated across classrooms within schools. Because student characteristics probably vary at the school level, this sort of correlation seems very likely.

In principle, the spurious correlation of the empirical scores could be either positive or negative. For the model (38), it is evidently positive and quite large. This explains why the bootstrap tests yield much larger $P$ values than the asymptotic tests. Equivalently, the

bootstrap critical values are greater than the asymptotic ones. For example, the test statistic for $H_N$ against $H_R$ for $\beta_a$ is 7.625. The asymptotic critical value for an upper-tail test at the 0.05 level is 1.645, but the bootstrap critical value is 3.423.

Whenever there is a dummy variable that affects only a few clusters (in this case the classrooms within each school), OLS residuals will be negatively correlated across those clusters, even when the disturbances are uncorrelated. This distortion of the residuals can cause cluster-robust inference to be severely misleading; see, among others, MacKinnon and Webb (2017, 2018) and de Chaisemartin and Ramirez-Cuellar (2022). However, $CV_3$ standard errors are almost certainly much more reliable in such cases than $CV_1$ standard errors. As MacKinnon et al. (2023b) explains, the cluster jackknife implicitly involves transforming the empirical scores in a way that undoes at least part of the distortion induced by least squares. This is evidently happening here.

With 330 clusters, we would normally expect $CV_1$ and $CV_3$ standard errors to be almost identical. But this is not the case for the model with fixed effects and classroom clustering. The $CV_1$ standard errors with classroom clustering for $\hat{\beta}_s$ and $\hat{\beta}_a$ are 2.322 and 2.109, respectively. These are much smaller than the $CV_3$ standard errors of 3.028 and 2.814 reported in Table 1. The latter are almost certainly much more reliable than the former. Note that the $CV_1$ standard error for $\hat{\beta}_a$ with school clustering is 2.422, which is almost identical to the $CV_3$ one in the table and greater than 2.109. Thus the ratio of the S and R standard errors is greater than one for $CV_1$ and less than one for $CV_3$. Because the former ratio is greater than one, the $\tau_\sigma$ statistic is positive.

In additional simulation experiments not reported here, we generated artificial samples using the actual regressors for the STAR model. When there are no school fixed effects, all the SV tests, both asymptotic and bootstrap, work very well. However, when there are fixed effects, the asymptotic tests over-reject severely (up to about 70% of the time). The bootstrap tests perform almost perfectly when testing $H_N$ against either $H_R$ or $H_S$, but they reject between 7% and 9% of the time for the tests of $H_R$ against $H_S$. We also performed some experiments in which the number of classrooms per school was doubled. All tests performed very much better in this case. These results suggest that, when there are fixed effects at the coarse level with few fine clusters per coarse cluster, and the asymptotic and bootstrap $P$ values differ sharply, the former should not be believed, and the latter should be taken with a grain of salt.

The IM tests are undoubtedly also affected by the odd properties of OLS residuals with school fixed effects. However, many of the differences between the score-variance tests and the IM tests in Table 1 probably arise because calculating the latter for the model (38) 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 $\beta_s$ and $\beta_a$ (17 schools in the case of $\beta_s$ and 21 schools in the case of $\beta_a$). 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 $\beta_s$ and $\beta_a$ 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.

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. Which of these levels we should cluster at depends on the model and the coefficient(s) of interest. With just a constant term, the sequential testing procedure, using either asymptotic or bootstrap tests, suggests that we should choose $\mathrm{H_R}$ and cluster at the classroom level. However, with school fixed effects, we should apparently choose $\mathrm{H_R}$ if interest focuses on $\beta_a$ and $\mathrm{H_S}$ if it focuses on $\beta_s$ or on both coefficients. Both choices lead us to conclude that the effect of small classes is positive and significant at the 0.05 level, while the effect of a teacher's aide is also positive but not significant at that level.

The fact that we obtain different results for the three SV tests should not be surprising. The test statistics depend on empirical scores, and they are different for the three tests because the $\boldsymbol{Z}$ matrices in (30), which are vectors for the $\tau_\sigma$ tests, are different; see Remark 17. For the model with fixed effects, the residuals are clearly correlated at the school level. While part of this correlation is evidently spurious and caused by the fixed effects, the bootstrap results suggest that the disturbances are surely correlated at the school level, because the $\tau_\sigma$ test for $\beta_s$ and the $\tau_\Sigma$ test for the two coefficients both reject quite strongly. For $\beta_a$ by itself, however, the scores are apparently not correlated, leading the $\tau_\sigma$ test not to reject in that case.

# 8 Conclusion

Empirical research that uses cluster-robust inference typically assumes that the level of clustering is known. When it is unknown, 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 (or SV) 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 an alternative 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; see Section 3.3.

The simplest of our two tests is based on the statistic $\tau_\sigma$. 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, upper-tail test or as a two-sided test. Since upper-tail $\tau_\sigma$ tests have more power than two-sided ones (Section 6.2), we believe that they will usually be the procedure of choice. The second variant, based on the Wald-like statistic $\tau_\Sigma$, 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 6, 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 7, 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. When the asymptotic tests are seriously over-sized, the bootstrap tests always perform much better.

Our score-variance tests are very different from the other tests for the correct level of clustering proposed in Ibragimov and Müller (2016) and Cai (2022); see Section 3.4. All these tests may provide valuable information, although we believe that SV tests are particularly intuitive. As we discuss in Section 3.5, SV tests can be used either as formal pre-tests for choosing the level at which to cluster or simply as robustness checks.

Both our simulation results and the empirical example suggest that SV 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; see Section 6.4.

The score-variance tests we have proposed are intended to provide guidance for applied researchers. In our view, it should be routine to report the results of SV tests whenever more than one level of clustering 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. In practice, however, it may be safest to report inferences based on more than one level of clustering, along with the outcomes of SV tests, as we did in Section 7.

# Appendix A: Proofs of Main Results

## A.1 Proof of Theorem 1

We give the proof for $\tau_\sigma$ only because the proof for $\tau_\Sigma$ is essentially the same but with more complicated notation. Also, because the factors $m_\mathrm{c}$ and $m_\mathrm{f}$ both converge to 1, we can ignore them in the proof.

Recall the contrast $\theta = \sum_{g=1}^{G} \sum_{h_1=1}^{M_g} \sum_{h_2 \neq h_1}^{M_g} s_{gh_1} s_{gh_2}$ defined in (16). 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 (18) 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

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

where we note that $w_{gh}$ is a martingale difference sequence with respect to the filtration $\mathcal{F}_{gh} = \sigma(\{s_{mn}\}_{m=1,\dots,g-1,n=1,\dots,M_m}, \{s_{gn}\}_{n=1,\dots,h})$, 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 $\mathrm{E}|s_{gh}|^{2\lambda} \leq cN_{gh}^{2\lambda}$ by Lemma A.1. We also find that

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

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

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

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 $q_{1,gh} = \sigma_{gh}^2 \sum_{j=1}^{h-1}(s_{gj}^2 - \sigma_{gj}^2)$ and $q_{2,gh} = \sigma_{gh}^2 \sum_{j_1=1}^{h-1} \sum_{j_2 \neq j_1}^{h-1} s_{gj_1} s_{gj_2}$. 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 $r_{1,gh} = (s_{gh}^2 - \sigma_{gh}^2) \sum_{j=h+1}^{M_g} \sigma_{gj}^2$ is mean zero and independent across both $g$ and $h$. We prove convergence in $L_\lambda$-norm. We find $\mathrm{E}|r_{1,gh}|^{\lambda} \leq c\mathrm{E}|s_{gh}|^{2\lambda}(\sum_{j=h+1}^{M_g} \sigma_{gj}^2)^{\lambda} \leq cN_{gh}^{2\lambda}(\sum_{j=1}^{M_g} \sigma_{gj}^2)^{\lambda} = cN_{gh}^{2\lambda}\sigma_g^{2\lambda}$ 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 $r_{2,gh} = 2s_{gh}\sum_{j_1=h+1}^{M_g}\sigma_{gj_1}^2\sum_{j_2=1}^{h-1}s_{gj_2} = w_{gh}\sum_{j=h+1}^{M_g}\sigma_{gj}^2$ 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 $\tau_\sigma$ only, and we ignore the asymptotically irrelevant factors $m_\mathrm{c}$ and $m_\mathrm{f}$. Under the conditions of Theorem 2, and specifically under Assumption 6, we find from (8) that $\boldsymbol{\Sigma}_\mathrm{c} = \sum_{g=1}^{G}\sigma_g^2$. However, it is important to note that, under the conditions of Theorem 2, $\sigma_g^2 = \mathrm{Var}(s_g) \neq \sum_{h=1}^{M_g}\sigma_{gh}^2$.

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 $\mathrm{E}(\theta)/\sum_{g=1}^{G}\sigma_g^2 = (\boldsymbol{\Sigma}_\mathrm{c} - \boldsymbol{\Sigma}_\mathrm{f})\boldsymbol{\Sigma}_\mathrm{c}^{-1}$ is non-zero in the limit under the alternative hypothesis, $\mathrm{H}_1$ in (9). 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

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

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

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

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 $\tau_\sigma$ only. The proof for $\tau_\Sigma$ is essentially the same but with slightly more complicated notation. The bootstrap probability measure is denoted $P^*$, and expectation under this measure is denoted $\mathrm{E}^*$. We define the bootstrap contrast $\theta^* = \sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}^{M_g} s_{gh_1}^* s_{gh_2}^*$, 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 $\sigma_g^2 = \text{Var}(s_g) \neq \sum_{h=1}^{M_g} \sigma_{gh}^2$ 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 $\sigma_g^4$ 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 $s_{gh}^* = \hat{s}_{gh} v_{gh}^*$, where $v_{gh}^*$ is independent across both $g$ and $h$, such that, c.f. (18) and (19),

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

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, $w_{gh}^* = 2 s_{gh}^* \sum_{j=1}^{h-1} s_{gj}^*$ 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 $r^*_{1,gh} = (s^{*2}_{gh} - \hat{s}^2_{gh}) \sum_{j=h+1}^{M_g} \hat{s}^2_{gj}$ and $r^*_{2,gh} = w^*_{gh} \sum_{j=h+1}^{M_g} \hat{s}^2_{gj}$. 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 $q^*_{3,gh} = (s^{*2}_{gh} - \hat{s}^2_{gh}) \sum_{j=1}^{h-1} s^{*2}_{gh}$. 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

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

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

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

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

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),

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

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 (15) and also define $\boldsymbol{\theta} = \sum_{g=1}^{G} \sum_{h_1=1}^{M_g} \sum_{h_2 \neq h_1}^{M_g} \mathrm{vech}(\boldsymbol{s}_{gh_1} \boldsymbol{s}_{gh_2}^{\top})$; c.f. (16). 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, $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| = O_P(\sup_g N_g^2)$.*

*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 note that

$$\hat{s}_{ghi} = x_{ghi}\hat{u}_{ghi} = x_{ghi}(u_{ghi} - x_{ghi}(\hat{\beta} - \beta_0)) = s_{ghi} - x_{ghi}^2(\hat{\beta} - \beta_0). \tag{A.24}$$

From (12), (14), and (16), we then find the difference

$$\hat{\theta} - \theta = \sum_{g=1}^{G} \sum_{h_1=1}^{M_g} \sum_{h_2 \neq h_1}^{M_g} \Big( \sum_{i=1}^{N_{gh_1}} x_{gh_1 i}^2 (\hat{\beta} - \beta_0) \Big) \Big( \sum_{i=1}^{N_{gh_2}} x_{gh_2 i}^2 (\hat{\beta} - \beta_0) \Big) \tag{A.25}$$

$$- 2 \sum_{g=1}^{G} \sum_{h_1=1}^{M_g} \sum_{h_2 \neq h_1}^{M_g} s_{gh_1} \sum_{i=1}^{N_{gh_2}} x_{gh_2 i}^2 (\hat{\beta} - \beta_0). \tag{A.26}$$

Using (2) we find that $\hat{\beta} - \beta_0 = (\boldsymbol{x}^\top \boldsymbol{x})^{-1} \sum_{g=1}^{G} \sum_{h=1}^{M_g} s_{gh}$. Under Assumption 1 we have

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

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

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

Hence, using also Assumption 3,

$$\begin{aligned} |\hat{\beta} - \beta_0| &= O_P\Big( N^{-1/2} \sup_{g,h} N_{gh}^{1/2} \Big) \quad \text{under Assumption 1,} \\ |\hat{\beta} - \beta_0| &= O_P\Big( N^{-1/2} \sup_{g} N_{g}^{1/2} \Big) \quad \text{under Assumption 6.} \end{aligned} \tag{A.29}$$

We also need the simple bounds

$$\sup_{g,h} N_{gh}^{-1} \sum_{i=1}^{N_{gh}} x_{ghi}^2 = O_P(1) \quad \text{and} \quad \sup_{g} N_g^{-1} \sum_{h=1}^{M_g} \sum_{i=1}^{N_{gh}} x_{ghi}^2 = O_P(1), \tag{A.30}$$

which follow from 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

$$(\hat{\beta} - \beta_0)^2 \sum_{g=1}^{G} \Big( \sum_{h=1}^{M_g} \sum_{i=1}^{N_{gh}} x_{ghi}^2 \Big)^2 = (\hat{\beta} - \beta_0)^2 O_P\Big( \sum_{g=1}^{G} N_g^2 \Big) = (\hat{\beta} - \beta_0)^2 O_P\Big( N \sup_{g} N_g \Big), \tag{A.31}$$

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

Next, we write (A.26) as

$$(A.26) = -2\sum_{g=1}^{G}\sum_{h_1=1}^{M_g} s_{gh_1} \sum_{h_2=1}^{M_g}\sum_{i=1}^{N_{gh_2}} x_{gh_2i}^2(\hat{\beta}-\beta_0) \tag{A.32}$$

$$+2\sum_{g=1}^{G}\sum_{h=1}^{M_g} s_{gh} \sum_{i=1}^{N_{gh}} x_{ghi}^2(\hat{\beta}-\beta_0). \tag{A.33}$$

Under Assumption 1, (A.27), (A.29), and (A.30) show that $|(A.32)| = O_P(\sup_{g,h} N_{gh} \sup_g N_g)$ and that $|(A.33)| = O_P(\sup_{g,h} N_{gh}^2)$. Under Assumption 6, (A.28), (A.29), and (A.30) show that $|(A.32)| = O_P(\sup_g N_g^2)$ and that $|(A.33)| = O_P(\sup_{g,h} N_{gh} \sup_g N_g)$. This proves the result for (A.26). □

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

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

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

As in (A.27), we find that

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

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 $\mathrm{E}^*|(A.34)| = O_P(\sup_{g,h} N_{gh} \sup_g N_g)$ as in (A.31). By the same argument, see also (A.32)–(A.33), we find that $\mathrm{Var}^*$ (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 (21). 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* $\boldsymbol{\delta}^\top \boldsymbol{H}_k = \boldsymbol{b}^\top = [\boldsymbol{b}_1^\top \otimes \boldsymbol{b}_2^\top]$,

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

*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 (19), and we show that

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

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

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

$$2\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}^{M_g} (\hat{s}_{gh_1}^2\hat{s}_{gh_2}^2 - s_{gh_1}^2 s_{gh_2}^2) \tag{A.38}$$
$$+ 2\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}^{M_g} (s_{gh_1}^2 s_{gh_2}^2 - \sigma_{gh_1}^2\sigma_{gh_2}^2). \tag{A.39}$$

We first prove the result for (A.39) under Assumption 1. We use (18) and (19) 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 $q_{1,gh} = \sigma_{gh}^2\sum_{j=1}^{h-1}(s_{gj}^2 - \sigma_{gj}^2)$ and $q_{3,gh} = (s_{gh}^2 - \sigma_{gh}^2)\sum_{j=1}^{h-1} s_{gj}^2$. Under Assumption 1 we have already proven in (A.10) that $\sum_{g=1}^{G}\sum_{h=1}^{M_g} q_{1,gh} = O_P(\sup_{g,h} N_{gh}^{3/2}\sup_g N_g^{1/2}(\sum_{g=1}^{G}\sigma_g^4)^{1/2})$. 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 $L_\lambda$-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 $\mathrm{E}|q_{3,gh}|^\lambda \leq \mathrm{E}|s_{gh}|^{2\lambda}\mathrm{E}\big|\sum_{j=1}^{h-1} s_{gj}^2\big|^\lambda$, 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 $\sum_{j=1}^{h-1} s_{gj}^2 \leq \sum_{j=1}^{M_g} s_{gj}^2$ is a non-negative random variable, and hence is of order $O_P(\mathrm{E}\sum_{j=1}^{M_g} s_{gj}^2) = O_P(\sum_{h=1}^{M_g}\sigma_{gh}^2)$. That is, there exists a constant $K < \infty$ and a set $\mathcal{A}$ with $P(\mathcal{A}) \to 1$ on which $\sum_{j=1}^{h-1} s_{gj}^2 \leq K\sum_{h=1}^{M_g}\sigma_{gh}^2$. Then, on the set $\mathcal{A}$,

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

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

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

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

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

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 $\sum_{j=1}^{h-1} s_{gj}^2 \leq K\sum_{j=1}^{M_g}\sigma_{gj}^2$. 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 (19) and (A.24) to write

$$\begin{aligned}
\text{(A.38)} &= 2\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}^{M_g}(\hat{s}_{gh_1}\hat{s}_{gh_2}+s_{gh_1}s_{gh_2})(\hat{s}_{gh_1}\hat{s}_{gh_2}-s_{gh_1}s_{gh_2}) \\
&= 2(\hat{\beta}-\beta_0)^2\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}^{M_g}(\hat{s}_{gh_1}\hat{s}_{gh_2}+s_{gh_1}s_{gh_2})\sum_{i=1}^{N_{gh_1}}x_{gh_1i}^2\sum_{j=1}^{N_{gh_2}}x_{gh_2j}^2 \qquad \text{(A.43)} \\
&\quad + 4(\hat{\beta}-\beta_0)\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}^{M_g}(\hat{s}_{gh_1}\hat{s}_{gh_2}+s_{gh_1}s_{gh_2})s_{gh_1}\sum_{i=1}^{N_{gh_2}}x_{gh_2i}^2. \qquad \text{(A.44)}
\end{aligned}$$

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

$$\begin{aligned}
\text{(A.44)} &= 8(\hat{\beta}-\beta_0)\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}^{M_g}s_{gh_1}^2 s_{gh_2}\sum_{i=1}^{N_{gh_2}}x_{gh_2i}^2 \qquad \text{(A.45)} \\
&\quad - 4(\hat{\beta}-\beta_0)^2\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}^{M_g}\Big(s_{gh_1}\sum_{i=1}^{N_{gh_2}}x_{gh_2i}^2+s_{gh_2}\sum_{j=1}^{N_{gh_1}}x_{gh_1j}^2\Big)s_{gh_1}\sum_{\ell=1}^{N_{gh_2}}x_{gh_2\ell}^2 \qquad \text{(A.46)} \\
&\quad - 4(\hat{\beta}-\beta_0)^3\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}^{M_g}s_{gh_1}\Big(\sum_{j=1}^{N_{gh_1}}x_{gh_1j}^2\Big)\Big(\sum_{i=1}^{N_{gh_2}}x_{gh_2i}^2\Big)^2. \qquad \text{(A.47)}
\end{aligned}$$

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,

$$\begin{aligned}
\text{(A.45)}^2 &\leq 64(\hat{\beta}-\beta_0)^2\Big(\sum_{g=1}^{G}\Big(\sum_{h_1=1}^{M_g}s_{gh_1}^2\sum_{i=1}^{N_{gh_1}}x_{gh_1i}^2\Big)^2\Big)\Big(\sum_{g=1}^{G}\Big(\sum_{h_2=1}^{M_g}s_{gh_2}\Big)^2\Big) \\
&\leq 64(\hat{\beta}-\beta_0)^2\Big(\sup_{g,h}\sum_{i=1}^{N_{gh}}x_{ghi}^2\Big)^2\Big(\sum_{g=1}^{G}\Big(\sum_{h_1=1}^{M_g}s_{gh_1}^2\Big)^2\Big)\Big(\sum_{g=1}^{G}\Big(\sum_{h_2=1}^{M_g}s_{gh_2}\Big)^2\Big). \qquad \text{(A.48)}
\end{aligned}$$

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 $O_P(\mathrm{E}\sum_{g=1}^{G}(\sum_{h=1}^{M_g}s_{gh})^2) = O_P(\sum_{g=1}^{G}\mathrm{E}s_g^2) = O_P(N\sup_g N_g)$ by Lemma A.1. Combining these results and using (A.29) and (A.30), we find that

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

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. □

# References


Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge (2023). When should you adjust standard errors for clustering? *Quarterly Journal of Economics 138*, 1–35.

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.

Bester, C. A., T. G. Conley, and C. B. Hansen (2011). Inference with dependent data using cluster covariance estimators. *Journal of Econometrics 165*, 137–151.

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

Cai, Y. (2022). A modified randomization test for the level of clustering. ArXiv e-prints 2105.01008v2, Northwestern University.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics 90*, 414–427.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics 29*, 238–249.

Cameron, A. C. and D. L. Miller (2015). A practitioner's guide to cluster-robust inference.

*Journal of Human Resources 50*, 317–372.

Carter, A. V., K. T. Schnepel, and D. G. Steigerwald (2017). Asymptotic behavior of a $t$ test robust to cluster heterogeneity. *Review of Economics and Statistics 99*, 698–709.

Chiang, H. D., K. Kato, and Y. Sasaki (2022). Inference for high-dimensional exchangeable arrays. *Journal of the American Statistical Association*, to appear.

Cho, J. S. and P. C. B. Phillips (2018). Pythagorean generalization of testing the equality of two symmetric positive definite matrices. *Journal of Econometrics 202*, 45–56.

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.

Davezies, L., X. D'Haultfœuille, and Y. Guyonvarch (2021). Empirical process results for exchangeable arrays. *Annals of Statistics 49*, 845–862.

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.

de Chaisemartin, C. and J. Ramirez-Cuellar (2022). At what level should one cluster standard errors in paired experiments, and in stratified experiments with small strata? Working Paper 27609, revised, 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.

Esarey, J. and A. Menger (2019). Practical and effective approaches to dealing with clustered data. *Political Science Research and Methods 7*, 541–559.

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. (2022). Jackknife standard errors for clustered regression. Working paper, University of Wisconsin.

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. Ø. Nielsen, and M. D. Webb (2021). Wild bootstrap and asymptotic inference with multiway clustering. *Journal of Business and Economic Statistics 39*, 505–519.

MacKinnon, J. G., M. Ø. Nielsen, and M. D. Webb (2022). Leverage, influence, and the jackknife in clustered regression models: Reliable inference using summclust. QED Working Paper 1483, Queen's University.

MacKinnon, J. G., M. Ø. Nielsen, and M. D. Webb (2023a). Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics 232*, 272–299.

MacKinnon, J. G., M. Ø. Nielsen, and M. D. Webb (2023b). Fast and reliable jackknife and bootstrap methods for cluster-robust inference. *Journal of Applied Econometrics*, to appear.

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). Clustering methods for statistical inference. In K. F. Zimmermann (Ed.), *Handbook of Labor, Human Resources and Population Economics*. Cham, Switzerland: Springer.

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

Menzel, K. (2021). Bootstrap with cluster-dependence in two or more dimensions. *Econometrica 89*, 2143–2188.

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. (2022). Reworking wild bootstrap based inference for clustered errors. *Canadian Journal of Economics*, to appear.

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.